

**В.В.УЧАЙКИН**

**ТЕОРИЯ ВЕРОЯТНОСТЕЙ  
И  
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА**

# Содержание

<b>1</b>	<b>Лекция 1. Случайные события и вероятности</b>	<b>11</b>
1.1	Исходы и события . . . . .	11
1.2	Постулаты вероятности . . . . .	12
1.3	Три теоремы о вероятности . . . . .	13
1.4	Правило сложения вероятностей . . . . .	14
1.5	Вычисление вероятностей . . . . .	14
1.6	Приложение: Алгебра событий . . . . .	17
<b>2</b>	<b>Лекция 2. Комбинаторика и вероятность</b>	<b>18</b>
2.1	Принцип умножения . . . . .	18
2.2	Перестановки . . . . .	19
2.3	Размещения . . . . .	20
2.4	Сочетания . . . . .	20
2.5	Разбиения . . . . .	21
2.6	Перестановки с повторяющимися элементами . . . . .	24
2.7	Выбор с возвращением . . . . .	24
2.8	Комбинаторные вероятности . . . . .	24
2.9	Спортлото . . . . .	26
2.10	Математическое дополнение: Биномиальные коэффициенты. Формула Стирлинга . . . . .	26
<b>3</b>	<b>Формула Стирлинга.</b>	<b>26</b>
<b>4</b>	<b>Примеры решения задач</b>	<b>27</b>
<b>5</b>	<b>Лекция 4. Условная вероятность и независимость</b>	<b>33</b>
5.1	Условная вероятность . . . . .	33
5.2	Правила умножения . . . . .	33
5.3	Пример применения формулы . . . . .	33
5.4	Теорема полной вероятности . . . . .	34
5.5	Теорема Байеса . . . . .	34
5.6	Примеры. . . . .	35
5.7	Независимые события . . . . .	35
<b>6</b>	<b>Лекция 5. Случайные величины</b>	<b>36</b>
6.1	Распределения вероятностей . . . . .	36
6.2	Плотность распределения вероятностей . . . . .	37
6.3	Кумулятивная функция распределения . . . . .	37
6.4	Графические представления распределений . . . . .	38
<b>7</b>	<b>Лекция 6. Моменты случайных величин</b>	<b>39</b>
7.1	Математическое ожидание . . . . .	39
7.2	Некоторые свойства математического ожидания . . . . .	39
7.3	Моменты случайных величин . . . . .	40
7.4	Дисперсия . . . . .	41
7.5	Производящая функция моментов . . . . .	42

<b>8</b>	<b>Лекция 7. Типичные задачи</b>	<b>43</b>
<b>9</b>	<b>Лекция 8. Совместно распределённые случайные величины</b>	<b>52</b>
9.1	Совместные распределения вероятностей . . . . .	52
9.2	Маргинальные распределения . . . . .	52
9.3	Условные распределения . . . . .	53
9.4	Независимые случайные величины . . . . .	54
9.5	Распределения сумм . . . . .	54
<b>10</b>	<b>Лекция 9. Суммирование случайных величин</b>	<b>55</b>
10.1	Математическое ожидание суммы . . . . .	55
10.2	Дисперсия суммы . . . . .	55
10.3	Свойства дисперсии и ковариации . . . . .	55
10.4	Коэффициент корреляции . . . . .	56
<b>11</b>	<b>Лекция 10. Дискретные распределения вероятностей</b>	<b>56</b>
11.1	Распределение Бернулли . . . . .	56
11.2	Равномерное дискретное распределение . . . . .	57
11.3	Биномиальное распределение . . . . .	57
11.4	Геометрическое распределение . . . . .	58
11.5	Распределение Пуассона . . . . .	59
<b>12</b>	<b>Лекция 11. Типичные задачи</b>	<b>61</b>
<b>13</b>	<b>Лекция 12. Непрерывные плотности вероятности</b>	<b>73</b>
13.1	Равномерная плотность . . . . .	73
13.2	Бета-плотность . . . . .	74
13.3	Экспоненциальная плотность . . . . .	74
13.4	Гамма-плотность . . . . .	75
<b>14</b>	<b>Лекция 13. Непрерывные плотности (окончание)</b>	<b>75</b>
14.1	Нормальная плотность . . . . .	75
14.2	Стандартная нормальная плотность . . . . .	76
14.3	Производящая функция моментов . . . . .	77
14.4	Центральная предельная теорема . . . . .	77
14.5	Двумерное нормальное распределение. . . . .	78
<b>15</b>	<b>Лекция 14. Решение типичных задач</b>	<b>78</b>
15.1	<b>Задача 2.</b> . . . . .	78
<b>16</b>	<b>Лекция 15. Random Sampling</b>	<b>90</b>
16.1	Генеральная совокупность и выборка . . . . .	90
16.2	Статистический метод . . . . .	90
16.3	Вероятностная модель случайной выборки . . . . .	91
16.4	Распределения генеральной совокупности и их характеристики . . . . .	91

<b>17 Лекция 16. Выборочные распределения</b>	<b>92</b>
17.1 Статистики и оценки . . . . .	92
17.2 Выборочное среднее . . . . .	92
17.3 Выборочное среднее . . . . .	93
17.4 Нормированное выборочное среднее . . . . .	94
17.5 Отношение двух выборочных дисперсий . . . . .	95
<b>18 Лекция 17. Типичные задачи</b>	<b>96</b>
18.1 Порядковые статистики . . . . .	96
18.2 Сравнение вычисления параметров генеральной совокупности и выборки	96
18.3 Выборочные распределения . . . . .	97
18.4 Примеры . . . . .	97
18.5 Пример 4. . . . .	98
<b>19 Лекция 18. Интервальные оценки</b>	<b>104</b>
19.1 Статистические выводы, оценки и проверка гипотез . . . . .	104
19.2 Понятие интервальной оценки . . . . .	104
19.3 Доверительные интервалы для $\mu$ ( $\sigma$ известны) . . . . .	105
19.4 Доверительные интервалы для $\mu$ : $\sigma$ неизвестна . . . . .	106
19.5 Доверительные интервалы для $\sigma$ . . . . .	106
<b>20 Лекция 19. Оценки по двум выборкам</b>	<b>108</b>
20.1 Разность двух средних (дисперсии известны) . . . . .	108
20.2 Разность между средними (Дисперсии одинаковы, но неизвестны) . . .	108
20.3 Разность между средними (дисперсии неизвестны, но не обязательно одинаковы) . . . . .	109
20.4 Отношение дисперсий . . . . .	109
<b>21 Lecture 20. Типичные задачи</b>	<b>111</b>
21.1 Оценка пропорции . . . . .	111
21.2 Пример с небольшой генеральной совокупностью . . . . .	111
21.3 Пример со среднеквадратичной ошибкой . . . . .	112
21.4 Пример со стандартным отклонением . . . . .	114
<b>22 Лекция 21. Проверка гипотез</b>	<b>122</b>
22.1 Два типа гипотез . . . . .	122
22.2 Два типа ошибок . . . . .	122
22.3 Процедура проверки гипотез (классическая, $\alpha$ -подход) . . . . .	123
22.4 Процедура проверки гипотез (подход на основе Р-значения) . . . . .	123
22.5 Пример . . . . .	123
<b>23 Lecture 22. Критерий согласия ????Goodness-of-Fit Test</b>	<b>124</b>
23.1 Столбчатые диаграммы . . . . .	124
23.2 Гистограммы . . . . .	125
23.3 Хи-квадрат распределение в выборках с известной $f(x)$ . . . . .	126
23.4 Критерий согласия $\chi^2$ . . . . .	127

<b>24 Lecture 23. Типичные задачи</b>	<b>129</b>
24.1 Пример 1. Односторонняя проверка гипотез ???One-sided HT . . . . .	129
24.2 Пример 2. Двусторонний тест . . . . .	130
24.3 Пример 3. Построение гистограмм . . . . .	130
24.4 Пример 4. Критерий согласия . . . . .	131
<b>25 Lecture 24. Корреляционный анализ</b>	<b>137</b>
25.1 Двухмерные распределения и корреляции . . . . .	137
25.2 Двухмерное нормальное распределение . . . . .	138
25.3 Корреляции в нормальных распределениях . . . . .	139
25.4 Оценка коэффициента корреляции . . . . .	139
25.5 Пример . . . . .	140
<b>26 Лекция 25. Регрессионный анализ</b>	<b>141</b>
26.1 Линейная регрессионная модель . . . . .	141
26.2 Два примера подгонки ???Fitting . . . . .	141
26.3 Метод наименьших квадратов . . . . .	142
26.4 Некоторые другие формы второго коэффициента . . . . .	143
26.5 Пример . . . . .	143
<b>27 Лекция 26. Вариационный анализ ???Analysis of Variance (ANOVA)</b>	<b>144</b>
27.1 Основная идея ANOVA . . . . .	144
27.2 Меры вариации . . . . .	144
27.3 Тестовая статистика для ANOVA . . . . .	145
27.4 Процедура проверки ANOVA . . . . .	145
27.5 Пример. . . . .	145
<b>28 The LAST Final exam problems (Fall 2005)</b>	<b>146</b>
28.1 Задача 1 . . . . .	146
28.2 Задача 2. . . . .	147
28.3 Задача 3. . . . .	147
28.4 Задача 4. . . . .	148
28.5 Задача 5. . . . .	148
<b>29 Лекция 27. Типичные задачи</b>	<b>156</b>
29.1 Пример . . . . .	156
29.2 Тест на корреляцию . . . . .	157
29.3 Пример (???page 394). . . . .	157
<b>30 Лекция 25. Корреляция в многомерных распределениях</b>	<b>157</b>
30.1 Биномиальное распределение . . . . .	157
30.2 Триномиальное распределение . . . . .	158
30.3 Двухмерное нормальное распределение . . . . .	158
30.4 Корреляции в нормальных распределениях . . . . .	159
30.5 Регрессия в нормальном распределении . . . . .	160

<b>31 Лекция 25. Регрессионный анализ</b>	<b>162</b>
31.1 Модель линейной регрессии . . . . .	162
31.2 Метод наименьших квадратов . . . . .	162
31.3 Второе представление для $b$ . . . . .	163
31.4 Доверительный интервал для $\beta$ . . . . .	163
31.5 Коэффициент корреляции . . . . .	163
31.6 Мультиномиальные коэффициенты . . . . .	165
31.7 Мультиномиальное распределение . . . . .	165
<b>32 Лекция 26. Двухмерное нормальное распределение</b>	<b>167</b>
32.1 Двухмерное нормальное распределение . . . . .	167
<b>33 Лекция 16. Выборочное распределение</b>	<b>170</b>
33.1 Выборочное среднее . . . . .	170
33.2 Свойства выборочного среднего . . . . .	170
33.3 Выборочная медиана и мода . . . . .	170
33.4 Размах выборки, выборочная дисперсия и выборочное стандартное отклонение . . . . .	171
33.5 Основные свойства выборочной дисперсии . . . . .	171
33.6 Нормальное распределение . . . . .	171
33.7 Хи-квадрат распределение . . . . .	172
33.8 $t$ -распределение . . . . .	172
33.9 $F$ -распределение . . . . .	173
33.10 Приложения статистических распределений . . . . .	173
33.11 Выборочное распределение среднего значения нормальной генеральной совокупности . . . . .	174
<b>34 Лекция 18. Выборочные распределения. [Учебник: Разделы 8.4; 8.5; 8.6; 8.7; 8.8]</b>	<b>174</b>
34.1 Выборочные распределения среднего: $\sigma$ известно . . . . .	174
34.2 Выборочное распределение величины $S^2$ . . . . .	175
34.3 Выборочное распределение среднего: $\sigma$ неизвестно . . . . .	176
34.4 Разница между средними . . . . .	176
34.5 Отношение двух дисперсий . . . . .	176
34.6 Центральная предельная теорема . . . . .	176
34.7 Применение ЦПТ к аппроксимации распределений . . . . .	177
34.8 Многомерное нормальное распределение . . . . .	178
34.9 Симметричные многомерные распределения . . . . .	178
34.10 Изотропные многомерные распределения . . . . .	178
34.11 Многомерные устойчивые распределения . . . . .	180
34.12 Примеры . . . . .	182
34.13 Выборочная оценка коэффициента корреляции . . . . .	182
34.14 Тест на корреляцию . . . . .	183
34.15 Пример (стр. 394). . . . .	183
34.16 Взаимосвязь между нормальными и стандартными нормальными случайными величинами . . . . .	184

<b>35 Лекция 14. Преобразования случайных величин</b>	<b>184</b>
35.1 Теорема преобразования . . . . .	184
35.2 Преобразование стандартных равномерных случайных величин . . . . .	184
35.3 Пример: преобразование стандартной равномерной с.в. в экспоненциальную с.в. . . . .	184
35.4 Пример: преобразование экспоненциальной с.в. в гамма-распределённую с.в. . . . .	185
35.5 Плотность Вейбулла . . . . .	185
35.6 Логнормальная плотность . . . . .	185
<b>36 Лекция 15. Производящая функция моментов</b>	<b>185</b>
36.1 Производящая функция моментов . . . . .	185
36.2 Пример: $M_X(t)$ для распределения Пуассона . . . . .	186
36.3 Пример: $M_X(t)$ для нормального распределения . . . . .	186
36.4 Основные свойства ПФМ . . . . .	186
36.5 Дискретные распределения . . . . .	187
36.6 Непрерывные распределения . . . . .	188
36.7 Закон больших чисел . . . . .	188
36.8 Устойчивость нормальных распределений . . . . .	188
<b>37 FOR INSERTING</b>	<b>189</b>
<b>38 Парадокс игры в кости и его разрешение.</b>	<b>189</b>
<b>39 Понятие вероятности.</b>	<b>189</b>
<b>40 Алгебра событий.</b>	<b>190</b>
<b>41 Аксиоматическое определение вероятности.</b>	<b>190</b>
<b>42 Множественно - вероятностный словарь.</b>	<b>191</b>
<b>43 Принцип умножения</b>	<b>191</b>
<b>44 Перестановки и размещения.</b>	<b>191</b>
<b>45 Сочетания</b>	<b>192</b>
<b>46 Перестановки объектов с повторениями</b>	<b>192</b>
<b>47 Формула Стирлинга.</b>	<b>193</b>
<b>48 Выборка без возвратов.</b>	<b>194</b>
<b>49 Выборка с возвратом.</b>	<b>194</b>
<b>50 Гипергеометрическое распределение.</b>	<b>195</b>
<b>51 Спортлото</b>	<b>195</b>

52 Условные вероятности.	195
53 Основное правило исчисления вероятностей.	196
54 Формула полной вероятности.	196
55 Формула Байеса	196
56 Дискретная случайная величина.	197
57 Схема Бернулли.	197
58 Теорема Пуассона	198
59 Функция распределения и ее свойства.	198
60 Плотность распределения.	199
61 Дельта-функция Дирака.	200
62 Моменты случайной величины.	201
63 Медиана и мода.	201
64 Дисперсия.	201
65 Симметричные случайные величины.	202
66 Характеристики формы распределения.	202
67 Совместное распределение случайной величины.	203
68 Условная плотность распределения.	203
69 Формула полного математического ожидания.	204
70 Независимые случайные величины.	204
71 Изотропный вектор с независимыми координатами.	204
72 Математическое ожидание (среднее значение).	205
73 Математическое ожидание и моменты суммы случайной величины.	205
74 Дисперсия суммы, ковариация.	206
75 Коэффициент корреляции.	206
76 Распределение суммы случайной величины.	207
77 Функция от случайной величины.	208



78 Математическое ожидание.	208
79 Интегральное представление $\delta$ -функции.	208
80 Характеристическая функция.	209
81 Свойства характеристической функции.	209
82 Характеристические функции и моменты.	210
83 Характеристическая функция нормального распределения.	210
84 Закон больших чисел.	210
85 Центральная предельная теорема.	211
86 Теорема Муавра-Лапласа	211
87 Устойчивость нормального закона.	211
88 Характеристическая функция симметричного устойчивого закона.	212
89 Многократные свертки распределений.	212
90 Многомерное нормальное распределение.	213
91 Гамма-распределения.	214
92 Бета-распределение.	214
93 $\chi^2$ -распределение.	215
94 Распределение Фишера ( $F$ -распределение Фишера).	216
95 Распределение Стьюдента.	216
96 Обобщенная ЦПТ.	217
97 Система Пирсона.	218
98 Таблица плотностей.	218
 I Математическая статистика.	 218
99 Понятие выборки.	219
100 Математическое ожидание и дисперсия частоты события.	219
101 Математическое ожидание и дисперсия выборочного среднего.	219
102 Математическое ожидание выборочной дисперсии.	220

103	Дисперсия выборочной дисперсии.	221
104	Распределение выборочной дисперсии.	221
105	Распределение относительной погрешности.	222
106	Распределение частоты событий.	222
107	Распределение выборочного среднего.	223
108	Распределение выборочной дисперсии.	223
109	Квантили.	224
110	Интервальные оценки.	224
111	Интервальные оценки среднего и дисперсии в нормальной выборке.	225
112	Доверительный интервал для среднего при известной дисперсии (нормальное значение).	225
113	Доверительный интервал для дисперсии при известном среднем (нормальный закон).	226
114	Метод наибольшего правдоподобия.	227
115	Оценка параметров нормального распределения.	227
116	Распределение выборочной дисперсии в нормальной выборке.	228
117	Эмпирические распределения.	228
118	Критерий .	229
119	Обобщенные функции (распределения)	229
120	Критерий $\chi^2$ для сравнения распределений.	230
121	Анализ оценок $\hat{a}$ и $\hat{b}$ .	230
122	Формула Стирлинга.	231
123	Центральная предельная теорема.	232
124	Другие теоремы	232
124.1	Проверка на двух средних . . . . .	233

# ТЕОРИЯ ВЕРОЯТНОСТЕЙ

## 1 Лекция 1. Случайные события и вероятности

### 1.1 Исходы и события

Любую операцию, результат которой не предопределен (то есть, ее повторение при сохранении всех условий не гарантирует повторение результата), назовем *статистическим экспериментом*. Простейшие примеры – бросание монеты, игрального кубика, вынимание карты из перетасованной колоды, игра в лотерею, стрельба по мишени.

Любой возможный результат статистического эксперимента называют *исходом* (синоним – *элементарное событие*). Будем обозначать его греческой буквой  $\omega$ .

Множество всех возможных исходов называется *пространством элементарных событий* данного статистического эксперимента и обозначается буквой  $\Omega = \{\omega\}$  (символ  $\{\omega\}$  означает *множество элементов  $\omega$* ).

Принадлежащее определенному классу (*борелевскому полю*) подмножество пространства  $\Omega$  называется *событием*. События обозначаются заглавными латинскими буквами  $A, B, C, \dots$ , кроме  $\Omega$ , называемого *достоверным событием* и пустого множества  $\emptyset$ , называемого *невозможным событием*.

На множестве событий можно ввести *алгебру* – операции, подобные сложению и умножению, в результате которых получаются новые события.

*Суммой двух событий  $A$  и  $B$  называется событие  $A + B$ , включающее в себя все исходы события  $A$  и все исходы события  $B$ : оно наступает всякий раз, когда наступает одно из событий  $A$  или  $B$ , или наступают оба.*

*Произведением двух событий  $A$  и  $B$  называется событие  $A \cdot B$ , включающее в себя все исходы, общие для событий  $A$  и  $B$ : оно наступает всякий раз, когда наступают оба события  $A$  и  $B$ .*

События  $\emptyset$  и  $\Omega$  играют роль нуля и единицы соответственно:

$$A + \emptyset = A, \quad A \cdot \emptyset = \emptyset, \quad A + \Omega = \Omega, \quad A \cdot \Omega = A.$$

*Событием, противоположным  $A$ , называется событие  $A'$ , включающее в себя все исходы, не входящие в  $A$ : оно наступает всякий раз, когда не наступает событие  $A$ . Очевидно,*

$$A + A' = \Omega, \quad A \cdot A' = \emptyset.$$

События  $A$  и  $B$  называются несовместными, если  $A \cdot B = \emptyset$ : несовместные события не могут произойти оба в одном статистическом эксперименте, появление одного из них исключает появление другого. События  $A$  и  $A'$  несовместны.

Конечный или счётный набор попарно несовместных событий  $A_k$ ,  $k = 1, 2, 3, \dots$ , представляющий собой разбиение множества  $\Omega$ , то есть, удовлетворяющий условию

$$A_1 + A_2 + A_3 + \dots = \Omega, \quad A_i \cdot A_j = \emptyset \text{ при } i \neq j,$$

называется **полной группой событий**. Результатом эксперимента может быть одно и только одно из событий полной группы и никакое другое. События  $A$  и  $A'$  образуют полную группу.

#### Пример 1.

Статистический эксперимент: двукратное бросание монеты, на одной стороне которой герб ( $\Gamma$ ), на другой – решка ( $P$ ). Пространство  $\Omega$  элементарных событий состоит из четырех элементов:  $(\Gamma, \Gamma)$ ,  $(\Gamma, P)$ ,  $(P, \Gamma)$ ,  $(P, P)$ . Событие "монета упала поразному" состоит из двух элементарных событий  $(\Gamma, P)$  и  $(P, \Gamma)$ , событие "монета упала одинаковым образом" состоит из двух других элементарных событий  $(\Gamma, \Gamma)$  и  $(P, P)$ .

**Пример 2.** Статистический эксперимент: двукратное бросание игрального кубика с гранями, пронумерованными от 1 до 6. Пространство  $\Omega$  элементарных событий состоит из тридцати шести элементов:  $(1, 1)$ ,  $(1, 2)$ ,  $(1, 3)$ , ...,  $(4, 6)$ ,  $(5, 6)$ ,  $(6, 6)$ , которые удобно представить в виде таблицы:

(1, 1)	(2, 1)	(3, 1)	(4, 1)	<b>(5, 1)</b>	(6, 1)
(1, 2)	(2, 2)	(3, 2)	<b>(4, 2)</b>	(5, 2)	(6, 2)
(1, 3)	(2, 3)	<b>(3, 3)</b>	(4, 3)	(5, 3)	(6, 3)
(1, 4)	<b>(2, 4)</b>	(3, 4)	(4, 4)	(5, 4)	(6, 4)
<b>(1, 5)</b>	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

Событие "сумма чисел равна 6" состоит из пяти элементарных исходов, отмеченных полужирным шрифтом.

Таблица эта представляет собой частный случай *клеточных диаграмм*, удобных, когда число элементарных событий конечно или счетно. В общем случае удобно пользоваться *диаграммами Венна*, на которых  $\Omega$  представляется в виде квадрата,  $\omega$  – точками этого квадрата, события – фигурами (кругами или их частями) внутри квадрата (рис 1.1).

## 1.2 Постулаты вероятности

В обыденной речи мы довольно часто употребляем слова "вероятно", "невероятно", "маловероятно", "вполне вероятно", "очень вероятно" по отношению к событию, которое может наступить, а может и не наступить в условиях данного эксперимента. Чтобы сделать эти высказывания по отношению к событиям  $A_1, A_2$  и т.д. более точными, вместо этих слов используются числа  $P(A_1)$ ,  $P(A_2)$  и т.д. Другими словами, на множестве возможных событий задается функция  $P(A)$ , называемая *вероятностью*. Естественно принять вероятность невозможного

события равной нулю и считать что достоверное событие характеризуется максимальным значением вероятности, в качестве которого удобно выбрать единицу. Этого, однако, недостаточно, чтобы находить вероятности одних событий по известным вероятностям других – нужны правила исчисления вероятностей.

Поскольку каждое событие интерпретируется как множество (исходов),  $P(A)$  является функцией множеств. Функциями множеств являются длина участка кривой, площадь плоской фигуры, объем и масса трёхмерной фигуры и др. Приведённые примеры принадлежат к определённому классу функций, называемых **мерами**. Функция  $M(A)$  называется мерой, если  $M(\emptyset) = 0$ ,  $M(A) \geq 0$  и  $M(A + B) = M(A) + M(B)$  для любых непересекающихся множеств  $A$  и  $B$ . А.Н.Колмогоров определил вероятность как **вероятностную меру** на множестве подмножеств пространства  $\Omega$  (на множестве событий, по вероятностной терминологии), то есть, функцию  $P(A)$ , удовлетворяющую **постулатам вероятности**:

- 1)  $P(A) \geq 0$  для любого события  $A$ ;
- 2)  $P(A + B) = P(A) + P(B)$ , если  $A$  и  $B$  несовместны;
- 3)  $P(\Omega) = 1$ .

Постулат 2 легко распространяется на произвольное число попарно несовместных событий:

$$2^*) P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

### 1.3 Три теоремы о вероятности

Пользуясь приведённым выше определением, нетрудно доказать следующие простые теоремы.

**Теорема 1.**  $P(A) + P(A') = 1$ .

**Доказательство.** События  $A$  и  $A'$  несовместны и  $A + A' = \Omega$ , поэтому  $P(A) + P(A') = P(A + A') = P(\Omega) = 1$ .

**Теорема 2.**  $P(A) \leq 1$  для любого  $A$ .

**Доказательство.** Из первого постулата следует, что  $P(A') \geq 0$ , а согласно теореме 1  $P(A) = 1 - P(A')$ , откуда и следует утверждение теоремы.

**Теорема 3.**  $P(\emptyset) = 0$ .

**Доказательство.** Достаточно положить в теореме 1  $A = \emptyset$ ,  $A' = \Omega$  и применить третий постулат.

**Замечание.** Теорема 1 обобщается на любую полную группу событий  $A_1, A_2, A_3, \dots$ :

$$P(A_1) + P(A_2) + P(A_3) + \dots = 1. \quad (1)$$

Равенство (1) часто называют **условием нормировки вероятности**. Можно сказать, что полная (равная единице) вероятность распределена между событиями полной группы подобно тому, как полная масса механической системы распределена

между составляющими её частями. Совокупность чисел  $p_k = P(A_k), k = 1, 2, 3, \dots$ , поэтому называют *распределением вероятностей*.

## 1.4 Правило сложения вероятностей

Постулат 2 представляет собой правило сложения вероятностей *несовместных событий*. В случае произвольной пары событий  $A$  и  $B$  **правило сложения вероятностей** гласит:

$$P(A + B) = P(A) + P(B) - P(A \cdot B).$$

**Доказательство.** Пусть  $A_1$  – множество всех элементов  $A$ , не принадлежащих  $B$ , а  $B_1$  – множество всех элементов  $B$ , не принадлежащих  $A$ . Очевидно,

$$A + B = A_1 + (A \cdot B) + B_1,$$

где правая часть представляет собой сумму несовместных событий, и поэтому

$$P(A + B) = P(A_1 + (A \cdot B) + B_1) = P(A_1) + P(A \cdot B) + P(B_1).$$

Учитывая, что

$$P(A_1) + P(A \cdot B) = P(A)$$

и

$$P(B_1) + P(A \cdot B) = P(B),$$

приходим к правилу сложения вероятностей.

Для трех событий оно записывается в виде

$$P(A + B + C) = P(A) + P(B) + P(C) - P(A \cdot B) - P(A \cdot C) - P(B \cdot C) + P(A \cdot B \cdot C).$$

**Следствие:**  $P(A) + P(B) - 1 \leq P(A \cdot B) \leq P(A) + P(B)$ .

## 1.5 Вычисление вероятностей

Продолжая комбинировать правила аксиомы алгебры событий и постулаты теории вероятностей, мы будем получать новые правила и теоремы о действиях с вероятностями различных событий. При этом каждый раз исходные вероятности должны быть известны. Так, чтобы воспользоваться выведенным выше правилом сложения для нахождения вероятности  $P(A + B)$ , мы должны знать вероятности  $P(A)$ ,  $P(B)$  и  $P(A \cdot B)$ . Откуда же берутся изначальные вероятности?

Существуют три метода их получения.

**1. Предположение о равновероятных исходах.** Существуют задачи с высокой степенью симметрии, позволяющей найти изначальные вероятности без обращения к теории или эксперименту. Например, идеальная ("правильная") монета является симметричной, отсюда делается вывод, что при надлежащем бросании выпадение орла и решки равновероятны. Они образуют полную группу событий, поэтому  $P(\Gamma) = P(P) = 1/2$ . Вероятность любого исхода при бросании игрального кубика  $1/6$ , вероятность любого исхода при бросании двух кубиков равна  $1/36$ . Событие  $A$ : сумма очков на обоих кубиках равна 6, включает в себя пять исходов

(1,5), (2,4), (3,3), (4,2) и (5,1), и правило сложения вероятностей даёт:  $P(A) = 5/36$  (рис.1.2). Вообще, если событие  $A$  состоит из  $n(A)$  равновероятных исходов, полное число которых есть  $n(\Omega)$ , то

$$P(A) = \frac{n(A)}{n(\Omega)}.$$

Это – **классическое определение вероятности в схеме с конечным числом равновероятных исходов**. Вычисление числа  $n(\Omega)$  *всех возможных* исходов и числа  $n(A)$  *благоприятных* (по отношению к событию  $A$ ) исходов в общем случае осуществляется с применением правил *комбинаторики*.

Если эксперимент таков, что  $\Omega$  – бесконечное множество или множество мощности континуум, например, гиперкуб в  $d$ -мерном пространстве, то приписывая всем его точкам равные вероятности, мы получим нулевые значения. Если всё-таки имеются достаточные основания считать все точки равноправными, например, если мы говорим о положении отмеченной молекулы газа в замкнутом объёме  $\Omega$  спустя длительное время после её введения в этот объём, то **равные вероятности приписываются областям с равными геометрическими мерами  $M$  (длинами, площадями, объёмами)**, и вероятность обнаружить случайную точку в области  $A$  определяется соотношением:

$$P(A) = \frac{M(A)}{M(\Omega)}.$$

Такое определение вероятности называется **геометрическим**.

**Пример.** В квадрате со стороной  $L$  случайным образом выбирается точка так, что все её положения равновозможны. Вероятность того, что она попадёт во вписанный в этот квадрат круг, равна отношению соответствующих площадей:

$$P(A) = \frac{S(A)}{S(\Omega)} = \frac{\pi(L/2)^2}{L^2} = \frac{\pi}{4}.$$

Нетрудно видеть, что оба эти определения удовлетворяют постулатам вероятности.

**Замечание.** Выбор системы равновероятных исходов не всегда очевиден. Даламбер считал в опыте с бросанием двух монет равновероятными три исхода: 1) на обеих монетах выпадают гербы, 2) на обеих монетах выпадают решки, 3) на монетах выпадают разные стороны. При этом каждому исходу следует приписать вероятность  $1/3$ . Другой способ рассуждений, представленный таблицей 1, приводит к четырём исходам, и, в предположении об их равных вероятностях, к вероятностям  $1/4$ :

Исход	1-я монета	2-я монета
1	Г	Г
2	Г	Р
3	Р	Г
4	Р	Р

Для окончательного выбора необходимые дополнительные соображения (см. ).

**2. Теоретическое определение вероятностей.** Когда случайность является специфическим свойством процесса, описываемого математическим уравнением, соответствующая вероятность может найдена из решения этого уравнения. Так,

решив уравнение Шредингера для атома водорода и найдя волновую функцию электрона, мы можем определить вероятность нахождения электрона в любой области атома интегрированием по этой области квадрата модуля волновой функции. Прямую вероятностную трактовку допускают уравнение диффузии, кинетические уравнения Лиувилля, Больцмана, Паули и др.

**3. Экспериментальное определение вероятностей.** Если организовать эксперимент и его повторение несложно, вероятность можно приближённо определить (оценить) эмпирическим путем. Пусть  $A_k$ ,  $k = 1, 2, \dots, n$  – полная группа событий,  $N \equiv N(\Omega)$  – число произведенных опытов, исходы которых  $\omega_1, \omega_2, \dots, \omega_N$  можно представить в виде отдельных точек в пространстве  $\Omega$ . Пусть в  $A_1$  попало  $N(A_1)$  точек, в  $A_2$  –  $N(A_2)$  точек, ... , в  $A_n$  –  $N(A_n)$  точек. Это означает, что в  $N(\Omega)$  экспериментах  $N(A_1)$  раз произошло событие  $A_1$ ,  $N(A_2)$  раз произошло событие  $A_2$ , ... ,  $N(A_n)$  раз произошло событие  $A_n$ . Поскольку никаких других событий, кроме перечисленных, произойти не может (по определению полной группы событий),

$$N(A_1) + N(A_2) + \dots + N(A_n) = N(\Omega).$$

Разделив обе части этого равенства на его правую часть, и введя обозначение

$$P_N(A_k) = \frac{N(A_k)}{N(\Omega)}, \quad k = 1, \dots, n,$$

получим:

$$P_N(A_1) + P_N(A_2) + \dots + P_N(A_n) = 1.$$

Нетрудно убедиться в том, что функция

$$P_N(A) = \frac{N(A)}{N(\Omega)}$$

удовлетворяет постулатам вероятности и может быть названа *эмпирической вероятностью* (распространённое название – *частота*). Ее отличие от "настоящей" вероятности в том, что она не является строго определённой функцией события: она зависит от  $N$ , а при повторении эксперимента может принять другое значение при том же самом  $N$  (это явление называется *статистическим разбросом*). И тем не менее, если  $N(A)$  достаточно велико, статистический разброс мал и эмпирическую вероятность можно использовать как приближённое значение "истинной" вероятности

$$P_N(A) \approx P(A).$$

Более того, можно показать также, что (в определённом смысле)

$$P_N(A) \rightarrow P(A), \quad N \rightarrow \infty.$$

Экспериментальное исследование эмпирической вероятности исходов опыта с бросанием двух монет недвусмысленно свидетельствует в пользу системы 4-х равновероятных исходов (рис. ).

Следует отметить, что обязательным условием для выполнения этого эксперимента является неизменность всех обстоятельств, от которых может зависеть его исход. Если с течением времени условия эксперимента изменяются, то изменяться может и сама вероятность. Для измерения такой вероятности надо



либо повторять весь цикл сначала, приводя систему в исходное состояние, либо построить (реально или мысленно) множество *копий* этой системы, разместив их в пространстве таким образом, чтобы исключить их друг на друга. Множество это называют *статистическим ансамблем*, оно и используется для определения вероятностей. Естественно *предположить*, что в случае неизменности внешних условий вероятности, определённые по ансамблю и по результатам повторных измерений одной копии ансамбля, должны совпадать. Предположение это называется **эргодической гипотезой**.

## 1.6 Приложение: Алгебра событий

Принадлежащая великому русскому математику А.Н.Колмогорову теоретико-множественная интерпретация случайных событий и построенная на её основе аксиоматика составляет фундамент современной теории вероятностей. Список взаимного соответствия терминов ("словарь") показан в таблице.

Символ	Теоретико-множественный смысл	Теоретико-вероятностный смысл
$\omega$	Элемент	Элементарное событие (исход)
$\Omega$	Множество всех элементов	Достоверное событие
$\emptyset$	Пустое множество	Невозможное событие
$\omega \in A$	Элемент $\omega$ принадлежит $A$	Произошло событие $A$
$A \cup B$	Объединение множеств	Произошло событие $A$ или $B$ или оба
$A'$	Дополнение к множеству $A$	Событие $A$ не произошло
$A \cap B$	Пересечение множеств	Произошли события $A$ и $B$
$A \cap B = \emptyset$	Множества $A$ и $B$ не пересекаются	События $A$ и $B$ несовместны

Операции над событиями образуют *алгебру*, то есть удовлетворяют следующей системе аксиом, которую удобно представить, введя обозначения  $A + B$  вместо  $A \cup B$  и  $A \cdot B$  вместо  $A \cap B$ .

1) **Аксиома замкнутости**: для каждой пары событий  $A$  и  $B$  существует единственное событие  $A + B$  и единственное событие  $A \cdot B$ .

2) **Аксиома коммутативности**:

$$A + B = B + A \quad (a)$$

и

$$A \cdot B = B \cdot A. \quad (b)$$

3) **Аксиома ассоциативности**:

$$(A + B) + C = A + (B + C) \quad (a)$$

и

$$(A \cdot B) \cdot C = A \cdot (B \cdot C). \quad (b)$$

4) **Аксиома дистрибутивности**:

$$(A + B) \cdot C = (A \cdot C) + (B \cdot C) \quad (a)$$

и

$$(A \cdot B) + C = (A + C) \cdot (B + C). \quad (b)$$

5) **Аксиома нуля и единицы**: существует единственное событие  $\emptyset$ , такое, что

$$A + \emptyset = A \quad (a)$$

для любого события  $A$  и существует единственное событие  $\Omega$ , такое, что

$$A \cdot \Omega = A \quad (b)$$

для любого события  $A$ .

6) **Аксиома дополненности:** для каждого события  $A$  существует единственное событие  $A'$ , такое, что  $A + A' = \Omega$  и  $A \cdot A' = \emptyset$ .

С помощью этих аксиом нетрудно доказать, а с помощью диаграмм Венна проверить справедливость следующих соотношений:

1.  $A + \Omega = A$ .
2.  $A \cdot \emptyset = \emptyset$ .
3.  $A + A' = \Omega$ .
4.  $A \cdot A' = \emptyset$ .
5.  $\Omega' = \emptyset$ .
6.  $\emptyset' = \Omega$ .
7.  $(A')' = A$ .
8.  $(A + B)' = A' \cdot B'$ .
9.  $(A \cdot B)' = A' + B'$ .

Два последних соотношения известны как *формулы де Моргана*.

Напомним, что доказать тождество  $A_1 = A_2$  означает доказать, что  $\omega \in A_1 \Rightarrow \omega \in A_2$  (из  $\omega \in A_1$  следует  $\omega \in A_2$ ) и наоборот,  $\omega \in A_2 \Rightarrow \omega \in A_1$ . Покажем, как это делается, на примере доказательства формулы 8. Пусть  $\omega \in (A + B)'$ , следовательно,  $\omega \notin (A + B)$ , то есть,  $\omega \notin A$  и в то же время  $\omega \notin B$ , другими словами  $\omega \in A'$  и одновременно  $\omega \in B'$ . Таким образом, мы показали, что  $\omega \in (A + B)' \Rightarrow \omega \in A' \cdot B'$ . Пусть теперь  $\omega \in A' \cdot B'$ . Это значит, что  $\omega \in A'$  и в то же время  $\omega \in B'$ , то есть  $\omega \notin A$  и  $\omega \notin B$ : элемент  $\omega$  не принадлежит множеству, содержащему элементы множества  $A$  и элементы множества  $B$ , что выражается формулой  $\omega \notin (A + B)$  или эквивалентной ей  $\omega \in (A + B)'$ . Тем самым, доказано обратное утверждение:  $\omega \in A' \cdot B' \Rightarrow \omega \in (A + B)'$ .

## 2 Лекция 2. Комбинаторика и вероятность

### 2.1 Принцип умножения

Сколько существует способов расставить 10 книг на книжной полке? Как можно разместить 10 шаров по трем ящикам? Сколько существует различных путей из точки  $(0, 0)$  в точку  $(6, 4)$  координатной плоскости, если делать только единичные шаги и только в положительных направлениях осей  $x$  или  $y$ ?. Задачи о числе различных последовательностей элементов множества, разбиений на классы, перестановок, сочетаний и других операций принадлежат специальному разделу математики – **комбинаторике**.

Общим методом решения комбинаторных задач является построение *дерева* возможных путей и применение к нему *принципа умножения*. Пусть необходимо выполнить одно за другим  $k$  действий, причем, первое из них можно выполнить  $m_1$  способами, второе –  $m_2$  способами, ...,  $k$ -е –  $m_k$  способами. **Принцип умножения гласит: последовательность  $k$  действий может быть выполнена**

$$m_{12...k} = m_1 m_2 \dots m_k$$

**различными способами.** Например, первый бочонок из мешочка для игры в лото мы можем вынуть 99-ю способами (именно столько бочонков в полном комплекте), а вот второй бочонок – 98-ю способами (один уже вынут). Общее число различных вариантов при этом

$$m_{12} = 99 \cdot 98 = 9702.$$

## 2.2 Две процедуры выбора

Множество различных объектов безотносительно к порядку их расположения будем называть **совокупностью** и обозначать  $\{a_1, a_2, a_3, \dots, a_m\}$ , исходное множество, содержащее все элементы, могущие участвовать в эксперименте, назовём **генеральной совокупностью**. Две совокупности считаются разными, если в одной из них содержится хотя бы один элемент, отсутствующий в другой. Символы  $\{a_1, a_2, a_3\}$  и  $\{a_1, a_3, a_2\}$  обозначают одну и ту же совокупность.

Выбрав последовательно, один за другим,  $k$  элементов генеральной совокупности и расположив их в порядке выбора, получим упорядоченную совокупность, которую будем называть **последовательностью** и обозначать через  $\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k} \rangle$ . Символы  $\langle a_1, a_2, a_3 \rangle$  и  $\langle a_1, a_3, a_2 \rangle$  обозначают разные последовательности. Числа элементов  $k$  и  $m$  будем называть **объемами** последовательности и совокупности соответственно.

Каждый из индексов  $j_n$ ,  $n = 1, 2, \dots, k$ , может принять любое целое значение от 1 до  $m$ , но ни одно из них не повторится в последовательности. Эта процедура называется **выбором без возвращения**.

Наряду с этой возможен и другой тип процедуры, когда номер выбранного из генеральной совокупности элемента записывается в протокол испытания, а сам элемент возвращается в генеральную совокупность и может участвовать в дальнейшем на равных правах с другими элементами. Этот тип называется **выбором с возвращением**.

Последовательность, полученную из исходной любым изменением порядка следования её элементов, называют её **перестановкой**. Число  $n(\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_m} \rangle | m)$  различных перестановок последовательности объёма  $m$  (включая исходную) обозначается через  $P_m$ :

$$n(\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_m} \rangle | m) = P_m.$$

Это число легко находится из следующих соображений. Каждая перестановка есть определённая последовательность, составленная из элементов совокупности  $(a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_m})$ . Рассмотрим эту процедуру. Первый элемент последовательности можно выбрать из исходной совокупности  $m_1 = m$  способами, второй выбирается из оставшейся части совокупности и может быть выбран  $m_2 = m - 1$  способами, третий может быть выбран  $m_3 = m - 2$  способами и т.д. В конце этой процедуры остаётся лишь один элемент, и её завершение может быть осуществлено единственным образом – присоединением этого оставшегося элемента к последовательности в качестве её последнего члена. Применяя правило умножения, получим

$$P_m = m(m - 1) \dots 1 \equiv m!. \quad (1)$$

Заметим, что при  $m = 1$  мы имеем дело с единственной совокупностью, состоящей из одного элемента, так что

$$1! = 1.$$

**Пример.** На книжной полке осталось три свободных места, а на столе – три книги одинаковой толщины с разными названиями  $A, B$  и  $C$ . Сколько существует способов разместить их на полке? Ответ даётся формулой (1):

$$P_3 = 3! = 6.$$

Полученные 6 последовательностей, называемых перестановками исходной (включая её саму) суть

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

## 2.3 Размещения

Видоизменим вышеприведённый пример.

**Пример.** На книжной полке осталось три свободных места, а на столе – *четыре* книги одинаковой толщины с разными названиями  $A, B, C$  и  $D$ . Сколько существует способов разместить на полке три книги из четырёх?

Теперь извлекаются не все элементы исходной совокупности, часть из них остаётся на столе. В общей постановке, из исходной совокупности объёмом  $m$  извлекается последовательность  $\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k} \rangle$ . Такие последовательности называются **размещениями** из  $m$  элементов по  $k$ , их число обозначается через  $A_m^k$  и находится по той же самой схеме рассуждений:

$$n(\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k} \rangle | m) = A_m^k = m(m-1) \dots (m-k+1) = \frac{m!}{(m-k)!}. \quad (2)$$

Заметим, что перестановки можно считать частным случаем размещений:

$$A_m^m = P_m.$$

Вернувшись к примеру, воспользуемся формулой (2):

$$A_4^3 = \frac{4!}{1!} = 24.$$

Полученные 24 размещения суть

$$\begin{array}{cccc} ABC & ABD & ACD & BCD \\ ACB & ADB & ADC & BDC \\ BAC & BAD & CAD & CBD \\ BCA & BDA & CDA & CDB \\ CAB & DAB & DAC & DBC \\ CBA & DBA & DCA & DCB. \end{array} \quad (3)$$

## 2.4 Сочетания

Заметим, что каждая колонка в приведенной выше таблице содержит последовательности, состоящие из одних и тех же книг, только расположенных в разном порядке. Если порядок нам безразличен (скажем, собираемся купить три книги из четырёх), то остаётся только 4 варианта:

$$ABC \quad ABD \quad ACD \quad BCD. \quad (4)$$

Подмножество объёмом  $k$  совокупности объёмом  $m$ , рассматриваемое без учёта порядка элементов, называется **сочетанием из  $n$  элементов по  $k$** , число различных сочетаний обозначается  $C_m^k$ . В приведённом выше примере имеем  $P_4^3 = 24$  размещений и  $C_4^3 = 4$  сочетаний. Очевидно, число размещений больше (точнее, не

меньше) числа сочетаний с теми же индексами, каждое сочетание порождает  $k!$  перестановок:

$$n(\langle a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k} \rangle | m) = k! n(\{a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k}\} | m).$$

В результате имеем:

$$C_m^k = n(\{a_{j_1}, a_{j_2}, a_{j_3}, \dots, a_{j_k}\} | m) = \frac{A_m^k}{k!} \equiv \binom{m}{k}.$$

Число

$$\binom{m}{k} \equiv \frac{m!}{k!(m-k)!},$$

называется **биномиальным коэффициентом**.

Биномиальные коэффициенты  $\binom{n}{k}$  удовлетворяют правилу Ньютона

$$\sum_{k=0}^n \binom{n}{k} a^k b^{n-k} = (a+b)^n,$$

составляют треугольник Паскаля

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}, \quad 1 \leq k \leq n,$$

обладают симметрией

$$\binom{n}{k} = \binom{n}{n-k}$$

и принимают наибольшее значение при  $k$ , равном  $n/2$  или ближайшему к  $n/2$  целому числу.

Проиллюстрируем эти понятия на примере игры в лото. Полный набор бочонков в мешочке лото образует генеральную совокупность объёмом 99; после вынимания 19-ти бочонков там останется 80. Если мы не имеем информации о том, какие именно номера были вынуты, мы не знаем и какие остались, но мы можем утверждать, что возможно  $C_{99}^{80}$  сочетаний оставшихся элементов. В мешочке находится только одно такое сочетание. Открыв мешочек, мы можем узнать, какое именно сочетание (т.е., из каких элементов оно состоит). Расположив теперь эти бочонки вдоль линейки, положенной на стол, получим последовательность – размещение. Это можно сделать  $k! = 80!$  способами, поэтому каждое сочетание объёмом  $k$  порождает  $k!$  размещений.

## 2.5 Разбиения

Вернёмся к лото и поместим выбранные  $m_1 = 19$  бочонков в другой мешочек. Теперь у нас два мешочка, пометим новый номером 1. Продолжим эксперимент: вынем из исходного мешочка  $m_2$  бочонков и поместим их в новый мешочек с номером 2. Будем повторять эту операцию до тех пор, пока у нас не появится новый мешочек с номером  $k-1$ . Тут мы и остановимся, наклеив на исходный мешочек номер  $k$ . Мы получили **разбиение совокупности  $m$  элементов по  $k$  ячейкам**, такое, что в первой ячейке оказалось  $m_1$  элементов, во второй –  $m_2$ , ... , в  $k$ -й –  $m_k$  элементов. Числа  $m_1, \dots, m_k$ ,

называемые **числами заполнения**, могут принимать любые значения, от 0 до  $m = 99$ , лишь бы их сумма была равна числу всех бочонков  $m$ ,  $m_1 + m_2 + \dots + m_k = m$ . Обозначим через  $C_m^{\langle m_1, \dots, m_k \rangle}$  число разбиений с фиксированной последовательностью чисел заполнения  $\langle m_1, m_2, \dots, m_k \rangle$ . Заметим, что полученное выше число сочетаний  $C_m^k$  соответствует частному случаю числа разбиений  $C_m^{\langle m_1, m_2 \rangle}$ , где  $m_1 = k$  и  $m_2 = m - k$ , так что

$$C_m^{\langle m_1, m_2 \rangle} = \binom{m}{m_1} = \frac{m!}{m_1! m_2!}, \quad m_1 + m_2 = m.$$

Применяя принцип умножения, получим

$$\begin{aligned} C_m^{\langle m_1, m_2, m_3 \rangle} &= \binom{m}{m_1} \binom{m - m_1}{m_2} = \frac{m}{m_1! (m - m_1)!} \frac{(m - m_1)!}{m_2! (m - m_1 - m_2)!} = \\ &= \frac{m!}{m_1! m_2! (m - m_1 - m_2)!}, \\ &\dots \\ C_m^{\langle m_1, \dots, m_k \rangle} &= \binom{m}{m_1} \binom{m - m_1}{m_2} \dots \binom{m - m_1 - \dots - m_{k-2}}{m_{k-1}} = \\ &= \frac{m!}{m_1! \dots m_{k-1}! (m - m_1 - \dots - m_{k-1})!}. \end{aligned}$$

Поскольку

$$m_1 + m_1 + \dots + m_{k-1} + m_k = m,$$

окончательно получаем

$$C_m^{\langle m_1, \dots, m_k \rangle} \equiv \binom{m}{m_1, \dots, m_k}.$$

Число

$$\binom{m}{m_1, \dots, m_k} \equiv \frac{m!}{m_1! \dots m_k!}$$

называется **полиномиальным коэффициентом**.

**Пример.** Рассмотрим разбиение совокупности трёх элементов  $\{a, b, c\}$  по три:

$C_3^{(3,0,0)} = 1$	$\{abc  \quad   \quad \}$	$\{\circ \circ \circ   \quad   \quad \}$
$C_3^{(0,3,0)} = 1$	$\{ \quad  abc  \quad \}$	$\{ \quad   \circ \circ \circ   \quad \}$
$C_3^{(0,0,3)} = 1$	$\{ \quad   \quad  abc\}$	$\{ \quad   \quad   \circ \circ \circ\}$
$C_3^{(2,1,0)} = 3$	$\{ab c  \quad \} \quad \{ac b  \quad \} \quad \{bc a  \quad \}$	$\{\circ \circ   \circ   \quad \}$
$C_3^{(2,0,1)} = 3$	$\{ab  \quad  c\} \quad \{ac  \quad  b\} \quad \{bc  \quad  a\}$	$\{\circ \circ   \quad   \circ\}$
$C_3^{(1,2,0)} = 3$	$\{a bc  \quad \} \quad \{b ac  \quad \} \quad \{c ab  \quad \}$	$\{\circ   \circ \circ   \quad \}$
$C_3^{(1,0,2)} = 3$	$\{a  \quad  bc\} \quad \{b  \quad  ac\} \quad \{c  \quad  ab\}$	$\{\circ   \quad   \circ \circ\}$
$C_3^{(0,2,1)} = 3$	$\{ \quad  ab c\} \quad \{ \quad  ac b\} \quad \{ \quad  bc a\}$	$\{ \quad   \circ \circ   \circ\}$
$C_3^{(0,1,2)} = 3$	$\{ \quad  a bc\} \quad \{ \quad  b ac\} \quad \{ \quad  c ab\}$	$\{ \quad   \circ   \circ \circ\}$
$C_3^{(1,1,1)} = 6$	$\{a b c  \quad \} \quad \{a c b  \quad \} \quad \{b a c  \quad \} \quad \{b c a  \quad \} \quad \{c a b  \quad \} \quad \{c b a  \quad \}$	$\{\circ   \circ   \circ   \quad \}$

В первой колонке приведены числа разбиений совокупности различных элементов, в центральной части таблицы показаны сами разбиения, в правой же колонке приведены разбиения совокупности *неразличимых элементов*.

Из представленной выше таблицы видно, что если стереть символы  $a, b, c$  на шарах, то полное число возможных разбиений сократится: вместо

$$\sum_{i+j+k=3} C_3^{(i,j,k)} = 27$$

получим всего 10. Эти десять разбиений и показаны в крайней правой колонке таблицы.

Существует простой способ вывода формулы для полного числа разбиений множества  $m$  неразличимых элементов ("шаров") по  $k$  подмножествам ("ящикам"). Расположим ящики последовательно друг за другом, так что соседние ящики имеют общую перегородку. Уберём теперь крайние перегородки (они ничего не разделяют), останется  $k - 1$  перегородок. Шары в ящиках расположим линейно, раздвинув перегородки, если это необходимо (мысленно мы всегда можем это сделать). В результате получим последовательность  $m + k - 1$  объектов, состоящую из перемешанных между собой  $m - 1$  перегородок и  $k$  шаров:

$$\circ \circ \circ \mid \circ \circ \mid \mid \circ \circ \circ \circ \mid \mid \mid \circ \quad .$$

В терминах чисел заполнения она имеет вид: ;

$$\langle 3|2|0|4|0|0|1 \rangle$$

Если бы объекты были различимы, из этой совокупности можно было бы получить  $(m + k - 1)!$  последовательностей. На самом деле, перестановки шаров (числом  $k!$ ) и перестановки перегородок (числом  $(m - 1)!$ ) не приводят к новым разбиениям, и число разбиений  $A_{m,k}$  оказывается в  $k!(m - 1)!$  раз меньше:

$$A_{m,k} = \frac{(m + k - 1)!}{k!(m - 1)!} = \binom{m + k - 1}{k}.$$

В статистической физике эта модель находит полезное применение при подсчете числа возможных распределений частиц ("шаров") по квантовым состояниям ("ящикам"). Все частицы делятся на два класса: бозоны (частицы, подчиняющиеся статистике Бозе-Эйнштейна: фотоны, альфа-частицы, и вообще, частицы, ядра и атомы с целым спином) и фермионы (частицы, подчиняющиеся статистике Ферми-Дирака: электроны, протоны и вообще, частицы с полуцелым спином). Бозоны могут занимать одно и то же состояние в неограниченном числе, приведённая выше формула справедлива именно для этого случая. Фермионы же не могут находиться в одном и том же состоянии: состояние может быть либо пустым, либо заполненным только одним фермионом. В этом случае обязательно выполнение условия  $k \leq m$ . Разбиение полностью описывается указанием занятых состояний:

$$\langle 0|0|1|1|0|1|0|0|0 \rangle.$$

Их число равно  $k$ , а полное число состояний  $m$ , так что число возможных разбиений в этом случае

$$A'_{m,k} = \binom{m}{k}.$$

## 2.6 Перестановки с повторяющимися элементами

Пусть теперь исходная совокупность содержит повторяющиеся элементы, так что число элементов 1-го типа равно  $m_1$ , второго –  $m_2$ , ..., последнего ( $k$ -го) типа  $m_k$ . Расположив их вдоль линейки и перенумеровав, получим последовательность

$$a_1, \dots, a_1, a_2, \dots, a_2, \dots, a_k, \dots, a_k.$$

Переставляя элементы этой последовательности, мы могли бы получить  $m!$  перестановок, но многие из них будут совпадать друг с другом, окажутся неразличимыми: перестановки элементов одного и того же типа не приводят к новым последовательностям. Суммарное число таких перестановок, по принципу умножения, равно  $m_1!m_2!\dots m_k!$ , так что  $m! = m_1!m_2!\dots m_k! \times \text{число различных перестановок}$ . Отсюда следует, что число различных перестановок в схеме с повторяющимися элементами выражается той же формулой, что и число разбиений:

$$C_m^{(m_1, \dots, m_k)} = \frac{m!}{m_1! \dots m_k!} \equiv \binom{m}{m_1, \dots, m_k}.$$

## 2.7 Выбор с возвращением

Рассмотренный способ выбора последовательности из генеральной совокупности называется **выбором без возвращения**.

Другой результат получается в случае **выбора с возвращением**, когда номер выбранного элемента заносится в протокол испытания, а сам элемент возвращается в генеральную совокупность и может быть выбран повторно. В этом случае число вариантов выбора оказывается одним и тем же на каждом шаге, и мы имеем

$$n'(\langle a_{j_1}, a_{j_2}, \dots, a_{j_k} \rangle | m) = m m \dots m = m^k. \quad (3)$$

Но теперь элементы последовательности могут повторяться и размер её может быть сколь угодно большим. Сравните множество последовательностей, образуемых из одной и той же совокупности  $\{a, b, c\}$  двумя разными способами выбора.

**Выбор без возвращения** ( $n(\langle a_{j_1}, a_{j_2}, a_{j_3} \rangle | 3) = 3! = 6$ ):

$$\langle a, b, c \rangle \quad \langle a, c, b \rangle \quad \langle b, a, c \rangle \quad \langle b, c, a \rangle \quad \langle c, a, b \rangle \quad \langle c, b, a \rangle$$

**Выбор с возвращением** ( $n'(\langle a_{j_1}, a_{j_2}, a_{j_3} \rangle | 3) = 3^3 = 27$ ):

$$\begin{array}{cccccccccc} \langle a, a, a \rangle & \langle a, a, b \rangle & \langle a, b, a \rangle & \langle a, a, c \rangle & \langle a, c, a \rangle & \langle a, b, c \rangle & \langle a, c, b \rangle & \langle a, b, b \rangle & \langle a, c, c \rangle & \\ \langle b, b, b \rangle & \langle b, b, c \rangle & \langle b, c, b \rangle & \langle b, b, a \rangle & \langle b, a, b \rangle & \langle b, c, a \rangle & \langle b, a, c \rangle & \langle b, c, c \rangle & \langle b, a, a \rangle & \\ \langle c, c, c \rangle & \langle c, c, a \rangle & \langle c, a, c \rangle & \langle c, c, b \rangle & \langle c, b, c \rangle & \langle c, a, b \rangle & \langle c, b, a \rangle & \langle c, a, a \rangle & \langle c, b, b \rangle & \end{array}$$

## 2.8 Комбинаторные вероятности

Пользуясь предположением о равной вероятности выбора каждого элемента, вычислим связанные с описанными выше схемами вероятности.



Вероятность того, что при случайном выборе из совокупности  $m$  различных элементов будет выбран заранее названный элемент  $a_1$ , равна

$$P(a = a_1) = \frac{n(a_1)}{n(a_j)} = \frac{1}{m}.$$

Вероятность того, что при случайном выборе без возвращения  $k$  элементов из совокупности  $m$  различных элементов будет выбрана заранее указанная последовательность  $\langle a_1, a_2, \dots, a_k \rangle$ , равна

$$P(\langle a_{j_1}, a_{j_2}, \dots, a_{j_k} \rangle = \langle a_1, a_2, \dots, a_k \rangle) = \frac{n(\langle a_1, a_2, \dots, a_k \rangle | m)}{n(\langle a_{j_1}, a_{j_2}, \dots, a_{j_k} \rangle | m)} = \frac{1}{A_m^k} = \frac{(m-k)!}{k!}.$$

При  $k = m$  она равна  $m!^{-1}$ .

Вероятность того, что при случайном выборе без возвращения  $k$  элементов из совокупности  $m$  различных элементов будет выбрана заранее указанная совокупность  $\{a_1, a_2, \dots, a_k\}$ , равна

$$\begin{aligned} P(\{a_{j_1}, a_{j_2}, \dots, a_{j_k}\} = \{a_1, a_2, \dots, a_k\}) &= \frac{n(\{a_1, a_2, \dots, a_k\} | m)}{n(\{a_{j_1}, a_{j_2}, \dots, a_{j_k}\} | m)} = \\ &= \frac{n(\{a_1, a_2, \dots, a_k\} | m) k!}{n(\langle a_{j_1}, a_{j_2}, \dots, a_{j_k} \rangle | m)} = \frac{k!}{A_m^k} = \frac{k!(m-k)!}{m!} = \binom{m}{k}^{-1}. \end{aligned}$$

При  $k = m$  она равна 1.

Вероятность того, что при случайном выборе *с возвращением*  $k$  элементов из совокупности  $m$  различных элементов будет выбрана совокупность попарно несовпадающих элементов, равна

$$P(a_{j_1} \neq a_{j_2} \neq \dots \neq a_{j_k}) = \frac{A_m^k}{m^k} = \frac{m!}{m^k(m-k)!}.$$

Здесь принято во внимание, что последовательностей с  $k$  неповторяющимися элементами здесь столько же, что и в случае выбора без возвращения ( $A_m^k$ ), а число всех последовательностей больше:  $m^k$ .

В урне находится  $m_1$  черных шаров и  $m_2$  белых ( $m_1 + m_2 = m$ ). Из неё наугад извлекается без возвращения  $k$  шаров. Найдём вероятность  $P(N_{\text{ч}} = k_1)$  того, что среди извлеченных шаров чёрных ровно  $k_1$ . В выборке, содержащей  $k_1$  черных шаров и  $k_2 = k - k_1$  белых, черные шары могут быть выбраны  $\binom{m_1}{k_1}$  различными способами, а белые —  $\binom{m_2}{k_2}$ . Так как любой способ выбора  $k_1$  черных шаров может комбинироваться с любым способом выбора  $k_2$  белых, то число таких комбинаций по принципу умножения равно произведению

$$\binom{m_1}{k_1} \binom{m_2}{k_2},$$

а искомая вероятность получается делением его на полное число выборок:

$$P(N_{\text{ч}} = k_1) = \binom{m_1}{k_1} \binom{m_2}{k_2} / \binom{m}{k}.$$

Это распределение вероятностей называется **гипергеометрическим**.

## 2.9 Спортлото

Участники лотереи из 49 наименований видов спорта (обозначенных просто цифрами) называют 6. Выигрыш определяется тем, сколько из них совпадет с шестью наименованиями, заранее выделенными комиссией. Назовем их черными шарами, остальные белыми. Вероятность угадать все шесть:  $k_1 = 6, k = 6, n_1 = 6, n = 49$

$$P = \frac{1}{\binom{49}{6}} = \frac{6!43!}{49!} \approx 7.2 * 10^{-8}$$

по формуле Стирлинга

$$n! \sim \sqrt{2\pi n} n^n e^{-n}$$

## 2.10 Математическое дополнение: Биномиальные коэффициенты. Формула Стирлинга

Рассмотрим интеграл

$$\int_1^n \ln x dx$$

возьмем этот интеграл по частям:

$$x \ln x \Big|_1^n - \int_1^n dx = n \ln n - n + 1$$

По формуле ???:

$$\approx \frac{\ln 1 + \ln n}{2} + \ln 2 + \dots + \ln(n-1) = \ln n! - \frac{1}{2} \ln n$$

Таким образом, получаем равенство:

$$n \ln n - n + 1 \approx \ln n! - \frac{1}{2} \ln n$$

пренебрежем единицей:

$$\ln n! \approx \left(n + \frac{1}{2}\right) \ln n - n$$

$$n! \approx n^{n+1/2} e^{-n}$$

Более точная формула:

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

И еще более точная:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \left(1 + \frac{1}{12n}\right)$$

## 3 Формула Стирлинга.

$$\int_1^n \ln x dx = x \ln x \Big|_1^n - \int_1^n dx = n \ln n - n + 1$$

$$\int_1^n \ln x dx \approx \text{по формуле трапеций} \approx \frac{\ln 1 + \ln 2}{2} + \ln 2 + \dots + \ln(n-1) = \ln 1 * 2 * \dots * n - \frac{1}{2} \ln n =$$

$$= \ln n! - \frac{1}{2} \ln n$$

приравниваем:

$$\ln n! \approx n \ln n - n + 1 + \frac{1}{2} \ln n = \left(n + \frac{1}{2}\right) \ln n - n$$

$$n! \approx n^{n+1/2} e^{-n}$$

или более точно,

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \quad \text{формула Стирлинга.}$$

А если еще точнее, то

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \left[1 + \frac{1}{12n}\right]$$

## 4 Примеры решения задач

**Задача 3.2.** Дано:  $\Omega = \{a, b, c, d, e\}$ ,  $A = \{b, c, d, e\}$ ,  $B = \{c, d, e\}$ ,  $C = \{d, e\}$ . Найти  $A + B$ ,  $A \cdot B$ ,  $A'$  и  $A + B \cdot C$ .

**Решение:**  $A + B = \{b, c, d, e\}$ ,  $A \cdot B = \{c, d, e\}$ ,  $A' = \{a\}$  and  $A + B \cdot C = \{b, c, d, e\}$ .

**Задача 3.3.** Дано:  $\Omega = \{x : 0 \leq x \leq 5\}$ ,  $A = \{x : 0 \leq x \leq 4\}$ ,  $B = \{x : 1 \leq x \leq 3\}$ ,  $C = \{x : 1 < x < 3\}$ ,  $D = \{x : 2 \leq x \leq 5\}$ . Найти  $A'$ ,  $C'$ ,  $A + B$ ,  $A \cdot B$ ,  $A \cdot C$ ,  $A \cdot D$ .

**Решение:**  $A' = \{x : 4 < x \leq 5\}$ ,  $C' = \{x : 0 \leq x \leq 1\} + \{x : 3 \leq x \leq 5\}$ ,  $A + B = \{x : 0 \leq x \leq 4\} = A$ ,  $A \cdot B = \{x : 1 \leq x \leq 3\} = B$ ,  $A \cdot C = \{x : 1 < x < 3\} = C$ ,  $A \cdot D = \{x : 2 \leq x \leq 4\}$ .

**Задача 3.4.** Монета бросается три раза. Определить пространство элементарных событий и найти вероятности, что

- (а) выпадут три герба,
- (б) выпадет один герб,
- (в) выпадут два герба,
- (г) первым выпадет герб, второй – решка,
- (д) первым выпадет герб,
- (е) два первых раза выпадет герб,
- (ж) по меньшей мере два раза подряд выпадет герб,
- (з) не более двух раз выпадет герб,
- (и) не более двух раз подряд выпадет герб,
- (к) не менее двух раз подряд выпадет герб,
- (л) хотя бы раз выпадет герб,
- (м) два первых бросания дадут одинаковый результат,
- (н) два первых бросания дадут одинаковый результат, отличный от последнего,
- (о) два бросания дадут одинаковый результат, отличный от третьего,
- (п) два бросания из трех дадут разные результаты,
- (р) не менее двух бросаний из трех дадут одинаковые результаты.

**Решение:** Пространство элементарных событий включает в себя  $2^3 = 8$  следующих исходов:

ГГГ, ГГР, ГРГ, РГГ, ГРР, РГР, РРГ, РРР.

Отмечая исходы, относящиеся к рассматриваемому событию  $A$ , полужирным шрифтом, найдем вероятности по формуле  $P(A) = n(A)/n(\Omega) = n(A)/8$ :

- (а) ГГГ, ГГР, ГРГ, РГГ, ГРР, РГР, РРГ, РРР;  $P(A) = 1/8$ .
- (б) ГГГ, ГГР, ГРГ, РГГ, **ГРР**, **РГР**, **РРГ**, РРР;  $P(A) = 3/8$ .
- (в) ГГГ, **ГГР**, **ГРГ**, **РГГ**, ГРР, РГР, РРГ, РРР;  $P(A) = 3/8$ .
- (г) ГГГ, ГГР, **ГРГ**, РГГ, **ГРР**, РГР, РРГ, РРР;  $P(A) = 1/4$ .
- (д) **ГГГ**, **ГГР**, **ГРГ**, РГГ, **ГРР**, РГР, РРГ, РРР;  $P(A) = 1/2$ .
- (е) **ГГГ**, **ГГР**, ГРГ, РГГ, ГРР, РГР, РРГ, РРР;  $P(A) = 1/4$ .
- (ж) **ГГГ**, **ГГР**, ГРГ, **РГГ**, ГРР, РГР, РРГ, РРР;  $P(A) = 3/8$ .
- (з) ГГГ, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, **РРР**;  $P(A) = 7/8$ .
- (и) ГГГ, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, **РРР**;  $P(A) = 7/8$ .
- (к) **ГГГ**, **ГГР**, ГРГ, **РГГ**, ГРР, РГР, РРГ, РРР;  $P(A) = 3/8$ .
- (л) **ГГГ**, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, РРР;  $P(A) = 7/8$ .
- (м) **ГГГ**, **ГГР**, ГРГ, РГГ, ГРР, РГР, **РРГ**, **РРР**;  $P(A) = 1/2$ .
- (н) ГГГ, **ГГР**, ГРГ, РГГ, ГРР, РГР, **РРГ**, РРР;  $P(A) = 1/4$ .
- (о) ГГГ, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, РРР;  $P(A) = 3/4$ .
- (п) ГГГ, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, РРР;  $P(A) = 3/4$ .
- (р) **ГГГ**, **ГГР**, **ГРГ**, **РГГ**, **ГРР**, **РГР**, **РРГ**, **РРР**;  $P(A) = 1$ .

**Задача 3.5** Дважды бросается игральный кубик и результат записывается в виде пары чисел  $(X, Y)$ . Найти вероятности событий (а)  $X + Y = 8$ , (б)  $X - Y = 2$ , (в)  $|X - Y| = 2$ , (г)  $X < Y$ , (д)  $X \geq Y$ , (е)  $X = Y$ , (ж)  $X \neq Y$ .

**Решение:** Пространство элементарных событий включает в себя  $6^2 = 36$  следующих исходов:

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

Отмечая исходы, относящиеся к рассматриваемому событию  $A$ , полужирным шрифтом, найдем вероятности по формуле  $P(A) = n(A)/n(\Omega) = n(A)/36$ :

(а)

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	<b>(6, 2)</b>
(1, 3)	(2, 3)	(3, 3)	(4, 3)	<b>(5, 3)</b>	(6, 3)
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	(5, 4)	(6, 4)
(1, 5)	(2, 5)	<b>(3, 5)</b>	(4, 5)	(5, 5)	(6, 5)
(1, 6)	<b>(2, 6)</b>	(3, 6)	(4, 6)	(5, 6)	(6, 6)

$$P(X + Y = 8) = \frac{5}{36}.$$

(6)

(1, 1)	(2, 1)	<b>(3, 1)</b>	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	<b>(4, 2)</b>	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	<b>(5, 3)</b>	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	<b>(6, 4)</b>
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

$$P(X - Y = 2) = \frac{1}{9}.$$

(B)

(1, 1)	(2, 1)	<b>(3, 1)</b>	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	<b>(4, 2)</b>	(5, 2)	(6, 2)
<b>(1, 3)</b>	(2, 3)	(3, 3)	(4, 3)	<b>(5, 3)</b>	(6, 3)
(1, 4)	<b>(2, 4)</b>	(3, 4)	(4, 4)	(5, 4)	<b>(6, 4)</b>
(1, 5)	(2, 5)	<b>(3, 5)</b>	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	<b>(4, 6)</b>	(5, 6)	(6, 6)

$$P(|X - Y| = 2) = \frac{2}{9}.$$

(r)

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
<b>(1, 2)</b>	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
<b>(1, 3)</b>	<b>(2, 3)</b>	(3, 3)	(4, 3)	(5, 3)	(6, 3)
<b>(1, 4)</b>	<b>(2, 4)</b>	<b>(3, 4)</b>	(4, 4)	(5, 4)	(6, 4)
<b>(1, 5)</b>	<b>(2, 5)</b>	<b>(3, 5)</b>	<b>(4, 5)</b>	(5, 5)	(6, 5)
<b>(1, 6)</b>	<b>(2, 6)</b>	<b>(3, 6)</b>	<b>(4, 6)</b>	<b>(5, 6)</b>	(6, 6)

$$P(X < Y) = \frac{5}{12}.$$

(d)

<b>(1, 1)</b>	<b>(2, 1)</b>	<b>(3, 1)</b>	<b>(4, 1)</b>	<b>(5, 1)</b>	<b>(6, 1)</b>
(1, 2)	<b>(2, 2)</b>	<b>(3, 2)</b>	<b>(4, 2)</b>	<b>(5, 2)</b>	<b>(6, 2)</b>
(1, 3)	(2, 3)	<b>(3, 3)</b>	<b>(4, 3)</b>	<b>(5, 3)</b>	<b>(6, 3)</b>
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	<b>(5, 4)</b>	<b>(6, 4)</b>
(1, 5)	(2, 5)	(3, 5)	(4, 5)	<b>(5, 5)</b>	<b>(6, 5)</b>
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	<b>(6, 6)</b>

$$P(X \geq Y) = \frac{7}{12}$$

(e)

<b>(1, 1)</b>	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	<b>(2, 2)</b>	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	<b>(3, 3)</b>	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	<b>(5, 5)</b>	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	<b>(6, 6)</b>

$$P(X = Y) = \frac{1}{6}$$

(ж)

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

$$P(X \neq Y) = \frac{5}{6}$$

**Задача 3.6.** (а) Ящик содержит карточки с буквами *а, к, н, о*. Какова вероятность того, что при случайном последовательном выборе карточек получится слово *окна*?

(б) Ящик содержит карточки с буквами *к, н, о, о*. Какова вероятность того, что при случайном последовательном выборе карточек получится слово *окно*?

(в) Ящик содержит карточки с буквами *а, а, а, а, а, б, б, д, к, р, р*. Какова вероятность того, что при случайном последовательном выборе карточек получится слово *абракадабра*?

**Решение.** (а) Пространство элементарных событий состоит из следующих элементов:

*акно акон анко анок аокн аонк*  
*кано каон кнао кноа коан кона*  
*нако наок нкао нкоа ноак нока*  
***окна*** *окан онка онак оакн оанк*

Отсюда вероятность

$$P(\text{окна}) = \frac{1}{24}.$$

Заметим, что элементы каждой строки построены по схеме:

1234 1243 1324 1342 1423 1432

(б) Пространство элементарных событий состоит из следующих элементов (мы просто заменили в предыдущем примере *а* на *о*):

***окно*** *окон онко онок оокн оонк*  
*коно коон кноо кноо коон коно*  
*ноко ноок нкоо нкоо ноок ноко*  
***окно*** *окон онко онок оокн оонк*

Отсюда вероятность

$$P(\text{окно}) = \frac{2}{24} = \frac{1}{12}.$$

Не имея под руками предыдущего примера, мы могли бы сначала пронумеровать совпадающие буквы, то есть, рассмотреть совокупность *к, н, о<sub>1</sub>, о<sub>2</sub>*, а после построения пространства элементарных событий индексы убрать.

(в) Число букв здесь слишком велико, чтобы работать с полным списком элементов. Заметим, однако, что если мы пронумеруем совпадающие буквы, то в результате всех перестановок мы получим

$$n(\Omega) = 11!$$

исходов, из которых

$$n(\text{абракадабра}) = 5!2!2!$$

дадут искомое слово (после снятия индексов). Таким образом,

$$P(\text{абракадабра}) = \frac{n(\text{абракадабра})}{n(\Omega)} = \frac{5!2!2!}{11!} = \frac{1}{83\,160}.$$

**Задача 3.7.** Дано  $P(A) = 2/3$ ,  $P(A \cdot B') = 1/3$ . Найти  $P(A \cdot B)$ .

**Решение:**

$$A = A \cdot \Omega, \quad \Omega = B + B', \quad A = A \cdot (B + B') = (A \cdot B) + (A \cdot B'),$$

откуда

$$P(A) = P((A \cdot B) + (A \cdot B')) = P(A \cdot B) + P(A \cdot B'),$$

поскольку  $A \cdot B$  and  $A \cdot B'$  несовместные события. Отсюда

$$P(A \cdot B) = P(A) - P(A \cdot B') = 2/3 - 1/3 = 1/3.$$

### ЗАДАНИЕ 1

1. Пусть  $\Omega = \{0, 1, 2, 3, 4, 5\}$ ,  $A = \{1, 3, 5\}$ ,  $B = \{0, 2, 4\}$ ,  $C = \{2, 3, 4\}$ ,  $D = \{1, 4, 5\}$ .  
Определить события:

$$(a) A + B; \quad (b) A \cdot B; \quad (c) C'; \quad (d) (C' \cdot D) + B; \quad (e) (\Omega \cdot C)'; \quad (f) A \cdot C \cdot D'.$$

2. Дано  $\Omega = \{x : 0 \leq x \leq 3\}$ ,  $A = \{x : 0 \leq x \leq 2\}$ ,  $B = \{x : 1 \leq x \leq 3\}$ . Найти

$$(a) A'; \quad (b) B'; \quad (c) A + B; \quad (d) A \cdot B; \quad (e) A \cdot B'; \quad (f) A' \cdot B'.$$

3. Правильная монета бросается пять раз. Найти вероятности событий

$$(a) A : \text{все испытания дадут } \Gamma;$$

$$(b) B : \text{одно из испытаний даст } \Gamma, \text{ остальные дадут } \Gamma;$$

$$(c) C : \text{два из испытаний дадут } \Gamma, \text{ остальные - } \Gamma;$$

$$(d) D : \text{первое испытание даст } \Gamma, \text{ следующее - } \Gamma;$$

$$(e) E : \text{первые три испытания дадут } \Gamma.$$

4. Бросается пара игральных кубиков (красный и зелёный) и результат записывается как  $X$  (для красного кубика) и  $Y$  (для зелёного). Найти вероятности событий:

$$A_k : X + Y = k, \quad k = 2, 3, \dots, 12.$$

5. Бросается пара игральных кубиков (красный и зелёный) и результат записывается как  $X$  (для красного кубика) и  $Y$  (для зелёного). Найти вероятности событий:

$$(a) A : X - Y = 3; \quad (b) B : |X - Y| = 3; \quad (c) C : X < Y; \quad (d) D : X \leq Y;$$

$$(e) E : X + Y < 6; \quad (f) F : X > 3; \quad (g) G : Y^2 = X.$$

6. Даны вероятности  $P(A \cap B') = 0.3$ ,  $P(A' \cap B) = 0.2$ ,  $P((A \cap B)') = 0.8$ . Найти:

$$(1) P(A \cap B); \quad (2) P(A); \quad (3) P(B); \quad (4) P(A \cup B).$$



## 5 Лекция 4. Условная вероятность и независимость

### 5.1 Условная вероятность

Рассмотрим урну  $S$ , содержащую 3 красных (red) и 7 белых (white) шаров. Один из шаров выбирается наугад и вероятность, что это будет красный равна

$$P(R) = \frac{\text{число красных шаров в } S}{\text{число всех шаров в } S} = \frac{n(R)}{n(S)} = \frac{3}{3+7} = \frac{3}{10}.$$

Поместим в урну перегородку, делящую её на две части:  $S = B_1 \cup B_2$ .  $B_1 \cap B_2 = \emptyset$ ,  $n_1 = n_2 = 5$ ,  $n_1(\text{red}) = 1$ ,  $n_2(\text{red}) = 2$ . Вероятность вытащить красный шар, *выбирая его из первой урны* равна

$$P(R|B_1) = \frac{\text{число красных шаров в } B_1}{\text{число всех шаров в } B_1} = \frac{n(R \in B_1)}{n(B_1)} = \frac{1}{5}.$$

Эта вероятность может быть представлена в виде

$$P(R|B_1) = \frac{n(R \cap B_1)}{n(B_1)} = \frac{n(R \cap B_1)}{n(S)} \frac{n(S)}{n(B_1)} = \frac{P(R \cap B_1)}{P(B_1)}.$$

**Определение.** Пусть  $B$  - любое событие, такое, что  $P(B) > 0$ . Тогда *условная вероятность* наступления события  $A$  при условии, что наступило событие  $B$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Условная вероятность  $P(A|B)$  может быть меньше чем  $P(A)$ , больше чем  $P(A)$  или быть равной этой вероятности. Событие  $B$  называется *условным событием*.

Условная вероятность обладает всеми свойствами вероятности:  $0 \leq P(A|B) \leq 1$ ,  $P(\emptyset|B) = 0$ ,  $P(S|B) = 1$ , и одним особым свойством  $P(B|B) = 1$ .

### 5.2 Правила умножения

Умножая формулу для  $P(A|B)$  на  $P(B)$ , получаем:

$$P(A \cap B) = P(A|B)P(B).$$

Это самый простой случай **правила умножения**: Вероятность, с которой происходят  $A$  и  $B$  равна вероятности, с которой происходит  $B$ , умноженной на вероятность, с которой происходит  $A$ , при условии, что произошло  $B$ .

Эту формулу можно обобщить на случай более чем двух событий:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

и так далее.

### 5.3 Пример применения формулы

Урна содержит два черных шара и три белых. Три шара выбираются последовательно наугад без возвращения. Какова вероятность, что все отобранные шары белые?

Эта задача может быть решена двумя способами. Первый способ: число всех перестановок пяти шаров  $n(S) = 5!$ , число тех, которые содержат на первых трех местах белые шары  $n(WWW) = 2!3!$ . Таким образом,

$$P(WWW) = \frac{n(WWW)}{n(S)} = \frac{2!3!}{5!} = \frac{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = \frac{1}{10}.$$

Второй способ использует правило умножения:

$$P(W_3 W_2 W_1) = P(W_3 | W_2 \cap W_1) P(W_2 | W_1) P(W_1) = \frac{1}{3} \cdot \frac{2}{4} \cdot \frac{3}{5} = \frac{1}{10}.$$

## 5.4 Теорема полной вероятности

Вернёмся к примеру, который был рассмотрен в разделе 4.1 и предположим, что процесс состоит из двух случайных шагов: выбор урны ( $B_1$  с вероятностью  $P(B_1)$ , или  $B_2$  с вероятностью  $P(B_2) = 1 - P(B_1)$ ) и извлечение шара из урны, выбранной на первом шаге. Вычислим вероятность, что выбранный шар окажется красным (событие  $R$ ):

$$\begin{aligned} P(R) &= \frac{n(R)}{n(S)} = \frac{n(R_1 \cup R_2)}{n(S)} = \frac{n(R_1) + n(R_2)}{n(S)} = \frac{n(R \cap B_1) + n(R \cap B_2)}{n(S)} = \\ &= \frac{n(R \cap B_1)}{n(S)} + \frac{n(R \cap B_2)}{n(S)} = P(R \cap B_1) + P(R \cap B_2) = P(R|B_1)P(B_1) + P(R|B_2)P(B_2). \end{aligned}$$

**Т.1.** Если события  $B_k$ ,  $k = 1, \dots, n$ , представляют разбиение пространства  $S$  (то есть  $B_1 \cup B_2 \cup \dots \cup B_n = S$  и  $B_k \cap B_j = \emptyset$ ,  $k \neq j$ ), тогда для любого события  $A$

$$P(A) = \sum_{k=1}^n P(A|B_k)P(B_k)$$

**Доказательство.**

$$\begin{aligned} A &= (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n), \\ P(A) &= P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n) = \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n) = \\ &= \sum_{k=1}^n P(A|B_k)P(B_k) \end{aligned}$$

## 5.5 Теорема Байеса

**Теорема.** Если события  $B_1, \dots, B_n$  представляют разбиение пространства  $S$ , где  $P(B_j) \neq 0$  при  $j = 1, 2, \dots, n$ , тогда для любого события  $A$  такого что  $P(A) \neq 0$ ,

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^n P(A|B_j)P(B_j)}, \quad j = 1, \dots, n.$$

**Доказательство.**

$$P(B_k|A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(A \cap B_k)}{P((A \cap B_1) \cup \dots \cup (A \cap B_n))} =$$

$$= \frac{P(A \cap B_k)}{P(A \cap B_1) + \dots + P(A \cap B_n)} = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^n P(A|B_j)P(B_j)}.$$

Эта теорема выражает отношение между условными вероятностями, где роль событий была изменена на обратную. Она показывает, как вероятность условного события может быть найдена с использованием известного результата эксперимента.

## 5.6 Примеры.

В эксперименте, рассмотренном в разделе 4.4.  $P(B_1) = \frac{1}{3}$  and  $P(B_2) = \frac{2}{3}$ . Разберём две задачи.

**Задача 1.** Найти вероятность, что шар, вытянутый таким способом будет красным. Мы должны использовать теорему полной вероятности:

$$P(R) = P(R|B_1)P(B_1) + P(R|B_2)P(B_2) = \frac{1}{5} \cdot \frac{1}{3} + \frac{2}{5} \cdot \frac{2}{3} = \frac{1}{3}.$$

**Задача 2.** Нам известно, что вытянутый шар оказался красным и необходимо найти вероятность, что он был вынут из первой урны. Теперь необходимо использовать теорему Байеса:

$$P(B_1|R) = \frac{P(R|B_1)P(B_1)}{P(R|B_1)P(B_1) + P(R|B_2)P(B_2)} = \frac{\frac{1}{5} \cdot \frac{1}{3}}{\frac{1}{5} \cdot \frac{1}{3} + \frac{2}{5} \cdot \frac{2}{3}} = \frac{1}{5}.$$

Аналогичные вычисления дают  $P(B_2|R) = 4/5$ .

## 5.7 Независимые события

Вернёмся к процессу, рассмотренному в разделе 4.3. Шары выбирались *без возвращения*: каждый извлечённый шар убирался совсем. В результате изменились условные вероятности:

$$P(W) = \frac{3}{5}, \quad P(W|W_1) = \frac{2}{4}, \quad P(W|W_2 \cap W_1) = \frac{1}{3}.$$

Теперь рассмотрим такой же эксперимент, но *с возвращением*. В этом случае

$$P(W) = \frac{3}{5}, \quad P(W|W_1) = \frac{3}{5}, \quad P(W|W_2 \cap W_1) = \frac{3}{5}.$$

Условная вероятность не зависит от предыдущих событий и совпадает с безусловной вероятностью.

**Определение 1.** Если  $P(B) > 0$  и условная вероятность  $P(A|B)$  не зависит от  $B$ , тогда говорят, что события  $A$  и  $B$  являются *независимыми*.

Применяя правило умножения мы можем получить второе определение.

**Определение 2.** События  $A$  и  $B$  называются независимыми, если  $P(A \cap B) = P(A)P(B)$ .

**Замечание:** Не путайте независимые события с несовместными. Несовместные события весьма зависимы: если одно из них происходит, тогда другое не может произойти. Если события независимы, то вероятность появления одного из них остаётся той же самой, независимо от того, произошло другое событие или нет.

## 6 Лекция 5. Случайные величины

### 6.1 Распределения вероятностей

Бросая два игральных кубика, мы производим статистический эксперимент с пространством  $S$  состоящим из 36 исходов  $s_j$ . Присвоим число  $x$  = сумма каждому исходу, так что у нас будет один исход с  $x = 2$ , два исхода с  $x = 3$ , и так далее:

число	$x_1$	$x_2$	...
исход	$s_1$	$s_2$	...

и

исход	$s_1$	$s_2$	...
вероятность	$P(A = s_1)$	$P(A = s_2)$	...

Теперь мы объединим таблицы, используя обозначения  $p(x_1)$ ,  $p(x_2)$ , ... для вероятностей  $P(A = s_1)$ ,  $P(A = s_2)$ , ... :

число	$x_1$	$x_2$	...
вероятность	$p(x_1)$	$p(x_2)$	...

Таким образом мы преобразуем случайные события в случайные числа, набор их возможных значений вместе с соответствующими вероятностями образуют новый объект, который называют *случайной величиной*.

**Определение.1.** *Случайная величина* - это функция, которая связывает действительное число  $x$  с каждым событием  $A$  из данного разбиения пространства элементарных событий. Она обозначается заглавной буквой  $X$  ( от  $Y, Z, \dots$ ), числа  $x$  называются *значениями случайно величины  $X$*  и обозначаются прописными буквами  $x$  (или  $y, z, \dots$ ). Если они образуют дискретное множество  $x_1, x_2, \dots$ , как в примере, рассмотренном выше, мы называем случайную величину *дискретной случайной величиной* и говорим, что она принимает значения  $x_1, x_2, \dots$  с вероятностями

$$p(x_k) = P(X = x_k), \quad k = 1, 2, 3 \dots$$

**Определение 2.** *Дискретная функция распределения* - это любое множество пар  $(x, p(x))$ , удовлетворяющее условиям:

$$1) p(x) \geq 0, \quad \text{and} \quad 2) \sum p(x) = 1.$$

**Пример.** Случайная величина  $X$ , представляющая сумму в эксперименте с бросанием двух игральных кубиков характеризуется дискретным распределением вероятностей

$x$	2	3	4	...	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	...	$\frac{1}{36}$

Часто оказывается удобным представлять распределение вероятностей в виде формулы. Например, вышеприведённое распределение может быть представлено формулой

$$p(x) = \frac{6 - |x - 7|}{36}$$

для  $x = 2, 3, \dots, 12$ . Здесь  $|x - 7|$  - абсолютное значение числа  $x - 7$ .

## 6.2 Плотность распределения вероятностей

Предположим теперь, что полная вероятность 1 непрерывно распределена по интервалу  $(a, b)$ , размазана по нему. Такая случайная величина называется *непрерывной*.

**Определение 1.** *Плотность распределения вероятности* непрерывной случайной величины  $X$  определяется с помощью предела:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x}$$

Другими словами, вероятность для  $X$  попасть в малый интервал  $\Delta x$  определяется произведением соответствующей плотности вероятности и длины интервала:

$$f(x)\Delta x \approx P(x < X \leq x + \Delta x).$$

Путём разбиения интервала  $(a, b]$  конечной длины, лежащего на оси  $x$ , на малые элементы  $\Delta x$

$$\bigcup_k (x_k, x_k + \Delta x] = (a, b],$$

получаем

$$P\left(\bigcup_k (x_k < X \leq x_k + \Delta x)\right) = \sum_k P(x_k < X \leq x_k + \Delta x) = P(a < X \leq b).$$

Переходя к пределу  $\Delta x$  с заменой просуммированных вероятностей на  $f(x_k)\Delta x$ , получаем:

$$\lim_{\Delta x \rightarrow 0} \sum f(x_k)\Delta x = \int_a^b f(x)dx = P(a < X \leq b).$$

При  $a = -\infty$  и  $b = \infty$  имеем  $P(a < X \leq b) = P(-\infty < X \leq \infty) = 1$  и следовательно

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

**Определение 2.** *Плотность распределения вероятности* - это любая функция  $f(x)$ , удовлетворяющая условиям:

$$1) f(x) \geq 0, \quad \text{и} \quad 2) \int_{-\infty}^{\infty} f(x)dx = 1.$$

## 6.3 Кумулятивная функция распределения

**Определение 1.** Вероятность события  $X \leq x$ , представленная в виде функции от переменной  $x$ ,  $-\infty < x < \infty$  называется *кумулятивной функцией распределения*  $F(x)$  или просто *кумулятивным распределением* случайной величины  $X$ :

$$F_X(x) = P(X \leq x).$$

**Некоторые свойства кумулятивных распределений**

1. Из определения следует, что

$$0 \leq F(x) \leq 1.$$

2. Если случайная величина  $X$  ограничена значениями  $x_{\min}$  и  $x_{\max}$ , тогда  $F(x) = 0$ , если  $x < x_{\min}$ , и  $F(x) = 1$ , если  $x > x_{\max}$ . В любом случае,  $F(-\infty) = 0$  и  $F(+\infty) = 1$ .

3.  $F(x)$  - неубывающая функция,

$$F(x_1) \leq F(x_2) \text{ if } x_1 < x_2,$$

и

$$F(x_2) - F(x_1) = P(x_1 < X \leq x_2).$$

4. Если  $X$  не может принимать значения внутри интервала  $(x_1, x_2)$  тогда  $F(x) = \text{const}$  внутри него.

5. Для дискретной случайной величины кумулятивная функция является ступенчатой функцией:  $F(x) = \sum_{x_k \leq x} f(x_k)$ .

6. Для непрерывной случайной величины кумулятивная функция выражается через её плотность вероятности с помощью интеграла:

$$F(x) = \int_{-\infty}^x f(x') dx'.$$

Плотность вероятности выражается через кумулятивную функцию с помощью производной:

$$f(x) = \frac{dF(x)}{dx}.$$

## 6.4 Графические представления распределений

Discrete distribution.

Continuous densities.

Mode, medians.

Cumulative distributions.

Histograms.

## 7 Лекция 6. Моменты случайных величин

### 7.1 Математическое ожидание

По аналогии с центром распределения масс в механике

$$\bar{x} = \frac{x_1 m(x_1) + x_2 m(x_2) + \dots + x_n m(x_n)}{m(x_1) + m(x_2) + \dots + m(x_n)},$$

величина

$$\mu = \frac{x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n)}{p(x_1) + p(x_2) + \dots + p(x_n)} = \sum x p(x)$$

показывает центр распределения вероятностей, называемый *средним значением* дискретной случайной величины. Среднее значение непрерывной случайной величины даётся интегралом

$$\mu = \int_{-\infty}^{\infty} x f(x) dx.$$

Удобно ввести общее обозначение для этих операций:

$$EX \equiv \begin{cases} \sum x p(x) & (\text{дискретный случай}), \\ \int_{-\infty}^{\infty} x f(x) dx & (\text{непрерывный случай}). \end{cases}$$

Эта операция и её результат называются *математическим ожиданием*, а величину  $EX$  часто называют *ожидаемым значением*. Это понятие может быть обобщено на случай произвольной функции случайной величины  $g(X)$ :

$$Eg(X) \equiv \begin{cases} \sum g(x) p(x) & \text{в дискретном случае,} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{в непрерывном случае.} \end{cases}$$

**Замечание:** Заметим, что нужно писать заглавную букву  $X$  после  $E$ , но маленькие буквы  $x$  после символов суммы и интеграла.

### 7.2 Некоторые свойства математического ожидания

1.  $Ec = c$ , если  $c = \text{const}$  не является случайной величиной.

2.  $E(cX) = cEX$ .

3.  $E(X + c) = EX + c$ .

4.  $E(X + c)^2 = EX^2 + 2cEX + c^2$ ,

или, в более общем виде

$$E(X + c)^n = \sum_{k=0}^n \binom{n}{k} c^k EX^{n-k},$$

где

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

**Доказательство.** Доказательство основано на формуле бинома Ньютона:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}.$$

5.  $EX \geq 0$ , if  $X \geq 0$ .

6. Если  $X$  имеет симметричное относительно  $x = c$  распределение, тогда  $EX = c$ .

**Доказательство.** С.в.  $X' = X - c$  симметрично распределена относительно нуля, её плотность  $f_{X'}(x)$  – чётная функция, и

$$\langle X' \rangle = \int_{-\infty}^{\infty} x f_{X'}(x) dx = 0$$

как интеграл от нечетной функции. Отсюда  $\langle X \rangle = \langle c \rangle = c$ . Примеры:

$$1. Ec = \sum xp(x) = cp(c) = c \cdot 1 = c.$$

$$2. E(cX) = \left\{ \begin{array}{l} \sum c xp(x) = c \sum xp(x) \\ \int_{-\infty}^{\infty} cx f(x) dx = c \int_{-\infty}^{\infty} x f(x) dx \end{array} \right\} = cEX.$$

7. Пусть  $A$  - некоторое подмножество возможных значений случайной величины  $X$ , и

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Тогда

$$E\mathbf{1}_A(X) = P(X \in A).$$

## 7.3 Моменты случайных величин

**Определение 1.**  $N$ -й момент случайной величины  $X$  около точки  $x = c$  определяется выражением  $E(X - c)^n$ ,  $n = 1, 2, 3, \dots$

**Определение 2.** Моменты около точки  $x = 0$  называются *начальными моментами* и обозначаются

$$\mu_n = EX^n, \quad n = 1, 2, 3, \dots$$

$$\mu_1 = \mu$$

**Определение 3.** Моменты около  $x = \mu$  называются *центральными моментами* и обозначаются  $\overset{\circ}{\mu}_n$ :

$$\overset{\circ}{\mu}_n = E(X - \mu)^n, \quad \overset{\circ}{\mu}_1 = 0, \quad \overset{\circ}{\mu}_2 = \sigma^2.$$

**Теорема 2.** Начальные и центральные моменты связаны соотношениями

$$\mu_n = \sum_{k=0}^n \binom{n}{k} \mu^k \overset{\circ}{\mu}_{n-k},$$

$$\overset{\circ}{\mu}_n = \sum_{k=0}^n \binom{n}{k} (-\mu)^k \mu_{n-k}.$$

**Доказательство.** Следует из (6.3.4).



## 7.4 Дисперсия

Среднее значение величины  $X$ , обозначаемое  $\mu$ , показывает положение распределения на оси  $x$ . Чтобы охарактеризовать ширину распределения, рассмотрим разность  $X - \mu$ . Но это случайная величина и её ожидаемое значение равно нулю,

$$E(X - \mu) = EX - E\mu = EX - \mu = \mu - \mu = 0.$$

Поэтому используют квадрат разности, ожидаемое значение которого отлично от нуля. Его называют *дисперсией* и обозначают  $\text{Var}X$ :

$$\text{Var}X = E(X - \mu)^2.$$

**Теорема 1.** Дисперсия - это наименьший из моментов  $E(X - c)^2$ ,  $-\infty < c < \infty$ , соответствующий  $c = \mu$ .

**Доказательство.** Найдём положение  $c_{\min}$  минимума функции  $\psi(c) = E(X - c)^2$  используя известный метод ( $\psi'(c_{\min}) = 0$ ,  $\psi''(c_{\min}) > 0$ ).

Явное выражение для дисперсии:

$$\text{Var}X = \sum (x - \mu)^2 p(x)$$

если  $X$  дискретная, и

$$\text{Var}X = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

если  $X$  непрерывная случайная величина.

**Теорема 2.** Дисперсия случайной величины может быть представлена в следующем виде, более удобным для использования:

$$\text{Var}X = EX^2 - \mu^2 = \begin{cases} \sum x^2 p(x) - \mu^2, & \text{если } X \text{ дискретная;} \\ \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2, & \text{если } X \text{ непрерывная.} \end{cases}$$

**Доказательство.**

**Замечание.** Поскольку величины  $\sum x^2 p(x)$  и  $\int x^2 f(x) dx$  называются вторыми моментами и обозначаются  $\mu'_2$ , эта формула может быть переписана в виде  $\text{Var}X = \mu'_2 - \mu^2$ .

Дисперсия сама по себе не может характеризовать ширину распределения, поскольку её размерность отличается от размерности  $x$ . Для этой цели используется так называемое *стандартное отклонение*, определяемое как положительный квадратный корень из дисперсии:

$$\sigma = \sqrt{\text{Var}X},$$

таким образом

$$\text{Var}X = \sigma^2.$$

В дальнейшем мы будем часто использовать для дисперсии обозначение  $\sigma^2$ .

**Основные свойства дисперсии:**

1)  $\text{Var}X \geq 0$ .

2)  $\text{Var}(X + c) = \text{Var}X$  (добавление или вычитание неслучайной величины не изменяет дисперсию случайной величины).

3)  $\text{Var}(cX) = c^2 \text{Var}X$ .

Каждое из свойств легко доказывается с использованием определения дисперсии. Например, последнее:

$$\begin{aligned} \text{Var}(cX) &= E(cX - \mu_{cX})^2 = E(cX - E(cX))^2 = E(cX - cEX)^2 = Ec^2(X - EX)^2 = \\ &= c^2 E(X - EX)^2 = c^2 E(X - \mu_X)^2 = c^2 \text{Var}X. \end{aligned}$$

## 7.5 Производящая функция моментов

**Определение.** Математическое ожидание функции  $e^{tX}$  от действительной величины  $t$  называется *производящей функцией моментов случайной величины  $X$*  (ПФМ) и обозначается  $M_X(t)$ :

$$M_X(t) = \mathbb{E}e^{tX}.$$

Эта функция после  $n$ -кратного дифференцирования в точке  $t = 0$  производит  $n$ -й момент случайной величины  $X$ :

$$M_X^{(n)}(0) \equiv \left. \frac{d^n \mathbb{E}e^{tX}}{dt^n} \right|_{t=0} = \mathbb{E} \left. \frac{d^n e^{tX}}{dt^n} \right|_{t=0} = \mathbb{E} X^n e^{tX} \Big|_{t=0} = \mathbb{E} X^n = \mu_n.$$

Обратное соотношение даётся формулой

$$M_X(t) = 1 + \mu t + \frac{\mu_2' t^2}{2} + \dots = \sum_{k=0}^{\infty} \frac{(\mu_k)^k}{k!} t^k,$$

если все моменты существуют (заметим, что  $\mu_0 = 1$ ).

Другие свойства ПФМ:

1)  $M_X(0) = 1$  для любой случайной величины  $X$ .

2)  $M_c(t) = e^{tc}$  для любой неслучайной  $X = c$ .

3)  $M_{a+X} = e^{at} M_X(t)$  для любой неслучайной  $a$ .

4)  $M_{bX}(t) = M_X(bt)$  для любой неслучайной  $b$ .

5) При определённых условиях существует однозначное соответствие между ПФМ и распределениями.

## 8 Лекция 7. Типичные задачи

**Задача 1.** Бросаются два игральных кубика: красный ( $R$ ) и зелёный ( $G$ ). **Найти:**  $P(R = 4)$  и  $P(R = 4 | R + G \geq 8)$ .

**Решение.** Нарисуем пространство элементарных исходов, обозначая каждый исход  $(R, G)$ :

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

Выделим исходы с  $R = 4$  жирным шрифтом:

(1, 1)	(2, 1)	(3, 1)	<b>(4, 1)</b>	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	<b>(4, 2)</b>	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	<b>(4, 3)</b>	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	<b>(4, 5)</b>	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	<b>(4, 6)</b>	(5, 6)	(6, 6)

Найдём вероятность  $P(R = 4, G \text{ произвольное})$  вычисляя отношение числа выделенных жирным шрифтом исходов к числу всех исходов:

$$P(R = 4, G \text{ произвольное}) = \frac{6}{36} = \frac{1}{6}.$$

Что касается условной вероятности, то есть два способа её нахождения. Один из них основан на прямом использовании определения и применим во всех случаях:

$$P(R = 4, G \text{ arbitrary} | R + G \geq 8) = \frac{P((R = 4, G \text{ arbitrary}) \cap (R + G \geq 8))}{P(R + G \geq 8)}.$$

Используя для нахождения числителя диаграмму

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	<b>(4, 5)</b>	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	<b>(4, 6)</b>	(5, 6)	(6, 6)

а для вычисления знаменателя диаграмму

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	<b>(6, 2)</b>
(1, 3)	(2, 3)	(3, 3)	(4, 3)	<b>(5, 3)</b>	<b>(6, 3)</b>
(1, 4)	(2, 4)	(3, 4)	<b>(4, 4)</b>	<b>(5, 4)</b>	<b>(6, 4)</b>
(1, 5)	(2, 5)	<b>(3, 5)</b>	<b>(4, 5)</b>	<b>(5, 5)</b>	<b>(6, 5)</b>
(1, 6)	<b>(2, 6)</b>	<b>(3, 6)</b>	<b>(4, 6)</b>	<b>(5, 6)</b>	<b>(6, 6)</b>

находим:

$$P((R = 4, G \text{ arbitrary}) \cap (R + G \geq 8)) = \frac{3}{36} = \frac{1}{12}, \quad P(R + G \geq 8) = \frac{15}{36} = \frac{5}{12},$$

таким образом:

$$P(R = 4 | R + G \geq 8) = \frac{\frac{1}{12}}{\frac{5}{12}} = \frac{1}{5}.$$

Другой способ применим к случаю равновероятных исходов. Удалим из диаграммы все исходы, которые не удовлетворяют условию  $R + G \geq 8$ ,

$$\begin{array}{ccccccc} & & & & & & (6, 2) \\ & & & & & (5, 3) & (6, 3) \\ & & & (4, 4) & (5, 4) & (6, 4) \\ & (3, 5) & (4, 5) & (5, 5) & (6, 5) \\ (2, 6) & (3, 6) & (4, 6) & (5, 6) & (6, 6), \end{array}$$

отметим исходы с  $R = 4$

$$\begin{array}{ccccccc} & & & & & & (6, 2) \\ & & & & & (5, 3) & (6, 3) \\ & & & (4, 4) & (5, 4) & (6, 4) \\ & (3, 5) & (4, 5) & (5, 5) & (6, 5) \\ (2, 6) & (3, 6) & (4, 6) & (5, 6) & (6, 6), \end{array}$$

и затем найдём условную вероятность, как отношение числа исходов:

$$P(R = 4, G \text{ arbitrary} | R + G \geq 8) = \frac{3}{15} = \frac{1}{5}.$$

**Задача 2.** Ящик содержит 4 красных и 6 зелёных шаров. A box contains 4 red balls and 6 green balls. (a) Случайным образом вынимаем первый шар, а затем, без возвращения первого, второй. Найти вероятности  $P(R_1)$ ,  $P(R_2 | R_1)$  и  $P(R_2 \cap R_1)$ . (b) Выполним такую же процедуру с возвращением. Найти вероятности тех же событий.

**Решение:**

(a)

$$P(R_1) = \frac{4}{10} = \frac{2}{5}; \quad P(R_2 | R_1) = \frac{4-1}{10-1} = \frac{1}{3}; \quad P(R_2 \cap R_1) = P(R_2 | R_1)P(R_1) = \frac{1}{3} \cdot \frac{2}{5} = \frac{2}{15}.$$

(b)

$$P(R_1) = \frac{4}{10} = \frac{2}{5}; \quad P(R_2 | R_1) = \frac{4}{10} = \frac{2}{5}; \quad P(R_2 \cap R_1) = P(R_2 | R_1)P(R_1) = \frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25}.$$

**Задача 3.** Три машины  $B_1, B_2$  и  $B_3$  производят 30%, 45% и 25% продукции с 2%, 3% и 1% брака соответственно. Конечный продукт выбирается случайным образом. Какова вероятность, что он окажется бракованным (событие  $D$ )?

**Решение:**

$$\begin{aligned} P(D) &= P(D | B_1)P(B_1) + P(D | B_2)P(B_2) + P(D | B_3)P(B_3) = \\ &= 0.02 \cdot 0.30 + 0.03 \cdot 0.45 + 0.01 \cdot 0.25 = 0.022. \end{aligned}$$

**Задача 4.**

В отношении предыдущего примера, если продукт был выбран случайным образом и оказался бракованным, какова вероятность, что он был изготовлен машиной  $B_1$ ?  $B_2$ ?  $B_3$ ?

**Решение.**

$$\begin{aligned} P(B_1|D) &= \frac{P(D|B_1)P(B_1)}{P(D|B_1)P(B_1) + P(D|B_2)P(B_2) + P(D|B_3)P(B_3)} = \\ &= \frac{0.02 \cdot 0.30}{0.02 \cdot 0.30 + 0.03 \cdot 0.45 + 0.01 \cdot 0.25} \approx 0.27. \\ P(B_2|D) &= \frac{P(D|B_2)P(B_2)}{P(D|B_1)P(B_1) + P(D|B_2)P(B_2) + P(D|B_3)P(B_3)} = \\ &= \frac{0.03 \cdot 0.45}{0.02 \cdot 0.30 + 0.03 \cdot 0.45 + 0.01 \cdot 0.25} \approx 0.61, \\ P(B_3|D) &= \frac{P(D|B_3)P(B_3)}{P(D|B_1)P(B_1) + P(D|B_2)P(B_2) + P(D|B_3)P(B_3)} = \\ &= \frac{0.01 \cdot 0.25}{0.02 \cdot 0.30 + 0.03 \cdot 0.45 + 0.01 \cdot 0.25} \approx 0.11 \end{aligned}$$

Проверка:  $P(B_1|D) + P(B_2|D) + P(B_3|D) = 0.99 \approx 1$  (небольшие погрешности могут являться результатом округления).

**Задача 5.** Экзаменационный билет содержит 8 вопросов. Считается, что студент успешно прошёл тест, если он правильно ответил на 6 или более вопросов. **Чему равна** вероятность того, что студент сдаст экзамен?

**Решение:** Обозначим событие (студент сдал экзамен) через  $A$ . Какие исходы оно в себя включает? Один из исходов, когда студент правильно ответил на все вопросы (обозначим его через  $c$ ), может быть представлен в виде:

$$\{c \ c \ c \ c \ c \ c \ c \ c\}$$

Существует 8 исходов соответствующих одному правильному ответу ( $i$ ):

$$\begin{aligned} &\{i \ c \ c \ c \ c \ c \ c \ c\} \\ &\{c \ i \ c \ c \ c \ c \ c \ c\} \\ &\{c \ c \ i \ c \ c \ c \ c \ c\} \\ &\{c \ c \ c \ i \ c \ c \ c \ c\} \\ &\{c \ c \ c \ c \ i \ c \ c \ c\} \\ &\{c \ c \ c \ c \ c \ i \ c \ c\} \\ &\{c \ c \ c \ c \ c \ c \ i \ c\} \\ &\{c \ c \ c \ c \ c \ c \ c \ i\} \end{aligned}$$

Наконец, существуют исходы, содержащие два неправильных ответа. Сколько таких исходов содержится в  $A$ ? Когда первый ответ оказывается неправильным, 7 вариантов возможно для другого. Если неправильным оказывается второй ответ, снова 7 вариантов возможны для другого неправильного ответа и так далее. Умножение даёт  $8 \cdot 7$ , но такое произведение учитывает каждую пару неправильных ответов дважды, так что правильное число будет  $\frac{8 \cdot 7}{2} = 28$ . Общее число исходов  $n(S) = 2^8$  и все они равновероятны. Таким образом, вероятность

$$P(A) = \frac{n(A)}{n(S)} = \frac{1 + 8 + 28}{256} = \frac{37}{256} \approx 0.14.$$

## STAT 312 Spring

### Домашняя работа 2

**Задача 1.** Одно из ста чисел 00, 01, 02, ..., 98, 99 выбирается случайным образом. Рассмотрите следующие события:

$A_1$ : обе цифры одинаковы;

$A_2$ : первая цифра больше, чем вторая;

$A_3$ : вторая цифра 1;

$A_4$ : сумма двух цифр равна 9;

$A_5$ : обе цифры превышают 3.

**Найти:** (a)  $P(A_1)$  и  $P(A_1|A_5)$ ; (b)  $P(A_2)$  и  $P(A_2|A_1)$ ; (c)  $P(A_3)$  и  $P(A_3|A_2)$ ; (d)  $P(A_4)$  и  $P(A_4|A_3)$ ; (e)  $P(A_5)$  и  $P(A_5|A_4)$ .

**Задача 2.** Урна содержит два чёрных шара и три белых. Последовательно и случайно выбираются два шара.

**Найти** вероятность каждого из четырёх исходов

$B_1B_2$ ,  $B_1W_2$ ,  $W_1B_2$ , и  $W_1W_2$  если: (a) первый шар не возвращается в урну перед тем, как выбирается второй шар (выборка *без возвращения*); (b) первый шар возвращается назад в урну перед тем, как извлекается второй шар (выборка *с возвращением*).

**Задача 3.** В урне находится пятнадцать шаров, 10 из которых белые. Случайным образом выбираются четыре, *без возвращения*.

**Вычислить** вероятность того, что (a) все четыре шара белые, (b) ни один из шаров не оказался белым, (c) ровно один из выбранных шаров белый.

**Задача 4.** Производитель калькуляторов покупает основной процессор у трёх разных поставщиков ( $B_1$ ,  $B_2$ ,  $B_3$ ). Практика показала, что 1% микросхем первого поставщика, 4% второго поставщика, и 2% от третьего являются бракованными. Зная, что этот производитель покупает 30% процессоров у первого, 10% у второго и остальные у третьего поставщика

**вычислить** вероятность, что процессор окажется бракованным (событие  $A$ ) при проверке перед установкой в калькулятор.

**Задача 4А.** В условиях, сформулированных в Задаче 4, все чипы, купленные производителем у разных поставщиков помещены в один контейнер без учёта названия поставщика. Случайным образом выбирается один из чипов и оказывается бракованным.

**Найти** вероятность, что он поступил: (a) от первого поставщика; (b) от второго поставщика; (3) от третьего поставщика.

**Задача 5.** Два друга тщательно исследовали монету. Первый из них говорит, что монета правильная, то есть вероятности выпадения герба ( $H$ ) или решки ( $T$ ) одинаковы ( $B_1$ ), второй говорит, что монета неправильная, так что вероятность выпадения герба в три раза больше, чем вероятность выпадения решки ( $B_2$ ). Без какой-либо дополнительной информации будем предполагать, что оба эти мнения с равной вероятностью могут оказаться верными (монета в действительности может быть деформирована).

Теперь монета подбрасывается один раз и (a) выпадает герб; (b) выпадает решка.

Используя теорему Байеса, **вычислите** вероятности событий  $B_1$  и  $B_2$  в свете представленной информации для каждого случая в отдельности, то есть **найдите**  $P(B_1|H)$ ,  $P(B_2|H)$ ,  $P(B_1|T)$  и  $P(B_2|T)$ .

**Задача 6.** Три команды,  $A, B$  и  $C$ , участвуют в соревнованиях, в которых каждая команда играет с каждой другой командой один раз, и ничейный результат не допускается. Допустим, что  $P(A \text{ победит } B) = 0.4$ ,  $P(B \text{ победит } C) = 0.5$ , и  $P(C \text{ победит } A) = 0.6$  и исходы всех игр независимы.

**Вычислить** вероятности следующих событий: (а)  $A$  выиграет соревнования (то есть  $A$  победит  $B$  но  $C$  не победит  $A$ ); (b)  $B$  выиграет соревнования; (с) никто не выиграет соревнования (заметьте, что в этом случае возможны два исхода).

**Задача 7.** Пропорция женщин в некоторой популяции равна  $p$  то есть, если  $N$  - это размер популяции, тогда она содержит  $pN$  женщин (F) и  $(1-p)N$  мужчин (M). Пять человек выбираются случайным образом *с возвращением*.

**Найти** вероятность исхода MFFFF: (а) если  $p = 0.5$ , (b) если  $p = 0.2$ .

**Задача 8.** Экзаменационный билет теста содержит 7 вопросов. Если считать, что студент успешно сдаёт тест при условии правильного ответа на 5 или более вопросов, **какова** вероятность, что студент сдаст экзамен?

## STAT 312 Spring

### Домашняя работа 2 - решение

**Задача 1.** Одно из ста чисел 00, 01, 02, ..., 98, 99 выбирается случайным образом.

Рассмотрите следующие события:

$A_1$ : обе цифры одинаковы;

$A_2$ : первая цифра больше, чем вторая;

$A_3$ : вторая цифра 1;

$A_4$ : сумма двух цифр равна 9;

$A_5$ : обе цифры превышают 3.

**Найти:** (a)  $P(A_1)$  и  $P(A_1|A_5)$ ; (b)  $P(A_2)$  и  $P(A_2|A_1)$ ; (c)  $P(A_3)$  и  $P(A_3|A_2)$ ; (d)  $P(A_4)$  и  $P(A_4|A_3)$ ; (e)  $P(A_5)$  и  $P(A_5|A_4)$ .

**Указание:** запишите все возможные исходы в виде таблицы

00	01	02	03	04	05	06	07	08	09
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

**Solution (a):** Отметим исходы с одинаковыми цифрами (то есть исходы, принадлежащие к  $A_1$ )

<b>00</b>	01	02	03	04	05	06	07	08	09
10	<b>11</b>	12	13	14	15	16	17	18	19
20	21	<b>22</b>	23	24	25	26	27	28	29
30	31	32	<b>33</b>	34	35	36	37	38	39
40	41	42	43	<b>44</b>	45	46	47	48	49
50	51	52	53	54	<b>55</b>	56	57	58	59
60	61	62	63	64	65	<b>66</b>	67	68	69
70	71	72	73	74	75	76	<b>77</b>	78	79
80	81	82	83	84	85	86	87	<b>88</b>	89
90	91	92	93	94	95	96	97	98	<b>99</b>

и найдём отношение:

$$P(A_1) = \frac{n(A_1)}{n(S)} = \frac{10}{100} = 0.1.$$

Уберём исходы, которые не отвечают условию  $A_5$



—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	<b>44</b>	45	46	47	48	49
—	—	—	—	54	<b>55</b>	56	57	58	59
—	—	—	—	64	65	<b>66</b>	67	68	69
—	—	—	—	74	75	76	<b>77</b>	78	79
—	—	—	—	84	85	86	87	<b>88</b>	89
—	—	—	—	94	95	96	97	98	<b>99</b>

и найдём отношение

$$P(A_1|A_5) = \frac{n(A_1 \cap A_5)}{n(A_5)} = \frac{6}{36} = \frac{1}{6} \approx 0.167.$$

**Остальные ответы:** (b) 0.45 and 0; (c) 0.10 and 0.178; (d) 0.10 and 0.10; (e) 0.36 and 0.2.

**Задача 2.** Урна содержит два чёрных шара и три белых. Последовательно и случайно выбираются два шара.

**Найти** вероятность каждого из четырёх исходов

$B_1B_2$ ,  $B_1W_2$ ,  $W_1B_2$ , и  $W_1W_2$  если: (a) первый шар не возвращается в урну перед тем, как выбирается второй шар (выборка *без возвращения*); (b) первый шар возвращается назад в урну перед тем, как извлекается второй шар (выборка *с возвращением*).

**Решение:**

$$\begin{aligned} P(B_2B_1) &\equiv P(B_2 \cap B_1) = P(B_2|B_1)P(B_1) = (a) \frac{1}{4} \cdot \frac{2}{5} = \mathbf{0.1}; \text{ or } (b) = \frac{2}{5} \cdot \frac{2}{5} = \mathbf{0.16}; \\ P(W_2B_1) &= P(W_2|B_1)P(B_1) = (a) \frac{3}{4} \cdot \frac{2}{5} = \mathbf{0.3}; \text{ or } (b) = \frac{3}{5} \cdot \frac{2}{5} = \mathbf{0.24}; \\ P(B_2W_1) &= P(B_2|W_1)P(W_1) = (a) \frac{2}{4} \cdot \frac{3}{5} = \mathbf{0.3}; \text{ or } (b) = \frac{2}{5} \cdot \frac{3}{5} = \mathbf{0.24}; \\ P(W_2W_1) &= P(W_2|W_1)P(W_1) = (a) \frac{2}{4} \cdot \frac{3}{5} = \mathbf{0.3}; \text{ or } (b) = \frac{3}{5} \cdot \frac{3}{5} = \mathbf{0.36}. \end{aligned}$$

**Задача 3.** В урне находится пятнадцать шаров, 10 из которых белые. Случайным образом выбираются четыре, *без возвращения*.

**Вычислить** вероятность того, что (a) все четыре шара белые, (b) ни один из шаров не оказался белым, (c) ровно один из выбранных шаров белый.

**Решение:**

$$\begin{aligned} (a)P(WWWW) &= P(W_4|W_3, W_2, W_1)P(W_3|W_2, W_1)P(W_2|W_1)P(W_1) = \\ &= \frac{7}{12} \cdot \frac{8}{13} \cdot \frac{9}{14} \cdot \frac{10}{15} \approx \mathbf{0.154}. \\ (b)P(W'W'W'W') &= P(W'_4|W'_3, W'_2, W'_1)P(W'_3|W'_2, W'_1)P(W'_2|W'_1)P(W'_1) = \\ &= \frac{2}{12} \cdot \frac{3}{13} \cdot \frac{4}{14} \cdot \frac{5}{15} \approx \mathbf{0.00366}. \\ (c)P(W'W'W'W) + P(W'W'WW') + P(W'WWW') + P(WW'W'W') &= \\ = \frac{3}{12} \cdot \frac{4}{13} \cdot \frac{5}{14} \cdot \frac{10}{15} + \frac{3}{12} \cdot \frac{4}{13} \cdot \frac{10}{14} \cdot \frac{5}{15} + \frac{3}{12} \cdot \frac{10}{13} \cdot \frac{4}{14} \cdot \frac{5}{15} + \frac{10}{12} \cdot \frac{3}{13} \cdot \frac{4}{14} \cdot \frac{5}{15} &\approx \\ \approx \mathbf{0.0733}. \end{aligned}$$

**Задача 4.** Производитель калькуляторов покупает основной процессор у трёх разных поставщиков ( $B_1, B_2, B_3$ ). Практика показала, что 1% микросхем первого поставщика, 4% второго поставщика, и 2% от третьего являются бракованными. Зная, что этот производитель покупает 30% процессоров у первого, 10% у второго и остальные у третьего поставщика, **вычислить** вероятность, что процессор окажется бракованным (событие  $A$ ) при проверке перед установкой в калькулятор.

**Решение:**

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) = \\ &= 0.01 \cdot 0.3 + 0.04 \cdot 0.1 + 0.02 \cdot 0.6 = \mathbf{0.019}. \end{aligned}$$

**Задача 4А.** В условиях, сформулированных в Задаче 4, все чипы, купленные производителем у разных поставщиков помещены в один контейнер без учёта названия поставщика. Случайным образом выбирается один из чипов и оказывается бракованным.

**Найти** вероятность, что он поступил: (а) от первого поставщика; (б) от второго поставщика; (3) от третьего поставщика.

**Решение:** используя теорему Байеса, находим:

$$(a) P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} = \frac{0.01 \cdot 0.30}{0.019} \approx \mathbf{0.158},$$

$$(a) P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} = \frac{0.04 \cdot 0.10}{0.019} \approx \mathbf{0.210},$$

$$(a) P(B_3|A) = \frac{P(A|B_3)P(B_3)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} = \frac{0.02 \cdot 0.60}{0.019} \approx \mathbf{0.632}.$$

Проверим нормировку:  $0.158 + 0.21 + 0.632 = 1$ . Заметим, что выполняя приближённые вычисления мы не можем ожидать точного выполнения условия нормировки.

**Задача 5.** Два друга тщательно исследовали монету. Первый из них говорит, что монета правильная, то есть вероятности выпадения герба ( $H$ ) или решки ( $T$ ) одинаковы ( $B_1$ ), второй говорит, что монета неправильная, так что вероятность выпадения герба в три раза больше, чем вероятность выпадения решки ( $B_2$ ). Без какой-либо дополнительной информации будем предполагать, что оба эти мнения с равной вероятностью могут оказаться верными (монета в действительности может быть деформирована).

Теперь монета подбрасывается один раз и (а) выпадает герб; (б) выпадает решка.

Используя теорему Байеса, **вычислите** вероятности событий  $B_1$  и  $B_2$  в свете представленной информации для каждого случая в отдельности, то есть **найдите**  $P(B_1|H)$ ,  $P(B_2|H)$ ,  $P(B_1|T)$  и  $P(B_2|T)$ .  $P(B_1|T)$  and  $P(B_2|T)$  и дайте устное описание.

**Решение:** используя теорему Байеса, находим:

$$(a) P(B_1|H) = \frac{P(H|B_1)P(B_1)}{P(H|B_1)P(B_1) + P(H|B_2)P(B_2)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2}} = \mathbf{0.4},$$

$$P(B_2|H) = 1 - P(B_1|H) = \mathbf{0.6}.$$

Таким образом, мы должны считать гипотезу  $B_1$  только на 40% верной, а  $B_2$  на 60% возможной.

$$(b) P(B_1|T) = \frac{P(T|B_1)P(B_1)}{P(T|B_1)P(B_1) + P(T|B_2)P(B_2)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}} = \mathbf{2/3},$$

$$P(B_2|T) = 1 - P(B_1|T) = 1/3.$$

В этом случае мы можем считать гипотезу  $B_1$  только приблизительно на 67% верной, а  $B_2$  возможной на 33%.

**Задача 6.** Три команды,  $A, B$  и  $C$ , участвуют в соревнованиях, в которых каждая команда играет с каждой другой командой один раз, и ничейный результат не допускается. Допустим, что  $P(A \text{ победит } B) = 0.4$ ,  $P(B \text{ победит } C) = 0.5$ , и  $P(C \text{ победит } A) = 0.6$  и исходы всех игр независимы.

**Вычислить** вероятности следующих событий: (а)  $A$  выиграет соревнования (то есть  $A$  победит  $B$  но  $C$  не победит  $A$ ); (б)  $B$  выиграет соревнования; (с) никто не выиграет соревнования (заметьте, что в этом случае возможны два исхода).

**Решение:**

$$(a) P(A \text{ выиграет}) = P(A \text{ победит } B)P(C \text{ не победит } A) = 0.4 \cdot 0.4 = \mathbf{0.16}.$$

$$(b) P(B \text{ выиграет}) = P(A \text{ не победит } B)P(B \text{ победит } C) = 0.6 \cdot 0.5 = \mathbf{0.30}.$$

$$\begin{aligned} (c) P(\text{не выиграет никто}) &= P(A \text{ победит } B)P(B \text{ победит } C)P(C \text{ победит } A) + \\ &+ P(A \text{ не победит } B)P(B \text{ не победит } C)P(C \text{ не победит } A) = \\ &= 0.4 \cdot 0.5 \cdot 0.6 + 0.6 \cdot 0.5 \cdot 0.4 = \mathbf{0.24}. \end{aligned}$$

**Задача 7.** Пропорция женщин в некоторой популяции равна  $p$  то есть, если  $N$  - это размер популяции, тогда она содержит  $pN$  женщин (F) и  $(1-p)N$  мужчин (M). Пять человек выбираются случайным образом *с возвращением*.

**Найти** вероятность исхода MFFFF: (а) если  $p = 0.5$ , (б) если  $p = 0.2$ .

**Ответ:** (а)  $1/32 \approx \mathbf{0.031}$ ; (б)  $\mathbf{0.00128}$ .

**Задача 8.** Экзаменационный билет теста содержит 7 вопросов. Если считать, что студент успешно сдаёт тест при условии правильного ответа на 5 или более вопросов, **какова** вероятность, что студент сдаст экзамен?

**Решение:** Пусть  $A_0$  обозначает событие, когда все ответы правильны,  $A_1$  означает, что только один из них неправильный,  $A_2$  - только два ответа неправильны. Тогда

$$\begin{aligned} P(\text{сдал}) &= P(A_0) + P(A_1) + P(A_2) = \\ &= \left(\frac{1}{2}\right)^7 + 7 \left(\frac{1}{2}\right)^7 + \frac{7(7-1)}{2} \left(\frac{1}{2}\right)^7 = \frac{29}{128} \approx \mathbf{0.227}. \end{aligned}$$

## 9 Лекция 8. Совместно распределённые случайные величины

### 9.1 Совместные распределения вероятностей

**Определение 1.** Пусть  $X$  - случайная величина, принимающая дискретные значения  $\{x\}$  и  $Y$  - другая дискретная случайная величина с возможными значениями  $\{y\}$ . Функция двух дискретных переменных  $p(x, y) \equiv P(X = x \cap Y = y)$  называется *совместным распределением вероятностей* случайных величин  $X$  и  $Y$ .

**Свойства:**

- 1)  $0 \leq p(x, y) \leq 1$ ,
- 2)  $\sum_{\{x\}} \sum_{\{y\}} p(x, y) = 1$ .

**Определение 2.** Пусть  $X$  и  $Y$  - непрерывные случайные величины. Функция двух непрерывных переменных  $f(x, y)$ , определяемая равенством  $f(x, y)\Delta x\Delta y \approx P(x < X \leq x + \Delta x \cap y < Y \leq y + \Delta y)$ ,  $\Delta x \rightarrow 0$ ,  $\Delta y \rightarrow 0$ , называется *совместной плотностью вероятности* случайных величин  $X$  и  $Y$ .

**Свойства:**

- 1)  $f(x, y) \geq 0$ .
- 2)  $\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) = 1$ .

Расширяя понятие математического ожидания

$$Eg(X) = \begin{cases} \sum_{-\infty}^{\infty} g(x)p(x) & \text{в дискретном случае;} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{в непрерывном случае} \end{cases}$$

на случай двух случайных величин, мы приходим к следующей теореме.

**Theorem.** Пусть  $g(X, Y)$  - некоторая функция двух случайных величин  $X$  и  $Y$ , тогда её математическое ожидание определяется выражением

$$Eg(X, Y) = \begin{cases} \sum_{\{x\}} \sum_{\{y\}} g(x, y)p(x, y) & \text{в дискретном случае;} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy & \text{в непрерывном случае.} \end{cases}$$

**Замечание:** иногда необходимо указывать случайные величины с помощью индексов:  $f_{X,Y}(x, y)$ .

### 9.2 Маргинальные распределения

Рассмотрим эксперимент по бросанию двух игральных кубиков в случае, когда кубики не сбалансированы и появление результатов на верхних гранях не является независимым. В этом случае  $x, y \in \{1, 2, 3, 4, 5, 6\}$ , как и прежде, но  $p(x, y) \neq \frac{1}{36}$ . Предположим, что распределение  $p(x, y)$  известно. Каким образом можно найти вероятность  $P(X = x)$  зная  $p(x, y)$ ?

Посмотрим на следующую цепочку равенств:

$$\begin{aligned} \{X = x\} &= \{X = x\} \cap S = \\ &= \{X = x\} \cap (\{Y = 1\} \cup \{Y = 2\} \cup \{Y = 3\} \cup \{Y = 4\} \cup \{Y = 5\} \cup \{Y = 6\}) = \end{aligned}$$

$$= (\{X = x\} \cap \{Y = 1\}) \cup (\{X = x\} \cap \{Y = 2\}) \cup \dots \cup (\{X = x\} \cap \{Y = 6\}).$$

Вычисление вероятности этого события даёт

$$\begin{aligned} P(X = x) &= P(\{X = x\} \cap \{Y = 1\}) \cup (\{X = x\} \cap \{Y = 2\}) \cup \dots \cup (\{X = x\} \cap \{Y = 6\}) = \\ &= P(\{X = x\} \cap \{Y = 1\}) + P(\{X = x\} \cap \{Y = 2\}) + \dots + P(\{X = x\} \cap \{Y = 6\}) = \sum_y f(x, y). \end{aligned}$$

**Определение 1.** Распределения  $p_X(x) \equiv \sum_y p(x, y)$  и  $p_Y(y) \equiv \sum_x p(x, y)$  в дискретном случае и плотности  $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$  и  $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$  в непрерывном случае называются *маргинальными распределениями* (*маргинальными плотностями*) по отношению к совместному распределению  $p(x, y)$  (плотности  $f(x, y)$ ).

### 9.3 Условные распределения

Вспомним определение условной вероятности (см. 4.1):

**Определение.** Пусть  $B$  - любое событие, такое что  $P(B) > 0$ . Тогда *условная вероятность* наступления события  $A$  при условии, что  $B$  произошло даётся выражением

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Заменяя здесь  $A$  на  $\{X = x\}$  и  $B$  на  $\{Y = y\}$  мы получаем условное распределение.

**Определение.1.** Условное распределение случайной величины  $X$  при условии, что  $Y = y$ , обозначаемое  $p(x|y)$  в дискретном случае или  $f(x|y)$  в непрерывном случае, определяется выражениями

$$p(x|y) = \frac{p(x, y)}{p(y)},$$

или

$$f(x|y) = \frac{f(x, y)}{f(y)},$$

при условии, что  $p(y) > 0$  или  $f(y) > 0$  соответственно.

Свойства:

1) **Неотрицательность:**  $p(x|y) \geq 0$ , or  $f(x|y) \geq 0$ .

2) **Условие нормировки:**  $\sum_x p(x|y) = 1$ , or  $\int_{-\infty}^{\infty} f(x|y) dx = 1$

**Доказательство.**

$$\sum_x p(x|y) = \frac{\sum_x p(x, y)}{p(y)} = \frac{p(y)}{p(y)} = 1.$$

3) **Правило умножения:**  $p(x, y) = p(x|y)p(y)$ . or  $f(x, y) = f(x|y)f(y)$ .

4) **Формула полного распределения:**  $p(x) = \sum_y p(x|y)p(y)$  or  $p(x) = \int_{-\infty}^{\infty} p(x|y)p(y) dy$ .

**Доказательство** Необходимо заменить  $p(x) = P(X = x)$ ,  $p(y) = P(Y = y)$  и  $p(x|y) = P(X = x|Y = y)$ , и вернуться к 4.4.

## 9.4 Независимые случайные величины

**Определение 1.** Случайные величины  $X$  и  $Y$  называются независимыми, если условное распределение одной из них не зависит от другой:  $p(x|y) = p(x)$  или  $f(x|y) = f(x)$ .

**Теорема 1.** Совместное распределение независимых случайных величин равно произведению их маргинальных плотностей:  $p_{X,Y}(x, y) = p_X(x)p_Y(y)$  от  $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ .

**Доказательство:**  $f_{X,Y}(x, y) = f_X(x|y)f_Y(y) = f_X(x)f_Y(y)$ .

**Замечание.** Это свойство часто используется как определение независимых случайных величин.

**Теорема 2.** Пусть  $X$  и  $Y$  - независимые случайные величины, а  $g(X)$  и  $h(Y)$  - некоторые функции от них, тогда

$$E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)].$$

**Доказательство.** Используя теорему из 8.1, получаем:

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy g(x)h(y)f(x, y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy g(x)h(y)f_X(x)f_Y(y) = \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \cdot \int_{-\infty}^{\infty} h(y)f_Y(y)dy = E[g(X)] \cdot E[h(Y)]. \end{aligned}$$

## 9.5 Распределения сумм

**Теорема 1.** Распределение суммы случайных величин  $X$  и  $Y$  даётся выражением

$$p_{X+Y}(z) = \sum_y p_{X,Y}(z - y, y) \quad (\text{дискретный случай})$$

$$f_{X,Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z - y, y)dy \quad (\text{непрерывный случай}).$$

**Теорема 2.** Распределение суммы *независимых* случайных величин  $X$  и  $Y$  даётся выражением

$$p_{X+Y}(z) = \sum_y p_X(z - y)p_Y(y) \quad (\text{дискретный случай})$$

$$f_{X,Y}(z) = \int_{-\infty}^{\infty} f_X(z - y)f_Y(y)dy \quad (\text{непрерывный случай}).$$

**Замечания.**

1. Если независимые непрерывные случайные величины являются неотрицательными, то

$$f_{X+Y}(z) = \int_0^z f_X(z - y)f_Y(y)dy.$$

2. Эти математические операции имеют специальное название - *свёртки распределений* - и специальное обозначение:

$$\int_0^z f_X(z - y)f_Y(y)dy \equiv f_X * f_Y(z).$$

## 10 Лекция 9. Суммирование случайных величин

### 10.1 Математическое ожидание суммы

Используя теорему из 8.1 с  $g(X, Y) = X + Y$  и маргинальными распределениями, мы получим:

$$\begin{aligned} E(X + Y) &= \sum_{\{x\}} \sum_{\{y\}} (x + y) f(x, y) = \sum_{\{x\}} \sum_{\{y\}} x f(x, y) + \sum_{\{x\}} \sum_{\{y\}} y f(x, y) = \\ &= \sum_{\{x\}} x \left( \sum_{\{y\}} f(x, y) \right) + \sum_{\{y\}} y \left( \sum_{\{x\}} f(x, y) \right) = \sum_{\{x\}} x f_X(x) + \sum_{\{y\}} y f_Y(y) = EX + EY. \end{aligned}$$

Мы доказали, что м.о. суммы *любых двух случайных величин* равно сумме математических ожиданий каждой из них:

$$E(X + Y) = EX + EY.$$

Более общий случай:

$$E(aX + bY) = aEX + bEY$$

где  $a$  и  $b$  - произвольные числа.

### 10.2 Дисперсия суммы

Дисперсия, определяемая как  $\sigma_X^2 = E(X - \mu_X)^2$ , может быть записана в виде  $\sigma_x^2 = E \overset{\circ}{X}^2$ , где  $\overset{\circ}{X} \equiv X - \mu_X$  - центрированная случайная величина  $X$ . Рассмотрим дисперсию суммы  $Z = X + Y$ :

$$\sigma_{X+Y}^2 = E(\overset{\circ}{X} + \overset{\circ}{Y})^2 = \sigma_X^2 + \sigma_Y^2 + 2E(\overset{\circ}{X}\overset{\circ}{Y}).$$

**Определение 1.** Величина  $E(\overset{\circ}{X}\overset{\circ}{Y})$  называется *ковариацией случайных величин  $X$  и  $Y$* , и обозначается  $\sigma_{XY}$

$$\sigma_{XY} = E(\overset{\circ}{X}\overset{\circ}{Y}) = E[(X - \mu_X)(Y - \mu_Y)].$$

**Теорема 1.** Дисперсия суммы двух случайных величин  $X$  и  $Y$  определяется выражением  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$ .

### 10.3 Свойства дисперсии и ковариации

- 1)  $\sigma_{X,X} = \sigma_X^2$ .
- 2)  $\sigma_{XY} = E(XY) - (EX)(EY)$ , для дисперсии  $\sigma_X^2 = E(X^2) - (EX)^2$ .
- 3)  $\sigma_{X+a,Y+b} = \sigma_{X,Y}$ , для дисперсии  $\sigma_{X+a}^2 = \sigma_X^2$ .
- 4)  $\sigma_{aX,bY} = ab\sigma_{X,Y}$ , для дисперсии  $\sigma_{aX}^2 = a^2\sigma_X^2$ .
- 5) Если  $X$  и  $Y$  независимы, тогда  $\sigma_{X,Y} = 0$  и  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ .

**Замечание.** Следует заметить, что независимость случайных величин подразумевает нулевую ковариацию, но обратное утверждение не верно: *нулевая ковариация не обязательно подразумевает их независимость*.

Пример.

## 10.4 Коэффициент корреляции

**Определение 1.** Коэффициент корреляции случайных величин  $X$  и  $Y$  определяется как

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

**Замечание.** Коэффициент корреляции не зависит от размерности величин  $X$  и  $Y$  и удовлетворяет неравенству  $-1 \leq \rho_{X,Y} \leq 1$ .

**Теорема 1** Если  $X$  и  $Y$  связаны между собой линейным соотношением  $Y = aX + b$  тогда  $\rho_{X,Y} = 1$  если  $a > 0$  и  $\rho_{X,Y} = -1$  если  $a < 0$ .

**Доказательство.**

По определению,

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{E(\overset{\circ}{X}\overset{\circ}{Y})}{\sqrt{E\overset{\circ}{X}^2 \cdot E\overset{\circ}{Y}^2}}.$$

Учитывая, что

$$\overset{\circ}{Y} = Y - EY = aX - aEX = a\overset{\circ}{X}$$

мы получаем

$$\rho_{X,Y} = \frac{aE\overset{\circ}{X}^2}{|a|E\overset{\circ}{X}^2} = \frac{a}{|a|} = \begin{cases} 1, & a > 0, \\ -1, & a < 0. \end{cases}$$

## 11 Лекция 10. Дискретные распределения вероятностей

### 11.1 Распределение Бернулли

Мы начали наш курс с подбрасывания монеты. Обобщение этого эксперимента носит название испытания Бернулли.

**Определение 1.** Испытания Бернулли - это эксперимент с двумя возможными исходами, которые называются *успех* (*success*)( $s$ ) и *неудача* (*failure*)( $f$ ). Если  $s$  имеет место с вероятностью  $p$ , тогда  $f$  имеет место с вероятностью  $1 - p$ , поскольку полная вероятность должна равняться 1.

Связывая 1 с  $s$  и 0 с  $f$  мы получим случайную величину Бернулли  $X$  с распределением

$$p(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1. \end{cases} \quad (1)$$

**Замечание.** Нетрудно проверить, что это распределение может быть представлено в форме:

$$p(x) = p^x(1 - p)^{1-x}, \quad x = 0, 1.$$

**Определение 2.** Распределение (1) называется *распределением Бернулли с параметром  $p$* , а соответствующая случайная величина носит название *случайной величины Бернулли*.

Graphs of the probability distribution and of the cumulative distribution. Computing the mean and the variance.





Для  $n = 1$ :

$$b(0; 1, p) = P(f) = (1 - p) = \binom{1}{0} p^0 (1 - p)^{1-0},$$

$$b(1; 1, p) = P(s) = p = \binom{1}{1} p^1 (1 - p)^{1-1}.$$

Для  $n = 2$ :

$$b(0; 2, p) = P(ff) = P(f)P(f) = (1 - p)(1 - p) = (1 - p)^2 = \binom{2}{0} p^0 (1 - p)^{2-0},$$

$$\begin{aligned} b(1; 2, p) &= P(fs \text{ or } sf) = P(fs) + P(sf) = P(f)P(s) + P(s)P(f) \\ &= (1 - p)p + p(1 - p) = 2p(1 - p) = \binom{2}{1} p^1 (1 - p)^{2-1}, \end{aligned}$$

$$b(2; 2, p) = P(ss) = P(s)P(s) = pp = p^2 = \binom{2}{2} p^2 (1 - p)^{2-2}.$$

Случаи при  $n > 2$  можно получить по индукции.

Среднее значение и дисперсию можно найти напрямую из распределения

$$EY_n = \mu = \sum_{x=0}^n x b(x; n, p) = \sum_{x=0}^n x \binom{n}{x} p^x (1 - p)^{n-x} = np,$$

$$\sigma_{Y_n}^2 = \sum_{x=0}^n (x - \mu)^2 b(x; n, p) = \sum_{x=0}^n (x - \mu)^2 \binom{n}{x} p^x (1 - p)^{n-x} = np(1 - p),$$

но самый короткий путь - использовать связь со случайной величиной Бернулли:

$$EY_n = E \sum_{j=1}^n X_j = \sum_{j=1}^n EX_j = \sum_{j=1}^n p = np,$$

$$\sigma_{Y_n}^2 = \sigma_{X_1 + \dots + X_n}^2 = \sigma_{X_1}^2 + \dots + \sigma_{X_n}^2 = n\sigma_{X_1}^2 = np(1 - p).$$

Биномиальное распределение
Функция $p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$
Значения $[0, 1, \dots, n]$
Параметры $0 \leq p \leq 1; n = 1, 2, 3, \dots$
Среднее значение $np$
Дисперсия $np(1 - p)$

## 11.4 Геометрическое распределение

Вновь рассматривая независимые испытания Бернулли, будем искать *распределение первого успешного номера*  $N_1$ . Очевидно,

$$P(N_1 = 1) = P(s) = p,$$

$$P(N_1 = 2) = P(fs) = P(f)P(s) = (1 - p)p,$$

$$P(N_1 = 3) = P(ffs) = P(f)P(f)P(s) = (1 - p)^2 p,$$

и так далее.

Результаты представлены в таблице.

Геометрическое распределение
Функция $p(x) = p(1-p)^{x-1}$
Значения $1, 2, 3, \dots$
Параметр $p$
Среднее значение $\frac{1}{p}$
Дисперсия $\frac{1-p}{p^2}$

Распределение номера  $N_m$   $m$ -го успеха в серии испытаний Бернулли называется *отрицательным биномиальным распределением* и даётся следующей таблицей:

Отрицательное биномиальное распределение
Функция $p(x) = \binom{x-1}{m-1} p^m (1-p)^{x-m}$
Значения $m, m+1, m+2, \dots$
Параметры $m = 1, 2, 3; 0 < p \leq 1$
Среднее значение $\frac{m}{p}$
Дисперсия $m \frac{1-p}{p^2}$

**Замечание:** Случайная величина  $N_m$  может рассматриваться как сумма из  $m$  независимых случайных величин  $N_1$ , так что  $EN_m = mEN_1$  и  $\sigma_{N_m}^2 = m\sigma_{N_1}^2$ .

## 11.5 Распределение Пуассона

Умножая ряд

$$e^\mu = 1 + \mu + \frac{\mu^2}{2} + \dots = \sum_{n=0}^{\infty} \frac{\mu^n}{n!}$$

на  $e^{-\mu}$  мы получим

$$\sum_{n=0}^{\infty} \frac{\mu^n}{n!} e^{-\mu} = 1.$$

Это может быть интерпретировано, как условие нормировки для дискретного распределения

$$p(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \dots$$

называемого *распределением Пуассона*. Вычислим среднее значение соответствующей случайной величины:

$$\sum_{x=0}^{\infty} x p(x) = \sum_{x=0}^{\infty} x \frac{\mu^x}{x!} e^{-\mu} = \sum_{x=1}^{\infty} \frac{\mu^x}{(x-1)!} e^{-\mu} = \mu \left( \sum_{n=0}^{\infty} \frac{\mu^n}{n!} \right) e^{-\mu} = \mu.$$

Похожие вычисления показывают, что

$$\sigma^2 = \sum_{x=0}^{\infty} x^2 f(x) - \mu^2 = \mu.$$

Распределение Пуассона - однопараметрическое распределение с параметром  $\mu$ , который одновременно является его *средним значением и дисперсией*.

Распределение Пуассона
Функция $p(x) = \frac{\mu^x}{x!} e^{-\mu}$
Значения $0, 1, 2, \dots$
Параметр $\mu > 0$
Среднее значение $\mu$
Дисперсия $\mu$

Следующая теорема, носящая название **теоремы Пуассона**, связывает распределение Пуассона и биномиальное распределение:

$$\lim_{n \rightarrow \infty} b(x; n, \mu/n) = \frac{\mu^x}{x!} e^{-\mu}.$$

**Доказательство:**

$$\begin{aligned} b(x; n, \mu/n) &= \binom{n}{x} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} = \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} = \\ &= \frac{\mu^x}{x!} \frac{[1 \cdot 2 \dots (n-x)](n-x+1) \dots n}{[1 \cdot 2 \dots (n-x)]n^x} \left(1 - \frac{\mu}{n}\right)^{n-x} = \\ &= \frac{\mu^x}{x!} \left[ \frac{(n-x+1) \dots n}{n \dots n} \right] \left(1 - \frac{\mu}{n}\right)^{n-x} = \\ &= \frac{\mu^x}{x!} \left[ \left(\frac{n-x+1}{n}\right) \dots \left(\frac{n}{n}\right) \right] \left(1 - \frac{\mu}{n}\right)^{-x} \left(1 - \frac{\mu}{n}\right)^n. \end{aligned}$$

Учитывая, что для  $n \rightarrow \infty$

$$\left[ \left(\frac{n-x+1}{n}\right) \dots \left(\frac{n}{n}\right) \right] \rightarrow 1,$$

$$\left(1 - \frac{\mu}{n}\right)^{-x} \rightarrow 1$$

и

$$\left(1 - \frac{\mu}{n}\right)^n \rightarrow e^{-\mu}$$

мы получим

$$b(x; n, \mu/n) \rightarrow \frac{\mu^x}{x!} e^{-\mu}, \quad n \rightarrow \infty.$$

## 12 Лекция 11. Типичные задачи

**Задача 1.** Совместное распределение вероятностей представлено в таблице:

$x =$	-1	0	1
$y = -1$	0	1/6	0
$y = 0$	1/3	0	1/3
$y = 1$	0	1/6	0

Чтобы найти маргинальные распределения, необходимо просуммировать  $p(x) = \sum_y p(x, y)$  и  $p(y) = \sum_x p(x, y)$ :

$x =$	-1	0	1	$p(y)$
$y = -1$	0	1/6	0	1/6
$y = 0$	1/3	0	1/3	2/3
$y = 1$	0	1/6	0	1/6
$p(x)$	1/3	1/3	1/3	

Используя эти маргинальные распределения можно найти средние значения и дисперсии величин  $X$  и  $Y$ :

$$\mu_X = EX = \sum xp(x) = (-1) \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0,$$

$$\sigma_X^2 = EX^2 - \mu_X^2 = EX^2 = \sum x^2 p(x) = (-1)^2 \cdot \frac{1}{3} + 0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} = \frac{2}{3};$$

$$\mu_Y = EY = \sum yp(y) = (-1) \cdot \frac{1}{6} + 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{6} = 0,$$

$$\sigma_Y^2 = EY^2 - \mu_Y^2 = EY^2 = \sum y^2 p(y) = (-1)^2 \cdot \frac{1}{6} + 0^2 \cdot \frac{2}{3} + 1^2 \cdot \frac{1}{6} = \frac{1}{3}.$$

Чтобы найти условные распределения, нужно разделить  $p(x, y)$  на соответствующие маргинальные распределения  $p(x|y) = p(x, y)/p(y)$ :

$x =$	-1	0	1
$y = -1$	0	1	0
$y = 0$	1/2	0	1/2
$y = 1$	0	1	0

и  $p(y|x) = p(x, y)/p(x)$ :

$x =$	-1	0	1
$y = -1$	0	1/2	0
$y = 0$	1	0	1
$y = 1$	0	1/2	0

Обратите внимание, что  $\sum_x p(x|y) = 1$  и  $\sum_y p(y|x) = 1$ , но суммы по условным значениям не равны 1.

Полученные условные распределения зависят от условных значений, таким образом  $X$  и  $Y$  не являются независимыми.

Чтобы найти распределение суммы  $Z = X + Y$  случайных величин, необходимо прежде всего определить область её возможных значений. В нашем случае  $z = \{-1, 0, 1\}$  (другие значения имеют нулевую вероятность). Тогда, выбирая некоторые возможные значения  $z$ , скажем  $z = -1$ , отметим соответствующие исходы в таблице  $p(x, y)$

$x =$	-1	0	1
$y = 0$	0	<b>1/6</b>	0
1	<b>1/3</b>	0	1/3
	0	1/6	0

и найдём сумму:

$$P(Z = -1) = p_{X+Y}(-1) = p_{XY}(0, -1) + p_{XY}(-1, 0) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.$$

Подобным же образом,

$$P(Z = 0) = p_{X+Y}(0) = 0$$

и

$$P(Z = 1) = p_{X+Y}(1) = \frac{1}{2}.$$

Наконец, вычислим ковариацию и коэффициент корреляции. По определению

$$\text{Cov}(X, Y) \equiv \sigma_{XY} = E[(X - \mu_x)(Y - \mu_y)].$$

Обычно более удобно использовать для вычислений эквивалентную формулу

$$\sigma_{XY} = E[XY] - \mu_X \mu_Y,$$

хотя в нашем случае обе формулы имеют один и тот же вид, поскольку  $\mu_X = \mu_Y = 0$ :

$$\sigma_{XY} = E[XY].$$

Согласно теореме из 8.1, имеем ( смотри таблицу для  $f(x, y)$ )

$$\sigma_{XY} = \sum_x \sum_y xyp(x, y) = 0$$

и следовательно

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = 0.$$

Таким образом случайные величины  $X$  и  $Y$  являются некоррелированными, но не являются независимыми.

### Задача 2.

Рассмотрим теперь треугольник с углами в точках  $(0,0)$ ,  $(0,1)$ ,  $(1,0)$  и будем считать, что вероятность распределена внутри него непрерывно и равномерно. Отметим, что для любого фиксированного  $x \in [0, 1]$  величина  $y \in [0, 1 - x]$ , а следовательно, маргинальная плотность даётся выражением

$$f(x) = \int_0^{1-x} f(x, y) dy.$$

Поскольку распределение равномерное

$$f(x, y) = \text{const} = \frac{1}{\text{Area of } \triangle} = 2.$$

Подставляя выражения под знак интеграла, получим:

$$f(x) = \int_0^{1-x} 2dy = 2(1 - x), \quad 0 < x < 1.$$

Среднее значение величины  $X$

$$\mu_X = \int_0^1 x f(x) dx = \int_0^1 x \cdot 2(1-x) dx = 2 \left( \frac{1}{2} - \frac{1}{3} \right) = \frac{1}{3}.$$

Дисперсия

$$\sigma_X^2 = \int_0^1 x^2 \cdot 2(1-x) dx - \mu_X^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}.$$

Условная плотность

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad 0 \leq y \leq 1-x.$$

Заметим, что

$$\int_0^{y_{\max}} f(y|x) dy = \int_0^{1-x} \frac{1}{1-x} dy = 1.$$

Из-за симметрии распределения,  $\mu_Y = \mu_X$ ,  $\sigma_Y^2 = \sigma_X^2$  и

$$\begin{aligned} \sigma_{XY} &\equiv \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} xy f(x,y) dy - \mu_X \mu_Y = \int_0^1 dx \int_0^{1-x} xy 2 dy - \frac{1}{9} = \\ &= \int_0^1 x(1-2x+x^2) dx = \frac{1}{2} - 2 \cdot \frac{1}{3} + \frac{1}{4} - \frac{1}{9} = -\frac{1}{36}. \end{aligned}$$

Коэффициент корреляции

$$\rho_{XY} = \frac{-\frac{1}{36}}{\frac{1}{18}} = -\frac{1}{2}.$$

Он отрицательный, поэтому случайные величины антикоррелируют.

Предположим, требуется найти  $f_{X+Y}(z)$ . Для этого удобно сначала вычислить кумулятивную функцию

$$F_{X+Y}(z) \equiv P(X+Y \leq z) = \int \int_{x+y < z} f(x,y) dx dy,$$

а затем её продифференцировать:

$$f_{X+Y}(z) = \frac{dF_{X+Y}(z)}{dz}.$$

В рассматриваемом случае, благодаря равномерности распределения

$$F_{X+Y}(z) = P(X+Y \leq z) = (\text{Area of } \triangle_z) / (\text{Area of } \triangle_1) = z^2$$

откуда

$$f_{X+Y}(z) = 2z, \quad 0 \leq z \leq 1.$$

**Задача 3.** Предположим, вы купили 5 лотерейных билетов. Допустим, вероятность выигрыша некоторого приза равна  $1/20$  для каждого билета и все эти билеты являются независимыми. Может случиться, что ни один из этих билетов не выиграет, возможно,

выиграет только один билет, но возможно, выиграют все 5 билетов. Но какова вероятность этого события? Она даётся биномиальным распределением:

$$b(x; 5, 0.05) = \binom{5}{x} p^x (1-p)^{5-x} = \binom{5}{x} 0.05^x 0.95^{5-x} = \frac{5!}{x!(5-x)!}, \quad x = 0, 1, 2, 3, 4, 5.$$

Теперь предположим, что вы покупаете билеты моментальной лотереи до тех пор, пока не попадётся выигрышный. Как в этом случае распределён номер билета, который вы купили? Согласно **геометрическому распределению**

$$p(x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \dots$$

Но если вы прекратите покупать билеты лотереи после второго или третьего выигрыша, какое распределение описывает случайный номер билета в этом случае? Это будет **отрицательное биномиальное распределение** с параметрами  $m = 2$  или  $3$  соответственно. Заметьте, что геометрическое распределение - ничто иное, как отрицательное биномиальное распределение с параметром  $m = 1$ .

Отрицательное биномиальное распределение
Функция $p(x) = \binom{x-1}{m-1} p^m (1-p)^{x-m}$
Значения $m, m+1, m+2, \dots$
Параметры $m = 1, 2, 3; 0 < p \leq 1$
Среднее значение $\frac{m}{p}$
Дисперсия $m \frac{1-p}{p^2}$

**Задача 4.** Распределение Пуассона - это однопараметрическое дискретное распределение:

$$p(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \dots$$

Здесь параметр  $\mu$  - среднее значение и дисперсия одновременно. Если известно  $\mu$ , то можно вычислить любую вероятность связанную с распределением Пуассона. You can also find binomial and Poisson probabilities in tables A.1 and A.2 in our main textbook (pages 661-668).

**Замечание.** Для вычисления бесконечных сумм, подобных  $\sum_{x=n}^{\infty} p(x)$  нет необходимости находить предел, гораздо проще использовать условие нормировки:

$$\sum_{x=0}^{\infty} p(x) = 1,$$

откуда

$$\sum_{x=0}^{n-1} p(x) + \sum_{x=n}^{\infty} p(x) = 1$$

и

$$\sum_{x=n}^{\infty} p(x) = 1 - \sum_{x=0}^{n-1} p(x).$$

Теперь нужно вычислить конечную сумму.



## STAT 312 Spring

### Домашняя работа 3

(must be returned on March 1, 12:30 p.m., CLPP 108)

**Задача 1.** Вероятность равномерно распределена по точкам с целочисленными координатами  $x, y$ , которые подчиняются условиям  $x, y \geq 1, x + y \leq 5$ .

**Найти:** (a)  $f(x, y)$ ; (b)  $f(x)$ ; (c)  $f(y|x)$ ; (d)  $f_{X+Y}(z)$ ; (e)  $EX$  and  $EY$ ; (f)  $\sigma_X^2$  и  $\sigma_Y^2$ ; (g)  $\sigma_{XY}$  и  $\rho_{XY}$ . (h) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Задача 2.** Вероятность равномерно и непрерывно распределена внутри треугольника с вершинами  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 2)$  (первая координата  $x$ , вторая -  $y$ ).

**Найти:** (a)  $f(x, y)$ ; (b)  $f(x)$ ; (c)  $f(y|x)$ ; (d) кумулятивную функцию распределения  $F(x)$ ; (e)  $F(y|x)$ ; (f)  $P(X + Y < z)$ ; (g)  $f_{X+Y}(z)$ ; (h)  $EX$  и  $EY$ ; (i)  $\sigma_X^2$  и  $\sigma_Y^2$ ; (j)  $\sigma_{XY}$  и  $\rho_{XY}$ . (k) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Задача 3.** Дано:  $f(x, y) = C(x + y^2 + 1)$  при  $0 < x < 1, 0 < y < 1$ , **найти:** (a) постоянную  $C$ ; (b) маргинальные плотности  $f(x)$  и  $f(y)$ ; (c) условные плотности  $f(x|y)$  и  $f(y|x)$ ; (d) средние значения и дисперсии величин  $X$  и  $Y$ ; (e) их ковариацию и коэффициент корреляции, (f) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Задача 4.** Предположим, что известна совместная плотность вероятности для  $X$  и  $Y$

$$f(x, y) = \begin{cases} 2e^{-x-2y} & \text{for } x > 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

**Найти:** (a) маргинальные плотности; (b) средние значения и дисперсии для  $X$  и  $Y$ ; (c) обе условные плотности; (d) ковариацию и коэффициент корреляции. (e) являются ли  $X$  и  $Y$  независимыми или нет?

**Задача 5.** Дана совместная плотность вероятности

$$f(x, y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

**Найти:** (a) маргинальные плотности; (b) ожидаемые значения величин  $X$  и  $Y$ ; (c) их дисперсии; (d) обе условные плотности; (e) ковариацию и коэффициент корреляции. (f) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Задача 6.** Предположим, вы купили 10 билетов моментальной лотереи. Допустим, вероятность выигрыша равна  $1/9$  для каждого билета и все эти билеты являются независимыми.

**Найти** вероятность, что (a) ни один из билетов не окажется выигрышным; (b) только один из билетов окажется выигрышным; (c) два или более билетов окажутся выигрышными.

**Задача 7.** Предположим, что вероятность выигрыша по билету моментальной лотереи равна 0.1. Вы покупаете билеты до тех пор пока не выиграете три приза.

**Найти** вероятность, что (a) вам придётся купить не более 5 билетов; (b) число билетов, которые вам придётся купить окажется между 6 и 10 включительно (то есть  $6 \leq N_3 \leq 10$ ).

**Задача 8.** Запросы на бронирование мест на катере для спортивной рыбалки на некоторую дату подчиняются распределению Пуассона с параметром  $\mu$ . Зная, что вероятность отсутствия запросов на эту дату равна 0.25, **найти** вероятность (a) отсутствия запросов; (b) получения двух запросов; (2) получения двух или более запросов.

# STAT 312 Spring

## Домашняя работа 3 - решения

**Задача 1.** Вероятность равномерно распределена по точкам с целочисленными координатами  $x, y$ , которые подчиняются условиям  $x, y \geq 1, x + y \leq 5$ .

**Найти:** (a)  $f(x, y)$ ; (b)  $f(x)$ ; (c)  $f(y|x)$ ; (d)  $f_{X+Y}(z)$ ; (e)  $EX$  and  $EY$ ; (f)  $\sigma_X^2$  и  $\sigma_Y^2$ ; (g)  $\sigma_{XY}$  и  $\rho_{XY}$ . (h) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Решение.**

(a) Нарисуем график или таблицу. Число точек равно 10, таким образом совместное распределение

$$p(x, y) = \frac{1}{10} = 0.1, \quad x, y \geq 1, \quad x + y \leq 5.$$

Таблица  $p(x, y)$  :

$x =$	1	2	3	4
$y=1$	0.1	0.1	0.1	0.1
2	0.1	0.1	0.1	0
3	0.1	0.1	0	0
4	0.1	0	0	0

(b) Маргинальное распределение  $p_X(x)$  получается путём суммирования  $p(x, y)$  по столбцам:

$$p_X(1) = \sum_{y=1}^4 p(1, y) = \frac{4}{10} = 0.4,$$

$$p_X(2) = \sum_{y=1}^3 p(2, y) = \frac{3}{10} = 0.3,$$

$$p_X(3) = \sum_{y=1}^2 p(3, y) = \frac{2}{10} = 0.2,$$

$$p_X(4) = \sum_{y=1}^1 p(4, y) = \frac{1}{10} = 0.1.$$

(c) Чтобы найти  $p(y|x)$ , каждое значение  $p(x, y)$  нужно разделить на соответствующее  $p(x)$ . Результаты представлены в таблице  $p(y|x)$ :

$x =$	1	2	3	4
$y=1$	1/4	1/3	1/2	1
2	1/4	1/3	1/2	0
3	1/4	1/3	0	0
4	1/4	0	0	0

Вы можете проверить результаты, сумма  $p(y|x)$  по каждому столбцу (то есть по  $y$ ) должна быть равна 1.

(d) Сначала нужно найти диапазон значений  $Z = X + Y : \{2, 3, 4, 5\}$ . Очевидно, что  $Z = 2$  наступает только если  $X = 1$  и  $Y = 1$ , то есть с вероятностью  $p(1, 1) = 0.1$  (см. таблицу для  $p(x, y)$ ). Событие  $Z = 3$  включает два исхода. Отмечая соответствующие вероятности в таблице  $p(x, y)$

$x =$	1	2	3	4
$y=1$	0.1	<b>0.1</b>	0.1	0.1
2	<b>0.1</b>	0.1	0.1	0
3	0.1	0.1	0	0
4	0.1	0	0	0

и суммируя их, получаем:  $p_Z(3) = p(1, 2) + p(2, 1) = 0.2$ . Аналогично  $p_Z(4) = 0.3$  и  $p_Z(5) = 0.4$ .

(е)  $\mu_X = EX = \sum_x p(x) = 1 \cdot 0.4 + 2 \cdot 0.3 + 3 \cdot 0.2 + 4 \cdot 0.1 = 2$ . Благодаря симметрии распределения по отношению к  $x$  и  $y$ ,  $EY = 2$ .

(ф)  $\sigma_X^2 = \sum_x p(x) - \mu_X^2 = 1^2 \cdot 0.4 + 2^2 \cdot 0.3 + 3^2 \cdot 0.2 + 4^2 \cdot 0.1 - 2^2 = 5 - 4 = 1$ ;  $\sigma_Y^2 = 1$ .

(г) Используем следующую формулу для вычисления ковариации:

$$\sigma_{XY} = \sum_x \sum_y xyp(x, y) - \mu_x \mu_y.$$

Для вычисления ожидаемого значения произведения  $XY$  (первый член) удобно составить таблицу для  $xyp(x, y)$

$x =$	1	2	3	4
$y=1$	$1 \cdot 1 \cdot 0.1$	$2 \cdot 1 \cdot 0.1$	$3 \cdot 1 \cdot 0.1$	$4 \cdot 1 \cdot 0.1$
2	$1 \cdot 2 \cdot 0.1$	$2 \cdot 2 \cdot 0.1$	$3 \cdot 2 \cdot 0.1$	0
3	$1 \cdot 3 \cdot 0.1$	$2 \cdot 3 \cdot 0.1$	0	0
4	$1 \cdot 4 \cdot 0.1$	0	0	0

и затем просуммировать эти числа:

$$\sum_x \sum_y xyp(x, y) = (10 + 12 + 9 + 4) \cdot 0.1 = 35 \cdot 0.1 = 3.5.$$

В результате получим

$$\sigma_{XY} = 3.5 - 4 = -0.5, \quad \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = -\frac{0.5}{1} = -0.5.$$

(h) Случайные величины  $X$  и  $Y$  не являются независимыми: таблица  $p(y|x)$  показывает, что условное распределение  $Y$  зависит от значений-условий  $x$  величины  $X$ .

**Задача 2.** Вероятность равномерно и непрерывно распределена внутри треугольника с вершинами  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 2)$  (первая координата  $x$ , вторая -  $y$ ).

**Найти:** (а)  $f(x, y)$ ; (б)  $f(x)$ ; (с)  $f(y|x)$ ; (д) кумулятивную функцию распределения  $F(x)$ ; (е)  $F(y|x)$ ; (ф)  $P(X + Y < z)$ ; (г)  $f_{X+Y}(z)$ ; (h)  $EX$  и  $EY$ ; (и)  $\sigma_X^2$  и  $\sigma_Y^2$ ; (j)  $\sigma_{XY}$  и  $\rho_{XY}$ . (k) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Решение.** (см. Лекцию 14)

**Задача 3.** Дано:  $f(x, y) = C(x + y^2 + 1)$  при  $0 < x < 1$ ,  $0 < y < 1$ , **найти:** (а) постоянную  $C$ ; (б) маргинальные плотности  $f(x)$  и  $f(y)$ ; (с) условные плотности  $f(x|y)$  и  $f(y|x)$ ; (д) средние значения и дисперсии величин  $X$  и  $Y$ ; (е) их ковариацию и коэффициент корреляции, (ф) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Решение.**

(а) Постоянную следует находить из условия нормировки:

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) = 1.$$

Подставляя сюда явное выражение для совместной вероятности и выполняя интегрирование

$$C \int_0^1 dx \int_0^1 dy (x + y^2 + 1) = C \int_0^1 dx (x + 1) \int_0^1 dy + C \int_0^1 dx \int_0^1 y^2 dy = C \left[ \left( \frac{1}{2} + 1 \right) + \frac{1}{3} \right] = C \frac{11}{6} = 1,$$

получим

$$C = \frac{6}{11}.$$

(b) Маргинальные плотности:

$$f(x) = \int_0^1 f(x, y) dy = (6/11) \int_0^1 (x + y^2 + 1) dy = (2/11)(3x + 4)$$

и

$$f(y) = \int_0^1 f(x, y) dx = (6/11) \int_0^1 (x + y^2 + 1) dx = (3/11)(2y^2 + 3).$$

(с) Условные плотности:

$$f(y|x) = \frac{f(x, y)}{f(x)} = 3 \frac{x + y^2 + 1}{3x + 4}$$

и

$$f(x|y) = \frac{f(x, y)}{f(y)} = 2 \frac{x + y^2 + 1}{3 + 2y^2}.$$

(d) Средние значения вычисляются с использованием маргинальных плотностей (b)

$$\mu_X = \int_0^1 x f(x) dx = (2/11) \int_0^1 x(3x + 4) dx = \frac{6}{11},$$

$$\mu_Y = \int_0^1 y f(y) dy = (3/11) \int_0^1 y(2y^2 + 3) dy = \frac{13}{22}.$$

Средние квадраты

$$EX^2 = \int_0^1 x^2 f(x) dx = (2/11) \int_0^1 x^2(3x + 4) dx = \frac{25}{66},$$

$$EY^2 = \int_0^1 y^2 f(y) dy = (3/11) \int_0^1 y^2(2y^2 + 3) dy = \frac{21}{55}.$$

Дисперсии

$$\sigma_X^2 = EX^2 - \mu_X^2 = \frac{25}{66} - \left(\frac{6}{11}\right)^2 =$$

и

$$\sigma_Y^2 = EY^2 - \mu_Y^2 = \frac{21}{55} - \left(\frac{13}{22}\right)^2 =$$

(е)

$$E(XY) = (6/11) \int_0^1 dx x \int_0^1 dy y(x+y^2+1) = (6/11) \left[ \int_0^1 dx x(x+1) \int_0^1 dy y + \int_0^1 dx x \int_0^1 dy y^3 \right] = \frac{13}{44}.$$

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{13}{44} - \frac{6}{11} \cdot \frac{13}{22} =$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} =$$

(f) Величины  $X$  и  $Y$  не являются независимыми: их условные распределения зависят от значений-условий.

**Задача 4.** Предположим, что известна совместная плотность вероятности для  $X$  и  $Y$

$$f(x, y) = \begin{cases} 2e^{-x-2y} & \text{for } x > 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

**Найти:** (а) маргинальные плотности; (b) средние значения и дисперсии для  $X$  и  $Y$ ; (с) обе условные плотности; (d) ковариацию и коэффициент корреляции. (е) являются ли  $X$  и  $Y$  независимыми или нет?

**Решение.**

(а) Маргинальные плотности:

$$f(x) = \int_0^\infty 2e^{-x-2y} dy = e^{-x}, \quad x > 0; \quad f(y) = \int_0^\infty 2e^{-x-2y} dx = 2e^{-2y}, \quad y > 0.$$

(b) Средние значения

$$\mu_X = \int_0^\infty x e^{-x} dx = 1, \quad \mu_Y = \int_0^\infty 2y e^{-2y} dy = \frac{1}{2} \int_0^\infty z e^{-z} dz = \frac{1}{2}.$$

Вспомним, что  $\int_0^\infty z^n e^{-z} dz = n!$  Дисперсии

$$\sigma_X^2 = 1, \quad \sigma_Y^2 = \frac{1}{4}.$$

(с) Условные плотности

$$f(x|y) = e^{-x}, \quad x > 0; \quad f(y|x) = 2e^{-2y}, \quad y > 0.$$

(d) and (e) Из вышеизложенного видно, что случайные величины  $X$  и  $Y$  независимы, таким образом  $\sigma_{XY}$  и  $\rho_{XY}$  равны нулю.

**Задача 5.** Дана совместная плотность вероятности

$$f(x, y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

**Найти:** (a) маргинальные плотности; (b) ожидаемые значения величин  $X$  и  $Y$ ; (c) их дисперсии; (d) обе условные плотности; (e) ковариацию и коэффициент корреляции. (f) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

**Решение.**

(a) Нарисуем область на плоскости  $xy$ , где  $f(x, y)$  отлична от нуля. Маргинальные плотности:

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^{\infty} e^{-y} dy = e^{-x}, \quad x > 0,$$

$$f(y) = \int_0^y e^{-y} dx = e^{-y} \int_0^y dx = ye^{-y}, \quad y > 0.$$

(b) Средние значения:

$$\mu_X = \int_0^{\infty} x \cdot e^{-x} dx = 1, \quad \mu_Y = \int_0^{\infty} y \cdot ye^{-y} dy = 2.$$

(c) Дисперсии:

$$\sigma_X^2 = \int_0^{\infty} x^2 \cdot e^{-x} dx - \mu_X^2 = 2 - 1 = 1, \quad \sigma_Y^2 = \int_0^{\infty} y^2 \cdot ye^{-y} dy - \mu_Y^2 = 6 - 2 = 4.$$

(d) Условные плотности:

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, \quad y > x, \text{ and } 0 \text{ } y < x;$$

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y}, \quad 1 < x < y, \text{ and } 0 \text{ otherwise.}$$

(e) Во-первых, вычислим среднее значения произведения

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy xyf(x, y) = \int_0^{\infty} dy \int_0^y dx xye^{-y} = \int_0^{\infty} dy ye^{-y} \int_0^y dx x = \\ &= \int_0^{\infty} dy \frac{y^3}{2} e^{-y} = \frac{3!}{2} = 3. \end{aligned}$$

Затем

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = 3 - 2 = 1$$

и

$$\rho = \frac{1}{\sqrt{2}}.$$

(f) Нет, поскольку  $f_{XY}(x, y) \neq f_X(x)f_Y(y)$ .

**Задача 6.** Предположим, вы купили 10 билетов моментальной лотереи. Допустим, вероятность выигрыша равна  $1/9$  для каждого билета и все эти билеты являются независимыми.

**Найти** вероятность, что (а) ни один из билетов не окажется выигрышным; (б) только один из билетов окажется выигрышным; (с) два или более билетов окажутся выигрышными.

**Решение.** Вероятность даётся биномиальным распределением:

$$b(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 1, 2, \dots, n.$$

В нашем случае  $n = 10$  и  $p = 1/9$ . Вычисления дают

$$(a) \ b(0; 10, 1/9) = \frac{10!}{0!10!} (1/9)^0 (8/9)^{10-0} = (8/9)^{10} \approx 0.308.$$

$$(b) \ b(1; 10, 1/9) = \frac{10!}{1!9!} (1/9)^1 (8/9)^{10-1} = 10(1/9)(8/9)^9 \approx 0.385.$$

$$(c) \ P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - b(0; 10, 1/9) - b(1; 10, 1/9) \approx 1 - 0.308 - 0.385 = 0.307$$

**Задача 7.** Предположим, что вероятность выигрыша по билету моментальной лотереи равна 0.1. Вы покупаете билеты до тех пор пока не выиграете три приза.

**Найти** вероятность, что (а) вам придётся купить не более 5 билетов; (б) число билетов, которые вам придётся купить окажется между 6 и 10 включительно (то есть  $6 \leq N_3 \leq 10$ ).

**Решение.** Здесь мы должны использовать отрицательно биномиальное распределение

$$p(x) = \binom{x-1}{m-1} p^m (1-p)^{x-m} = \frac{(x-1)!}{(m-1)!(x-m)!} p^m (1-p)^{x-m}, \quad x = m, m+1, m+2, \dots \text{ with } m = 3 \text{ and } p = 0.1$$

Результаты:

$$(a) \ P(N_3 \leq 5) = P(N_3 = 3) + P(N_3 = 4) + P(N_3 = 5) = p(3) + p(4) + p(5) \approx 0.00856.$$

$$(b) \ P(6 \leq N_3 \leq 10) = P(N_3 = 6) + P(N_3 = 7) + P(N_3 = 8) + P(N_3 = 9) + P(N_3 = 10) = p(6) + p(7) + p(8) + p(9) + p(10) \approx 0.0612.$$

**Задача 8.** Запросы на бронирование мест на катере для спортивной рыбалки на некоторую дату подчиняются распределению Пуассона с параметром  $\mu$ . Зная, что вероятность отсутствия запросов на эту дату равна 0.25, **найти** вероятность (а) отсутствия запросов; (б) получения двух запросов; (в) получения двух или более запросов.

**Solution.** Распределение Пуассона

$$p(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \dots$$

- однопараметрическое распределение, его единственный параметр  $\mu$  имеет смысл среднего значения. В этой задаче оно может быть определено из условия

$$p(0) = \frac{\mu^0}{0!} e^{-\mu} = e^{-\mu},$$

и мы получаем:

$$\mu = -\ln(0.25) \approx 1.386,$$

(a)  $p(1) = \frac{1.386}{1!} 0.25 \approx 0.347,$

(b)  $p(2) = \frac{(1.386)^2}{2!} 0.25 \approx 0.240,$

(c)  $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) \approx 1 - 0.250 - 0.347 = 0.303.$



## 13 Лекция 12. Непрерывные плотности вероятности

### 13.1 Равномерная плотность

Обсуждение непрерывных распределений мы начнём с равномерного распределения в некотором интервале  $(a, b)$ . Мы не можем напрямую распространить на этот случай определение, данное для равномерного дискретного распределения: число точек в  $(a, b)$  бесконечно, и вероятность  $P(X = x) = \frac{1}{\infty} = 0$  для каждой точки, хотя полная вероятность  $P(X \in (a, b)) = 1$ . Согласие между этими двумя фактами может быть достигнуто путём введения вероятности события  $X \in (x_1, x_2)$ , пропорциональной длине интервала:

$$P(X \in (x_1, x_2)) = \frac{x_2 - x_1}{b - a}. \quad (1)$$

Очевидно, что для  $(x_1, x_2) = (b - a)$

$$P(X \in (a, b)) = \frac{b - a}{b - a} = 1$$

значение для  $x_1 \uparrow x, x_2 \downarrow x$

$$P(X = x) = \lim_{x_1 \uparrow x; x_2 \downarrow x} P(X \in (x_1, x_2)) = \lim_{x_2 - x_1 \rightarrow 0} \frac{x_2 - x_1}{b - a} = 0.$$

**Определение.** Непрерывная случайная величина  $X \in (a, b)$  называется *равномерно распределённой* в этом интервале, если равенство (1) имеет место для любых  $x_1, x_2 \geq x_1$ .

Соответствующие кумулятивная функция распределения и плотность вероятности имеют вид

$$F(x) = \frac{x - a}{b - a}, \quad f(x) = \frac{1}{b - a}, \quad a < x < b.$$

Среднее значение и дисперсия случайной величины  $X$  равномерно распределённой в  $(a, b)$  легко вычисляются:

$$\mu_X = \int_a^b x \frac{1}{b - a} dx = \frac{b^2 - a^2}{2(b - a)} = \frac{a + b}{2},$$

$$\sigma_X^2 = \int_a^b x^2 \frac{1}{b - a} dx - \mu_X^2 = \frac{b^3 - a^3}{3(b - a)} - \left(\frac{a + b}{2}\right)^2 = \frac{(b - a)^2}{12}.$$

Равномерная плотность
Функция $f(x) = \frac{1}{b-a}$
Значения $a < x < b$
Параметры $a, b$
Среднее значение $\frac{a+b}{2}$
Дисперсия $\frac{(b-a)^2}{12}$

При  $a = 0$  и  $b = 1$  равномерное распределение называется *стандартным равномерным распределением*. В этом случае

$$f(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 < x < 1, \\ 0, & x > 1, \end{cases}$$
$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 < x < 1, \\ 1, & x > 1, \end{cases}$$
$$\mu_X = \frac{1}{2}, \quad \sigma_X^2 = \frac{1}{12}.$$

## 13.2 Бета-плотность

Стандартная равномерная плотность является частным случаем семейства *Бета-плотностей* определяемого выражением:

$$f(x) = Cx^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

Постоянная  $C$  находится из условия нормировки:

$$\int_0^1 f(x)dx = C \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = 1.$$

Интеграл

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

называется *Бета-функцией* и обозначается  $B(\alpha, \beta)$ , так что  $C = 1/B(\alpha, \beta)$ .

Бета-плотность
Функция $f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$
Значения $0 < x < 1$
Параметры $\alpha > 0; \beta > 0$
Среднее значение $\frac{\alpha}{\alpha+\beta}$
Дисперсия $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Её среднее значение и другие моменты вычисляются с использованием основного свойства гамма-функции  $\Gamma(z+1) = z\Gamma(z)$ . Например:

$$\mu_X = \int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} = \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}.$$

## 13.3 Экспоненциальная плотность

Предположим, мы наблюдаем за поступлением телефонных вызовов на телефонную станцию ( police station????). Пусть первый звонок поступает в момент времени  $x = 0$ , а следующий поступает в случайный момент  $T$ . Вероятность, что  $T > x + \Delta x$  может быть представлена в виде

$$P(T > x + \Delta x) = P(T > x + \Delta x | T > x)P(T > x).$$

Заметим, что

$$P(T > x) = 1 - P(T < x) = 1 - F_T(x) \equiv \bar{F}(x).$$

Предполагая, что  $P(x < T < x + \Delta x | T > x) \approx a\Delta x$  для малых  $\Delta x$ , мы получим:

$$\bar{F}(x + \Delta x) - \bar{F}(x) = -a\Delta x \bar{F}(x).$$

Деля обе части на  $\Delta x$  и переходя к пределу, приходим к следующему дифференциальному уравнению:

$$\frac{d\bar{F}(x)}{dx} = -a\bar{F}(x)$$

с очевидным начальным условием  $\bar{F}(0) = 1$ . Это уравнение имеет решение

$$\bar{F}(x) = e^{-ax}, \quad x \geq 0.$$

теперь мы можем вычислить плотность вероятности:

$$f(x) = \frac{dF(x)}{dx} = \frac{d[1 - \bar{F}(x)]}{dx} = -\frac{d\bar{F}(x)}{dx} = ae^{-ax}, \quad x \geq 0,$$

называемой *экспоненциальной плотностью*. Прямые вычисления показывают, что  $\int_0^{\infty} xf(x)dx = 1/a$ , то есть единственный параметр распределения - это *обратное среднее значение* к  $T$ :  $a = 1/\mu_T$ .

Экспоненциальная плотность
Функция $f(x) = \frac{1}{\mu} e^{-x/\mu}$
Значения $x > 0$
Параметр $\mu > 0$
Среднее значение $\mu$
Дисперсия $\mu^2$

## 13.4 Гамма-плотность

Гамма-плотность описывает сумму независимых экспоненциально распределённых случайных величин.

Гамма-плотность
Функция $f(x) = \frac{1}{\Gamma(\alpha)\mu} \left(\frac{x}{\mu}\right)^{\alpha-1} e^{-x/\mu}$
Значения $x > 0$
Параметры $\alpha > 0; \mu > 0$
Среднее значение $\alpha\mu$
Дисперсия $\alpha\mu^2$

По определению, гамма-плотность обладает важным свойством - *воспроизводимостью*. Ясно, что если  $g_1(x)$  - гамма-плотность с параметрами  $(\mu, \alpha = n_1)$  и  $g_2(x)$  - гамма-плотность с параметрами  $(\mu, \alpha = n_2)$ , тогда их свёртка - гамма-плотность с параметрами  $(\mu, \alpha = n_1 + n_2)$ .

**Для дальнейшего чтения:** разделы 6.1, 6.6, 6.7 из основного учебника.

# 14 Лекция 13. Непрерывные плотности (окончание)

## 14.1 Нормальная плотность

Это самое знаменитое из вероятностных распределений, его называют также распределением Гаусса. Нормальная плотность имеет симметричный, колоколообразный вид и простирается от  $-\infty$  до  $\infty$ . Эта форма может быть описана с помощью экспоненциальной функции с квадратичным членом в показателе, и пиком в точке среднего значения  $\mu$

$$f(x) = A \exp\{-b(x - \mu)^2\}, \quad A > 0, b > 0.$$

Постоянные  $A$  и  $b$  определяются из условия нормировки и формулы для дисперсии:

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

$$\int_{-\infty}^{\infty} f(x)(x - \mu)^2 dx = \sigma^2.$$

Вычисления проделаем, производя замену переменных  $z = b(x - \mu)^2$  и используя гамма-функцию:

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt, \quad \Gamma(z) = (z-1)\Gamma(z-1), \quad \Gamma(1/2) = \sqrt{\pi}.$$

В результате получаем *нормальную плотность со средним значением  $\mu$  и дисперсией  $\sigma^2$* :

Нормальная плотность
Функция $n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x - \mu)^2/2\sigma^2\}$
Значения $-\infty < x < \infty$
Параметры $-\infty < \mu < \infty, \sigma^2 > 0$
Среднее значение $\mu$
Дисперсия $\sigma^2$

Соответствующая нормальная случайная величина обозначается  $N(\mu, \sigma)$

## 14.2 Стандартная нормальная плотность

Рассмотрим случайную величину

$$Z = \frac{N(\mu, \sigma) - \mu}{\sigma}.$$

Очевидно,  $EZ = (1/\sigma)[EX - \mu] = 0$  и  $\text{Var}Z = (1/\sigma^2)\text{Var}N = \sigma^2/\sigma^2 = 1$ . Чтобы найти её плотность, вычислим сначала её кумулятивную функцию распределения:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P\left(\frac{N(\mu, \sigma) - \mu}{\sigma} \leq z\right) = P(N(\mu, \sigma) \leq \mu + \sigma z) = \\ &= \int_{-\infty}^{\mu + \sigma z} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \end{aligned}$$

Таким образом, случайная величина  $Z$  имеет плотность

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

Отметим, что это ничто иное как частный случай нормальной плотности с  $\mu = 0$  и  $\sigma^2 = 1$ , называемой *стандартной нормальной плотностью*:

$$n(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

Нормальная случайная величина выражается через стандартную нормальную величину с помощью соотношения

$$N(\mu, \sigma) = \mu + \sigma Z.$$

### 14.3 Производящая функция моментов

Вспомним определение производящей функции моментов (см. 6.5):

$$M_X(t) = Ee^{tX} = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

Для стандартной нормальной величины  $Z$

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2+t^2/2} dx = e^{t^2/2}.$$

Для нормальной величины  $N$  со средним значением  $\mu$  и дисперсией  $\sigma^2$

$$M_N(t) = M_{\mu+\sigma Z}(t) = e^{\mu t + \sigma^2 t^2/2}.$$

Используем далее важное свойство ПФМ:

$$M_{X_1+\dots+X_n}(t) = [M_X(t)]^n$$

для любых независимых копий  $X$ . Это легко может быть доказано с использованием Теоремы 2 из 8.4.

### 14.4 Центральная предельная теорема

**Теорема.** Если  $X_1, X_2, \dots, X_n$  - независимые копии случайной величины  $X$  с конечной дисперсией  $\sigma^2$  и средним значением  $\mu$  и  $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$  тогда

$$F_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(x) \rightarrow F_Z(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

при  $n \rightarrow \infty$ . Этот факт иногда записывают как

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, \quad n \rightarrow \infty.$$

**Доказательство.** Мы должны доказать, что  $M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) \rightarrow e^{t^2/2}$ . Представляя эту случайную величину в виде суммы

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n \overset{\circ}{X}_j,$$

где  $\overset{\circ}{X}_j = X_j - \mu$  - центрированные копии величин  $X_j$ , и используя соответствующие свойства ПФМ, получим:

$$M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) = M_{\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n \overset{\circ}{X}_j}(t) = M_{\sum_{j=1}^n \overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right) = \left[M_{\overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right)\right]^n.$$

При  $n \rightarrow \infty$

$$M_{\overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right) \sim 1 + \frac{\mu_2' t^2}{2\sigma^2 n}.$$

Так как  $\mu = 0$ ,  $\mu_2' = \sigma^2$ . Используя асимптотические соотношения  $1 + \varepsilon \sim e^\varepsilon$ ,  $\varepsilon \rightarrow 0$ , получаем желаемый результат:

$$M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) \rightarrow e^{t^2/2}, \quad n \rightarrow \infty.$$

## 14.5 Двумерное нормальное распределение.

Пусть  $Z_1$  и  $Z_2$  - две независимые стандартные нормальные случайные величины. Их совместная плотность имеет вид:

$$f_{Z_1, Z_2}^{(0)}(z_1, z_2) = f_{Z_1}(z_1)f_{Z_2}(z_2) = \frac{1}{\sqrt{2\pi}}e^{-z_1^2/2} \frac{1}{\sqrt{2\pi}}e^{-z_2^2/2} = A_0 e^{-Q_0(z_1, z_2)/2}$$

где  $Q_0(z_1, z_2) = z_1^2 + z_2^2$ . Если мы пересечём трёхмерный график плотности плоскостью, параллельной  $xOy$ ,

$$f(x, y) = \text{const}$$

получим фигуру, описываемую уравнением

$$Q_0(z_1, z_2) = z_1^2 + z_2^2 = \text{const}.$$

Это окружность, или, для разных констант, - семейство концентрических окружностей. Но если *центрированные* нормальные случайные величины имеют разные дисперсии, результирующей фигурой будет эллипс. Эти случайные величины всё ещё независимы. Но линейные преобразования, соответствующие вращению в плоскости  $xOy$ , преобразуют их в *пару коррелированных нормальных величин* с новой совместной плотностью

$$f_{Z_1, Z_2}(z_1, z_2) = A e^{-Q(z_1, z_2)/2}$$

где

$$Q(z_1, z_2) = \frac{1}{1 - \rho^2} [z_1^2 - 2\rho z_1 z_2 + z_2^2]$$

и

$$A = \frac{1}{2\pi\sqrt{1 - \rho^2}}.$$

Это совместная плотность двух *коррелированных* (с коэффициентом корреляции  $\rho$ ) стандартных нормальных величин ( $\sigma_1 = \sigma_2 = 1$ ). В общем случае имеем:

$$Q(z_1, z_2) = \frac{1}{1 - \rho^2} \left[ \left( \frac{z_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{z_1 - \mu_1}{\sigma_1} \right) \left( \frac{z_2 - \mu_2}{\sigma_2} \right) + \left( \frac{z_2 - \mu_2}{\sigma_2} \right)^2 \right]$$

и

$$A = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}}.$$

Можно показать, что условные и маргинальные плотности, полученные из этой совместной плотности, снова являются нормальными плотностями.

## 15 Лекция 14. Решение типичных задач

### 15.1 Задача 2.

Вероятность непрерывно и равномерно распределена по треугольнику с вершинами  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 2)$  (первая координата -  $x$ , вторая -  $y$ ).

**Найти:** (a)  $f(x, y)$ ; (b)  $f(x)$ ; (c)  $f(y|x)$ ; (d) кумулятивную функцию распределения  $F(x)$ ; (e)  $F(y|x)$ ; (f)  $P(X + Y < z)$ ; (g)  $f_{X+Y}(z)$ ; (h)  $EX$  и  $EY$ ; (i)  $\sigma_X^2$  и  $\sigma_Y^2$ ; (j)  $\sigma_{XY}$  и  $\rho_{XY}$ . (k) Являются ли они независимыми случайными величинами или нет? Объясните, почему.

(a) Начертим рисунок. Найдём область  $A$  треугольника  $\triangle A = \frac{1}{2} \cdot 1 \cdot 1 = \frac{1}{2}$  и затем совместную плотность вероятности:  $f(x, y) = \frac{1}{A} = 2$ . Записывая ответ, не забудьте указать область изменения аргументов:

$$f(x, y) = \begin{cases} 2, & (x, y) \in \triangle, \\ 0, & \text{otherwise.} \end{cases}$$

(b) Как можно видеть из рисунка, пределы интеграла для маргинальной плотности

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{y_1(x)}^{y_2(x)} 2 dy$$

определяются уравнениями

$$y_1(x) = 2x, \quad y_2(x) = 1 + x.$$

В результате мы получим

$$f(x) = 2[(1 + x) - 2x] = 2(1 - x), \quad 0 \leq x \leq 1.$$

(c) Условная плотность

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{1}{1 - x}, \quad 2x \leq x \leq 1 + x.$$

Проверим нормировку:

$$\int_{2x}^{1+x} \frac{dy}{1 - x} = \frac{1}{1 - x} \int_{2x}^{1+x} dy = 1.$$

(d) Кумулятивная функция  $F(x)$  описывается формулой

$$F(x) = \int_{-\infty}^x f(x') dx' = \int_0^x 2(1 - x') dx' = x(2 - x), \quad 0 \leq x \leq 1.$$

Проверим нормировку:  $F(x_{\min}) = F(0) = 0$ ,  $F(x_{\max}) = F(1) = 1$ .

(e) Кумулятивная условная функция распределения

$$F(y|x) = \int_{-\infty}^y f(y'|x) dy' = \int_{2x}^y \frac{dy'}{1 - x} = \frac{1}{1 - x} \int_{2x}^y dy' = \frac{y - 2x}{1 - x}, \quad 2x \leq y \leq 1 + x.$$

Отметим, что  $F(y_{\min}|x) = F(2x|x) = 0$  и  $F(y_{\max}|x) = F(1 + x|x) = 1$ .

(f) Искомая вероятность даётся формулой  $P(X + Y \leq z) = A_z/A$ , где  $A_z$  - площадь треугольной части под прямой  $y + x = z$  и  $A$  - полная площадь треугольника.

Если  $z < 1$ , тогда (см. рисунок)  $A_z = \frac{1}{2}zh$ , где  $h$  - высота треугольника, даваемая  $x$ -координатой точки пересечения двух прямых линий:  $x + y = z$  и  $y = 2x$ . Решая эту систему, мы получим:  $h = x = z/3$ . Таким образом,

$$P(X + Y < z) = z^2/3, \quad 0 \leq z \leq 1.$$

При  $z > 1$  более удобно рассматривать верхнюю часть базисного треугольника, которая тоже имеет треугольную форму. Её площадь пропорциональна  $(3 - z)^2$ , поэтому вероятность  $P(X + Y > z) = C(3 - z)^2, z > 1$ . Постоянная  $C$  может быть найдена из условия нормировки:

$$P(X + Y < 1) + P(X + Y > 1) = \frac{1}{3} + C(3 - 1)^2 = 1.$$

Тогда,  $C = 1/6$  и

$$P(X + Y < z) = \begin{cases} (1/3)z^2, & 0 \leq z \leq 1, \\ 1 - (1/6)(3 - z)^2, & 1 \leq z \leq 3. \end{cases}$$

(g) Вероятность, найденная выше - это кумулятивная функция распределения суммы

$$P(X + Y < z) \equiv F_{X+Y}(z),$$

откуда

$$f_{X+Y}(z) = \frac{dF_{X+Y}(z)}{dz} = \begin{cases} (2/3)z, & 0 \leq z \leq 1, \\ (1/3)(3 - z), & 1 \leq z \leq 3. \end{cases}$$

(h) Используя результат, полученный в (b), получим

$$\mu_X = EX = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 2(1-x)x dx = 2 \left[ \frac{1}{2} - \frac{1}{3} \right] = \frac{1}{3}.$$

Чтобы найти  $\mu_Y$  нет необходимости вычислять другую маргинальную плотность  $f(y)$ . Можно использовать правило  $E(X + Y) = EX + EY$ , левую часть которого можно вычислить, используя  $f_{X+Y}(z)$ :

$$E(X + Y) = \int_0^3 zf_{X+Y}(z)dz = (2/3) \int_0^1 z^2 dz + (1/3) \int_1^3 (3 - z)z dz = \frac{4}{3}.$$

Тогда мы имеем:

$$\mu_Y = \frac{4}{3} - \frac{1}{2} = 1.$$

Можно проверить этот результат, используя вышеупомянутый способ:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx = \begin{cases} y, & 0 < y < 1, \\ 2 - y, & 1 \leq y \leq 2, \end{cases}$$

$$\mu_Y = \int_{-\infty}^{\infty} yf(y)dy = \int_0^1 y^2 dy + \int_1^2 y(2 - y)dy = 1.$$

(i) Вторые моменты

$$EX^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 2(1-x)x^2 dx = 2 \left[ \frac{1}{3} - \frac{1}{4} \right] = \frac{1}{6},$$

и

$$EY^2 = \int_{-\infty}^{\infty} y^2 f(y)dy = \int_0^1 y^3 dy + \int_1^2 y^2(2 - y)dy = \frac{7}{6}.$$

Дисперсии

$$\sigma_X^2 = EX^2 - \mu_X^2 = \frac{1}{6} - \left( \frac{1}{3} \right)^2 = \frac{1}{18}$$

and

$$\sigma_Y^2 = EY^2 - \mu_Y^2 = \frac{7}{6} - 1^2 = \frac{1}{6}.$$



(j) Начиная с вычисления среднего значения произведения  $XY$

$$\begin{aligned} EXY &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy xy f(x, y) = \int_0^1 dx \int_{2x}^{1+x} dy xy \cdot 2 = 2 \int_0^1 dx x \int_{2x}^{1+x} y dy = \\ &= \int_0^1 x[(1+x)^2 - (2x)^2] dx = \int_0^1 x[2x - 3x^2 + 1] dx = \frac{5}{12}, \end{aligned}$$

получим

$$\sigma_{XY} = EXY - \mu_X \mu_Y = \frac{1}{12}$$

и

$$\rho_{XY} = \frac{\frac{1}{12}}{\sqrt{\frac{1}{18 \cdot 6}}} = \frac{\sqrt{3}}{2}.$$

(k) Каждого из результатов, полученных выше, достаточно, чтобы сделать заключение, что  $X$  и  $Y$  не являются независимыми:

$$f(x, y) \neq f(x)f(y);$$

$$f(y|x) \text{ depends on } x;$$

$$EXY \neq EX \cdot EY;$$

$$\sigma_{XY} \neq 0;$$

$$\rho_{XY} \neq 0.$$

**Замечание.** Три последних критерия не могут быть использованы в обратную сторону:  $EXY = EX \cdot EY$  (или  $\sigma_{XY} = 0$ ) не означает, что  $X$  и  $Y$  независимы.



STUDENT NAME..... March 8 2006 1230 1400 CLPP

**MIDTERM EXAM 18100 STAT 312**  
**STATISTICS FOR ENG AND SCIENCE**  
ENR: 41 M W 1230 1345 CLPP 108 INSTR: UCHAYKIN, V

**Задача 1.** Пусть в шляпе содержится 10 свёрнутых листочков бумаги, на одном из которых написано число 1, на двух написано число 2, на трёх - число 3 и на оставшихся четырёх - число 4. Пусть  $X$  - число случайно выбранных листочков.

**Найти:**

- (a) распределение вероятностей  $f(x)$  ;
- (b) кумулятивную функцию распределения  $F(x)$  ;
- (c) среднее значение  $\mu$  и дисперсию  $\sigma^2$ ;
- (d) вероятности  $P(X = 0 \cup X = 1 \cup X = 5)$  и  $P(X = 2 \cup X = 3 \cup X = 4)$ .

**Задача 2.** Дана кумулятивная функция распределения

$$F(x) = \begin{cases} 0, & x < 0, \\ x^2, & 0 \leq x < 1, \\ 1, & x \geq 1 \end{cases}.$$

**Найти**

- (a) плотность вероятности  $f(x)$ ;
- (b) среднее значение  $\mu$ ;
- (c) дисперсию и стандартное отклонение;
- (d) вероятности  $P(X < \mu)$ ,  $P(X \leq \mu)$ ,  $P(X > \mu)$ .

**Задача 3.** Пусть время поступления первого телефонного вызова  $T_1$  и интервалы времени между поступлениями следующих вызовов  $T_2, T_3, \dots$  распределены согласно экспоненциальной плотности с тем же средним значением  $\mu = 4$ . Предположим, что эти времена взаимно независимы.

**Найти**

- (a) плотность вероятности  $f(x)$  и кумулятивную функцию распределения  $F(x)$  от  $T_1$ ;
- (b) вероятность, что не будет вызовов к моменту  $x = \ln(5)$ ;
- (c) среднее значение и стандартное отклонение времени 5-го вызова (будьте внимательны: это не промежуток времени  $T_5$ !);
- (d) плотность вероятности времени 5-го вызова.

**Задача 4.** Вероятность равномерно распределена по треугольнику с вершинами в точках  $(0, 0)$ ,  $(0, 1)$  и  $(1, 0)$ . Рассматривая случайные точки со случайными координатами  $X$  и  $Y$

**найти:**

- (a) маргинальные плотности величин  $X$  и  $Y$ ;
- (b) условные плотности величин  $X$  и  $Y$ ;
- (c) кумулятивную функцию распределения и плотность вероятности для суммы  $Z = X + Y$ .
- (d) Являются ли  $X$  и  $Y$  независимыми случайными величинами или нет? Объясните, почему.

**Задача 5.** Совместная плотность вероятности для двух случайных величин  $X$  и  $Y$  имеет вид:

$$f(x, y) = \begin{cases} ye^{-(x+y)}, & x \geq 0, y \geq 0, \\ 0, & \text{by fxt.} \end{cases}$$

**Найти:**

- (a) маргинальные плотности для  $X$  и  $Y$ ;
- (b) средние значения  $\mu_X$  и  $\mu_Y$ ;
- (c) дисперсии  $\sigma_X^2$  и  $\sigma_Y^2$ ;
- (d) являются ли  $X$  и  $Y$  независимыми случайными величинами или нет? Объясните, почему.

## Список определений, теорем и формул Allowed to the Use in Midterm Exam

**Теорема.** Если в эксперименте появляется  $n = n(S)$  различных одинаково вероятных исходов, и если ровно  $n(A)$  из этих исходов соответствует событию  $A$ , тогда вероятность события  $A$  равна

$$P(A) = \frac{n(A)}{n(S)}.$$

**Кумулятивная функция и плотность вероятности.** Для непрерывной случайной величины **кумулятивная функция** выражается через её плотность вероятности через интеграл:

$$F(x) = \int_{-\infty}^x f(x') dx'.$$

Плотность вероятности выражается через кумулятивную функцию через производную:

$$f(x) = \frac{dF(x)}{dx}.$$

**Теорема.** Дисперсия случайной величины может быть представлена в следующем виде, более удобном для приложений:

$$\text{Var}X = EX^2 - \mu^2 = \begin{cases} \sum x^2 f(x) - \mu^2, & \text{если } X \text{ дискретна;} \\ \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2, & \text{если непрерывна.} \end{cases}$$

**Определение.** Распределения  $f_X(x) \equiv \sum_y f(x, y)$  и  $f_Y(y) \equiv \sum_x f(x, y)$  в дискретном случае и плотности  $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$  и  $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$  в непрерывном случае называются **маргинальными распределениями** (**маргинальными плотностями**) отвечающими совместному распределению (плотности)  $f(x, y)$ .

**Определение.** **Условное распределение** случайной величины  $X$ , при условии что  $Y = y$ , и обозначаемое  $f(x|y)$  определяется формулой

$$f(x|y) = \frac{f(x, y)}{f(y)},$$

при  $f(y) > 0$ .

### Распределение Пуассона

Распределение Пуассона
Функция $f(x) = \frac{\mu^x}{x!} e^{-\mu}$
Значения $0, 1, 2, \dots$
Параметр $\mu > 0$
Среднее значение $\mu$
Дисперсия $\mu$

### Экспоненциальная плотность

Exponential density
Функция $f(x) = \frac{1}{\mu} e^{-x/\mu}$
Значения $x > 0$
Параметр $\mu > 0$
Среднее значение $\mu$
Дисперсия $\mu^2$

**Гамма-плотность** Гамма-плотность описывает сумму  $n$  независимых экспоненциально распределённых случайных величин.

Гамма-плотность
Функция $f(x) = \frac{1}{\Gamma(n)\mu} \left(\frac{x}{\mu}\right)^{n-1} e^{-x/\mu}$
Значения $x > 0$
Параметры $n > 0; \mu > 0$
Среднее значение $n\mu$
Дисперсия $n\mu^2$

### Интеграл

$$\int_0^{\infty} x^n e^{-x} dx = n!$$

## MIDTERM EXAM 18100 STAT 312 - SOLUTIONS

**Задача 1.** Пусть в шляпе содержится 10 свёрнутых листочков бумаги, на одном из которых написано число 1, на двух написано число 2, на трёх - число 3 и на оставшихся четырёх - число 4. Пусть  $X$  - число случайно выбранных листочков.

**Найти:**

(а) распределение вероятностей  $f(x)$  ;

**Решение:**

$$f(x) = P(X = x) = \frac{\text{число листочков с } x}{\text{полное число листочков}} =$$

$$= \begin{array}{c|c|c|c|c} x & 1 & 2 & 3 & 4 \\ \hline f(x) & 0.1 & 0.2 & 0.3 & 0.4 \end{array} = \begin{cases} 0.1x, & x = 1, 2, 3, 4, \\ 0, & \text{иначе.} \end{cases}$$

(b) кумулятивную функцию распределения  $F(x)$  ;

**Решение:**

$$F(x) \equiv P(X \leq x) = \sum_{x_i \leq x} f(x_i) = \begin{cases} 0.0, & x < 1, \\ 0.1, & 1 \leq x < 2, \\ 0.3, & 2 \leq x < 3, \\ 0.6, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

(с) среднее значение  $\mu$  и дисперсию  $\sigma^2$ ;

**Решение:**

$$\mu_X = \sum x f(x) = 1 \cdot 0.1 + 2 \cdot 0.2 + 3 \cdot 0.3 + 4 \cdot 0.4 = 3,$$

$$\sigma_X^2 = \sum x^2 f(x) - \mu_X^2 = 1^2 \cdot 0.1 + 2^2 \cdot 0.2 + 3^2 \cdot 0.3 + 4^2 \cdot 0.4 - 3^2 = 1.$$

(d) вероятности  $P(X = 0 \cup X = 1 \cup X = 5)$  и  $P(X = 2 \cup X = 3 \cup X = 4)$ .

**Решение:**

$$P(X = 0 \cup X = 1 \cup X = 5) = f(0) + f(1) + f(5) = 0 + 0.1 + 0 = 0.1;$$

$$P(X = 2 \cup X = 3 \cup X = 4) = f(2) + f(3) + f(4) = 0.9.$$

**Задача 2.** Дана кумулятивная функция распределения

$$F(x) = \begin{cases} 0, & x < 0, \\ x^2, & 0 \leq x < 1, \\ 1, & x \geq 1 \end{cases}$$

**найти**

(а) плотность вероятности  $f(x)$ ;

**Решение:**  $f(x) = \frac{dF(x)}{dx} = \begin{cases} 0, & x < 0, \\ 2x, & 0 \leq x < 1, \\ 0, & x \geq 1. \end{cases}$

(b) среднее значение  $\mu$ ;

**Решение:**  $\mu_X = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 2x dx = 2/3.$

(с) дисперсию и стандартное отклонение;

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_X^2 = \int_0^1 x^2 \cdot 2x dx - (2/3)^2 = 1/2 - 4/9 = 1/18,$$

$$\sigma_X = \sqrt{1/18} \approx 0.236.$$

(д) вероятности  $P(X < \mu)$ ,  $P(X \leq \mu)$ ,  $P(X > \mu)$ .

$$P(X < \mu) = P(X \leq \mu) = \int_{-\infty}^{\mu} f(x) dx = \int_0^{2/3} 2x dx = 2 \frac{(2/3)^2}{2} - 0 = \frac{4}{9},$$

$$P(X > \mu) = \int_{\mu}^{\infty} f(x) dx = \int_{2/3}^1 2x dx = 2 \frac{x^2}{2} \Big|_{2/3}^1 = 1 - \frac{4}{9} = \frac{5}{9}.$$

**Задача 3.** Пусть время поступления первого телефонного вызова  $T_1$  и интервалы времени между поступлениями следующих вызовов  $T_2, T_3, \dots$  распределены согласно экспоненциальной плотности с тем же средним значением  $\mu = 4$ . Предположим, что эти времена взаимно независимы.

**найти**

(а) плотность вероятности  $f(x)$  и кумулятивную функцию распределения  $F(x)$  от  $T_1$ ;

**Решение:**

$$f_T(x) = \begin{cases} 0, & x < 0, \\ (1/\mu)e^{-x/\mu} = 0.25e^{-0.25x}, & x \geq 0, \end{cases}$$

$$F_T(x) = \int_{-\infty}^x f(x') dx' = \begin{cases} 0, & x < 0, \\ \int_0^x (1/\mu)e^{-x'/\mu} dx' = 1 - e^{-0.25x}, & x \geq 0 \end{cases}.$$

(b) вероятность, что не поступит ни одного вызова к моменту  $x = \ln(5)$ ;

Это означает, что первый вызов поступит позже, чем  $x = \ln(5)$ :

$$\begin{aligned} P(T_1 > \ln(5)) &= 1 - P(T_1 \leq \ln(5)) = 1 - F_T(\ln(5)) = \\ &= e^{-0.25 \ln(5)} = \left(e^{\ln(5)}\right)^{-0.25} = 5^{-0.25} = \frac{1}{\sqrt[4]{5}} \approx 0.668. \end{aligned}$$

(с) среднее значение и стандартное отклонение времени 5-го вызова (будьте внимательны: это не промежуток времени  $T_5$ !);

Время 5-го вызова - это сумма пяти независимых интервалов времени между вызовами, таким образом

$$\mu_{\sum_{j=1}^5 T_j} = 5\mu = 5 \cdot 4 = 20, \quad \sigma_{\sum_{j=1}^5 T_j}^2 = 5\sigma_T^2 = 5\mu^2 = 5 \cdot 4^2 = 80, \quad \sigma_{\sum_{j=1}^5 T_j} = 4\sqrt{5}.$$

Здесь мы использовали свойство экспоненциального распределения: его дисперсия равна квадрату его среднего значения.

(d) плотность вероятности времени 5-го вызова.

Эта сумма подчиняется гамма-распределению с параметрами  $\mu = 4$  и  $n = 5$

$$f_{\sum_{j=1}^5 T_j} = \frac{1}{\Gamma(n)\mu} \left(\frac{x}{\mu}\right)^{n-1} e^{-x/\mu} = \frac{1}{24 \cdot 4} \left(\frac{x}{4}\right)^{5-1} e^{-x/4} = \frac{x^4 e^{-0.25x}}{24576}.$$

**Задача 4.** Вероятность равномерно распределена по треугольнику с вершинами в точках  $(0, 0)$ ,  $(0, 1)$  и  $(1, 0)$ . Рассматривая случайные точки со случайными координатами  $X$  и  $Y$

**найти:**

(а) маргинальные плотности величин  $X$  и  $Y$ ;

**Решение:** Площадь этого треугольника равна  $1/2$ , тогда плотность распределения

$$f(x, y) = \begin{cases} \frac{1}{\text{Площадь}} = 2, & (x, y) \in \Delta, \\ 0, & \text{иначе.} \end{cases}$$

Нарисовав этот треугольник, можно заметить, что при фиксированном  $x$  переменная  $y$  изменяется от 0 до  $1 - x$  и, следовательно,

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^{1-x} 2 dy = 2(1 - x), \quad 0 < x < 1.$$

Аналогично,

$$f(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^{1-y} 2 dx = 2(1 - y), \quad 0 < y < 1.$$

(b) условные плотности величин  $X$  и  $Y$ ;

**Решение:**

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{2}{2(1 - y)} = \frac{1}{1 - y}, \quad 0 < x < 1 - y, \quad 0 < y < 1;$$

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{2}{2(1 - x)} = \frac{1}{1 - x}, \quad 0 < y < 1 - x, \quad 0 < x < 1.$$

(с) кумулятивную функцию распределения и плотность вероятности для суммы  $Z = X + Y$ .

**Решение:** Кумулятивная функция определяется через вероятность

$$F_{X+Y}(z) = P(X + Y < z),$$

а эта вероятность определяется как отношение площадей

$$P(X + Y < z) = \frac{\text{Площадь } \Delta(x + y < z)}{\text{Площадь } \Delta(x + y < 1)} = z^2.$$

Следовательно,

$$F_{X+Y}(z) = \begin{cases} 0, & z < 0, \\ z^2, & 0 \leq z < 1, \\ 1, & z \geq 1 \end{cases}$$

и

$$f_{X+Y}(z) = \begin{cases} 0, & z < 0, \\ 2z, & 0 \leq z < 1, \\ 0, & z \geq 1. \end{cases}$$

(d) Являются ли  $X$  и  $Y$  независимыми случайными величинами или нет? Объясните, почему.



Они не являются независимыми, поскольку условные вероятности зависят от значений-условий.

**Задача 5.** Совместная плотность вероятности для двух случайных величин  $X$  и  $Y$  имеет вид:

$$f(x, y) = \begin{cases} ye^{-(x+y)}, & x \geq 0, y \geq 0, \\ 0, & \text{иначе.} \end{cases}$$

**Найти:**

(а) маргинальные плотности для  $X$  и  $Y$ ;

**Решение:**

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^{\infty} ye^{-(x+y)} dy = e^{-x} \int_0^{\infty} ye^{-y} dy = e^{-x}, \quad x \geq 0,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^{\infty} ye^{-(x+y)} dx = ye^{-y} \int_0^{\infty} e^{-x} dx = ye^{-y}, \quad y \geq 0.$$

(b) средние значения  $\mu_X$  и  $\mu_Y$ ;

**Решение:**  $\mu_X = \int_0^{\infty} xe^{-x} dx = 1, \quad \mu_Y = \int_0^{\infty} ye^{-y} dy = 1.$

(c) дисперсии  $\sigma_X^2$  и  $\sigma_Y^2$ ;

**Решение:**  $\sigma_X^2 = \int_0^{\infty} x^2 e^{-x} dx - 1^2 = 2, \quad \sigma_Y^2 = \int_0^{\infty} y^2 e^{-y} dy - 1^2 = 2.$

(d) являются ли  $X$  и  $Y$  независимыми случайными величинами или нет? Объясните, почему.

**Solution:** Они являются независимыми, поскольку их совместная плотность равна произведению их маргинальных плотностей:

$$f(x, y) = f(x)f(y).$$

**Замечание:** В (а)-(с) использован интеграл  $\int_0^{\infty} x^n e^{-x} dx = n!.$

# МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

## 16 Лекция 15. Random Sampling

### 16.1 Генеральная совокупность и выборка

### 16.2 Статистический метод

Основные термины:

**Генеральная совокупность** – совокупность  $N$  объектов (множество элементов), обладающих определенными признаками.

**Выборка** – подмножество генеральной совокупности, выбранное таким образом, что все ее элементы имели равные шансы быть представленными в выборке.

**Объем выборки** – число объектов  $n$  в выборке.

**Статистический вывод** – заключение о свойствах генеральной совокупности, сделанное на основе изучения выборки.

**Пример 1.** Завод изготовил партию из 1000 деталей (генеральная совокупность). Эксперт выбрал 100 из них (выборка), чтобы сделать заключение о качестве данной партии (статистический вывод). Он обнаружил 10 дефектных деталей, хотя в течение долгого времени считалось, что уровень брака данной производственной линии составляет 5%. Какой вывод должен сделать эксперт? Проблема в том, что другая выборка может дать другой результат, и это вносит некоторую неопределенность, в условиях которой и действует эксперт. Он должен понять, является ли данный результат свойством конкретной выборки, или отражает свойство всей генеральной совокупности. В последнем случае эксперт должен заявить, что качество выпускаемой данной линией продукции ухудшилось: процент брака вырос вдвое.

**Пример 2.** Предположим, мы знаем количество рыбы  $N$  в озере Эри и нужно определить её полную массу  $M$ . Чтобы сделать это, мы должны выбрать *случайным образом* некоторое число рыб  $n$ , измерить их массы  $m_1, \dots, m_n$ , найти среднюю массу (*выборочное среднее*)

$$\bar{m} = \frac{1}{n} \sum_{j=1}^n m_j,$$

и вычислить  $\bar{m} \cdot N$ . Если  $n = N$  (измерены все рыбы), то мы получим точный результат

$$\frac{1}{N} \sum_{j=1}^N m_j \cdot N = \sum_{j=1}^N m_j = M. \quad (1)$$

Заметим, что  $\frac{1}{N} \sum_{j=1}^N m_j$  – это *среднее генеральной совокупности*, которое будем обозначать  $\mu$ .

Однако в действительности  $n \ll N$  и равенство становится приближительным:

$$\frac{1}{n} \sum_{j=1}^n m_j \cdot N \approx M. \quad (2)$$

Более того, повторяя измерения, мы получим то же  $M$  в первом случае (1) и некоторое другое число во втором случае (2). Левая часть выражения (2) – это *случайная величина*.

В рассматриваемом случае, все рыбы в озере образуют *генеральную совокупность*;

$N$  – *объём генеральной совокупности*;

множество  $\{t_1, \dots, t_n\}$  – это *случайная выборка*;

$n$  – *объём выборки*;

заключение "полная масса приблизительно равна  $\bar{m} \cdot N$ " – это *статистический вывод*.

## 16.3 Вероятностная модель случайной выборки

Чтобы обосновать правомерность процедур такого рода, используется *вероятностная модель*, в рамках которой генеральная совокупность характеризуется распределением вероятностей, а случайная выборка рассматривается как набор *независимых одинаково распределённых случайных величин*  $X_1, \dots, X_n$ , имеющих это распределение.

Чтобы проиллюстрировать продуктивность вероятностной модели, мы *докажем*, что уравнение (2) справедливо при больших  $n$ . Строго говоря, для этого ничего не потребуется делать, поскольку результат следует непосредственно из Центральной Предельной Теоремы (смотри 13.4). Надо просто переписать утверждение теоремы

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, \quad n \rightarrow \infty$$

в виде

$$\bar{X} \approx \mu + \frac{\sigma}{\sqrt{n}}Z, \quad n \rightarrow \infty \quad (1)$$

добавляя, что выборочное среднее  $\bar{X}$  совпадает со средним генеральной совокупности  $\mu \equiv EX$  с точностью до *статистической погрешности*

$$\varepsilon \approx \frac{\sigma}{\sqrt{n}}Z, \quad n \gg 1.$$

Множитель  $\frac{\sigma}{\sqrt{n}}$  убывает с ростом  $n$  как  $n^{-1/2}$  и для больших  $n$  мы имеем

$$\bar{X} \equiv \frac{1}{n} \sum_{j=1}^n X_j \approx \mu.$$

**Замечание:** Центральная Предельная Теорема не уточняет распределение величины  $X$ , она справедлива для любого распределения с конечной дисперсией  $\sigma^2$ . Но если случайные величины  $X_j$  сами являются нормально распределёнными, тогда утверждение теоремы становится справедливым для любого целого числа  $n$ , то есть

$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}}Z, \quad n = 1, 2, 3, \dots \quad (2)$$

вместо асимптотического равенства (1).

## 16.4 Распределения генеральной совокупности и их характеристики

Вспомнить: распределение Бернулли, дискретные распределения, нормальное распределение, логнормальное распределение.

Разные формы:

колоколообразная (как нормальное распределение), треугольная, прямоугольная (всё это симметричные распределения);

асимметричное вправо (коэффициент асимметрии  $\gamma_1 \equiv \frac{1}{\sigma^3} E(X - \mu)^3 > 0$ ), асимметричное влево ( $\gamma_1 < 0$ ) (это асимметричные распределения),

для симметричных распределений, включая центрированное нормальное,  $\gamma_1 = 0$

эксцесс:  $\gamma_2 = \frac{1}{\sigma^3} E(X - \mu)^3$ , для нормального распределения  $\gamma_2 = 3$ .

*Мода*  $x_m$  – положение максимума  $f(x)$ . Существуют унимодальные, бимодальные и мультимодальные распределения.

*Медиана*  $x_{1/2}$  – точка, удовлетворяющая условию  $P(X \leq x_{1/2}) = 1/2$ .

## 17 Лекция 16. Выборочные распределения

### 17.1 Статистики и оценки

С вероятностной точки зрения, формулировка статистической задачи выглядит следующим образом. Есть некоторая генеральная совокупность, описываемая распределением  $f(x) = f(x; \mu, \sigma, \gamma_1, \gamma_2, \dots)$ , но ни функция распределения ни её параметры не известны. Мы производим случайную выборку  $X_1, \dots, X_n$  для того, чтобы сделать некоторое заключение о генеральной совокупности, то есть о некоторых её параметрах  $\mu, \sigma, \dots$  или даже о функции распределения  $f(x)$ . для этой цели мы должны выполнить некоторые математические операции над числами  $X_1, \dots, X_n$ .

**Определение 1.** Дан набор случайных величин  $X_1, \dots, X_n$ , *статистика* – это функция этих случайных величин, не использующая никаких неизвестных параметров.

**Определение 2.** Статистика, используемая для оценивания параметра  $\theta$  генеральной совокупности называется *оценкой* величины  $\theta$  и обозначается  $\hat{\theta}$ .

**Примеры.**  $\sum_{j=1}^n X_j$  и  $\sum_{j=1}^n X_j^2$  – статистики, статистика  $\bar{X} = (1/n) \sum_{j=1}^n X_j$  – это оценка для среднего генеральной совокупности  $\mu$ .

**Определение 3.** Оценка  $\hat{\theta}$  для параметра  $\theta$  называется *несмещённой*, если  $E\hat{\theta} = \theta$ .

### 17.2 Выборочное среднее

Статистика  $\bar{X} = (1/n) \sum_{j=1}^n X_j$  – оценка генерального среднего  $\mu$ . Она является несмещённой, поскольку

$$E\bar{X} = E\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n} \sum_{j=1}^n EX_j = \frac{1}{n} \sum_{j=1}^n \mu = \mu.$$

Найдём её дисперсию:

$$\text{Var}\bar{X} = \text{Var}\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n^2} \sum_{j=1}^n \text{Var}X_j = \frac{1}{n^2} \sum_{j=1}^n \sigma^2 = \frac{1}{n} \sigma^2,$$

где  $\sigma^2$  – дисперсия генеральной совокупности.

Формула (2) из 15.2 отвечает на вопрос о распределении величины  $\bar{X}$ : она распределена нормально со средним значением  $\mu$  и дисперсией  $\sigma/\sqrt{n}$ . Используя эту формулу, мы для любого заданного  $x$  можем найти

$$P(\bar{X} < x) = P\left(\mu + \frac{\sigma}{\sqrt{n}}Z < x\right) = P\left(Z < \frac{x - \mu}{\sigma/\sqrt{n}}\right) = P(Z < z_x),$$

где  $Z$  – стандартная нормальная случайная величина и  $z_x$  определяется данной  $x$  через соотношение  $z_x = \frac{x-\mu}{\sigma/\sqrt{n}}$ . Вы можете найти эту вероятность из таблицы **A.3** (стр. 670 учебника). Например,  $P(Z < 0.95) = 0.8289$ .

### 17.3 Выборочное среднее

Мы видели, что генеральное среднее  $\mu = EX$  имеет несмещённую оценку  $\hat{\mu} \equiv \bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ . Таким образом, можно ожидать что генеральная дисперсия

$$\sigma^2 = EX^2 - \mu^2$$

получается путём подобной замены  $E$  на  $\frac{1}{n} \Sigma$  и  $\mu$  by  $\hat{\mu} = \bar{X}$ :

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{j=1}^n X_j^2 - \hat{\mu}^2 = \frac{1}{n} \left( \sum_{j=1}^n X_j^2 - n\bar{X}^2 \right).$$

Однако, вычисления показывают, что

$$E\widehat{\sigma^2} = \frac{n-1}{n} \sigma^2$$

то есть  $\widehat{\sigma^2}$  – смещённая оценка для генеральной дисперсии  $\sigma^2$ . Переписанная в виде

$$E \frac{n}{n-1} \widehat{\sigma^2} = \sigma^2$$

эта формула показывает, что несмещённая оценка для  $\sigma^2$  имеет вид

$$s^2 = \frac{n}{n-1} \widehat{\sigma^2} = \frac{1}{n-1} \left( \sum_{j=1}^n X_j^2 - n\bar{X}^2 \right).$$

Заметим, что разница между  $\widehat{\sigma^2}$  и  $s^2$  становится существенной только при малых  $n$ .

Величина  $s^2$  – случайная величина. Как распределена её вероятность? Прежде всего, эта вероятность распределена по положительной полуоси, потому что  $s^2$  не может принимать отрицательных значений. Специальные вычисления показывают, что в случае нормальной генеральной совокупности масштабированная выборочная дисперсия  $(n-1)s^2/\sigma^2$  имеет гамма-распределение с параметрами  $\alpha = (n-1)/2$  и  $\beta = 2$ . Такая случайная величина имеет специальное обозначение:  $\chi^2(v)$  (и называется *хи-квадрат с  $v$  степенями свободы*). То есть, масштабированная выборочная дисперсия имеет хи-квадрат распределение с  $n-1$  степенями свободы:

$$(n-1)s^2/\sigma^2 \stackrel{d}{=} \chi^2(n-1).$$

Выборочная дисперсия выражается соотношением:

$$s^2 \stackrel{d}{=} \frac{\sigma^2}{n-1} \chi^2(n-1). \quad (1)$$

Пусть  $\alpha$  – некоторое положительное число между 0 и 1. Величина  $x_\alpha$ , подчиняющаяся уравнению  $P(X > x_\alpha) = \alpha$  называется *критическим значением*. Таблица критических значений  $\chi^2$  представлена на странице 674 (Таблица **A.5**).

Например, известны  $\sigma$  и  $n$ , требуется найти такое число  $x$ , что вероятность для оценки  $s^2$  превысить его равна  $\alpha = 0.1$ . Это число обычно обозначается символом  $s_\alpha^2$  с индексом  $\alpha$ , так что мы можем записать

$$P(s^2 > s_\alpha^2) = \alpha$$

или для заданного значения  $\alpha$

$$P(s^2 > s_{0.1}^2) = 0.1.$$

Используя формулу (1), получим:

$$P(s^2 > s_{0.1}^2) = P\left(\frac{\sigma^2}{n-1}\chi^2(n-1) > s_{0.1}^2\right) = P\left(\chi^2(n-1) > \frac{n-1}{\sigma^2}s_{0.1}^2\right) = 0.1.$$

Обозначим число  $\frac{n-1}{\sigma^2}s_{0.1}^2$  через  $\chi_{0.1}^2$  и перепишем предыдущее уравнение в виде двух:

$$P(\chi^2(n-1) > \chi_{0.1}^2) = 0.1 \quad (2)$$

и

$$\chi_{0.1}^2 = \frac{n-1}{\sigma^2}s_{0.1}^2. \quad (3)$$

Решение уравнения (2) представлено в таблице **A.5** (страница 647). Используя известный объём выборки  $n$ , например  $n = 12$ , определим число степеней свободы  $v = n - 1 = 11$ , и найдём в таблице число, размещённое в строке с  $v = 11$  и в столбце с  $\alpha = 0.1$ , откуда получаем  $\chi_{0.1}^2 = 17.275$ . Используя (3), получим ответ:

$$s_{0.1}^2 = \frac{\sigma^2}{n-1}\chi_{0.1}^2 = \frac{\sigma^2}{11}17.275.$$

## 17.4 Нормированное выборочное среднее

Вы знаете, что если  $X$  – нормальная случайная величина со средним значением  $\mu$  и стандартным отклонением  $\sigma$ , тогда нормированная случайная величина  $\frac{X-\mu}{\sigma}$  распределена согласно стандартному нормальному закону:

$$\frac{X - \mu}{\sigma} \stackrel{d}{=} Z.$$

Теперь мы рассмотрим выборочное среднее  $\bar{X}$ . В случае нормальной генеральной совокупности  $\bar{X}$  будет нормально распределена со средним значением  $\mu$  и стандартным отклонением  $\sigma/\sqrt{n}$ , таким образом,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{d}{=} Z.$$

Но если мы не знаем  $\sigma$  и используем вместо неё выборочное стандартное отклонение  $s = \sqrt{s^2}$ , как будет распределена случайная величина

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}?$$

Ответ таков: её распределение совпадает с так называемым  $t$ -распределением с  $v = n - 1$  числом степеней свободы. Обозначая соответствующую случайную величину через  $T(n-1)$ , мы можем записать:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \stackrel{d}{=} T(n-1).$$

Это двухстороннее симметричное относительно 0 распределение, оно становится близким к стандартному нормальному распределению при  $n > 30$ . Критические значения для этого распределения представлены в таблице **A.4**(стр. 672).

**Замечание.** Это распределение также называется *распределением Стьюдента*, поскольку его автор, W.S.Gossett, опубликовал свой результат под псевдонимом "Стьюдент".

## 17.5 Отношение двух выборочных дисперсий

Предположим, что случайные выборки объёмов  $n_1$  и  $n_2$  выбираются из двух независимых нормальных генеральных совокупностей с дисперсиями  $\sigma_1^2$  и  $\sigma_2^2$  соответственно. Из **16.3** мы знаем, что

$$s_1^2 \stackrel{d}{=} \frac{\sigma_1^2}{n_1 - 1} \chi^2(n_1 - 1)$$

и

$$s_2^2 \stackrel{d}{=} \frac{\sigma_2^2}{n_2 - 1} \chi^2(n_2 - 1).$$

Как будет распределено их отношение? Ответ на этот вопрос:

$$\frac{s_1^2}{s_2^2} \stackrel{d}{=} \frac{\sigma_1^2}{\sigma_2^2} F(n_1 - 1, n_2 - 1),$$

где  $F(v_1, v_2)$  – случайная величина с распределением Фишера (F-распределение) с числом степеней свободы  $v_1 (= n_1 - 1)$  и  $v_2 (= n_2 - 1)$ . Их соответствующие критические значения могут быть найдены в таблице **A.6**(стр. 676).

## 18 Лекция 17. Типичные задачи

### 18.1 Порядковые статистики

В некоторых случаях удобно располагать элементы выборки в порядке их возрастания (*порядковая статистика*). Например, если  $\{2.7, 2.2, 3.9, 1.9, 2.5\}$  – исходная случайная выборка, тогда  $\{1.9, 2.2, 2.5, 2.7, 3.9\}$  – упорядоченная выборка (вариационный ряд????). Статистики, получаемые из упорядоченных выборок называются *порядковыми статистиками*. Например, если число элементов выборки нечётно, как в приведённом случае, *выборочная медиана* определяется как средний элемент упорядоченной выборки (вариационного ряда). В нашем случае  $\{1.9, 2.2, \mathbf{2.5}, 2.7, 3.9\}$ , так что медиана – 2.5. Если число элементов выборки чётное, то выборочная медиана определяется как арифметическое среднее от пары элементов, находящихся в середине выборки. Например, для  $\{1.9, 2.2, \mathbf{2.5}, \mathbf{2.7}, 3.9, 3.9\}$  выборочная медиана равна  $(2.5 + 2.7)/2 = 2.6$ .

Другая порядковая статистика – это выборочная мода. Если ни одно значение в выборке не появляется чаще одного раза, тогда мы говорим, что выборка не имеет моды. Иначе, любое значение, которое появляется с максимальной частотой называется *выборочной модой*. Например, выборка  $\{1.9, 2.2, 2.5, 2.7, \mathbf{3.9}, \mathbf{3.9}\}$  имеет моду 3.9, выборка  $\{1.9, \mathbf{2.2}, \mathbf{2.2}, \mathbf{2.2}, 2.5, 2.7, 3.9, 3.9\}$  имеет моду 2.2, выборка  $\{1.9, \mathbf{2.2}, \mathbf{2.2}, \mathbf{2.2}, 2.5, 2.7, \mathbf{3.9}, \mathbf{3.9}, \mathbf{3.9}\}$  имеет две моды 2.2 и 3.9.

Сравним три характеристики выборки  $\{1.9, 2.2, 2.2, 2.2, 2.5, 2.7, 2.7, 2.7, 3.9\}$ : среднее

$$\bar{X} = \frac{1.1 + 2.2 + \dots + 2.7 + 3.9}{9} \approx 2.56.$$

медиану

$$\hat{X}_{1/2} = 2.5,$$

моды

$$m_1 = 2.2, \text{ и } m_2 = 2.7.$$

### 18.2 Сравнение вычисления параметров генеральной совокупности и выборки

Параметр	Генеральная совокупность $f(x)$	Выборка $\{X_1, \dots, X_n\}$
Среднее	$\mu = \int_{-\infty}^{\infty} x f(x) dx$	$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$
Дисперсия	$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$	$s^2 = \frac{1}{n-1} \left[ \sum_{j=1}^n X_j^2 - n \bar{X}^2 \right]$
Асимметрия	$\gamma_1 = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x - \mu)^3 f(x) dx$	$\hat{\gamma}_1 = \frac{1}{s^3} \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^3$
Медиана	$x_{1/2}$ – это решение уравнения $\int_{-\infty}^{x_{1/2}} f(x) dx = \frac{1}{2}$	$\hat{X}_{1/2}$ середина или середина пары
Мода	$m$ – решение уравнения $f'(x) = 0, f''(x) < 0$	$\hat{m}$ – максимальная частота



### 18.3 Выборочные распределения

Параметр генеральной совокупности	Выборочная оценка	Связь со стандартными случайными величинами
Среднее $\mu$	$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$	$\bar{X} \stackrel{d}{=} \mu + \frac{\sigma}{\sqrt{n}}Z$ Tab A.3(P.670)
Дисперсия $\sigma^2$	$s^2 = \frac{1}{n-1} [\sum_{j=1}^n X_j^2 - n\bar{X}^2]$	$s^2 \stackrel{d}{=} \frac{\sigma^2}{(n-1)}\chi^2(n-1)$ Tab A.5(P.674)
Выборочная $z$ -score ????	$\frac{\bar{X}-\mu}{s/\sqrt{n}}$	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \stackrel{d}{=} T(n-1)$ Tab A.4(P.672)
Отношение дисперсий $\sigma_1^2/\sigma_2^2$	$s_1^2/s_2^2$	$s_1^2/s_2^2 \stackrel{d}{=} (\sigma_1^2/\sigma_2^2)F(n_1-1, n_2-1)$ Tab A.6(676)

### 18.4 Примеры

**Пример 1.** Пусть  $X$  – нормальная случайная величина со средним  $\mu = 2$  и стандартными отклонением  $\sigma = 4$ . Найти вероятность  $P(-1 < X < 2)$ .

**Решение:** Во-первых,

$$P(-1 < X < 2) = P(X < 2) - P(X < -1/2).$$

Во-вторых, любая нормальная случайная величина с параметрами  $\mu$  и  $\sigma$  может быть представлена как

$$X = \mu + \sigma Z,$$

где  $Z$  – стандартная нормальная случайная величина с распределением, представленным в таблице Table A.3(Page 630). Используя это соотношение, получим

$$P(X < x) = P(\mu + \sigma Z < x) = P\left(Z < \frac{x - \mu}{\sigma}\right) = P(Z < z_x)$$

где  $z_x = (x - \mu)/\sigma$ . Подставляя сюда числа и используя таблицу, получим:

$$z_2 = (2 - 2)/4 = 0, \quad P(X < 2) = P(Z < 0) = 1/2;$$

$$z_{-1} = (-1 - 2)/4 = -0.75, \quad P(X < -1) = P(Z < -0.75) = 0.2266.$$

Таким образом

$$P(-1 < X < 2) = 0.5 - 0.2266 = 0.2734.$$

**Пример 2.** Пусть независимые случайные величины  $X_j$  распределены как в предыдущем примере и  $Y = X_1 + X_2 + X_3 + X_4$ . Найти  $P(-1 < Y < 2)$ .

**Решение.**

Сумма независимых нормальных величин – вновь нормальная случайная величина, и and

$$Y = \mu_Y + \sigma_Y Z.$$

Среднее суммы равно сумме средних, и дисперсия *независимых* случайных величин равна сумме дисперсий,  $\mu_Y = n\mu = 4 \cdot 2 = 8$ ,  $\sigma_Y^2 = n\sigma^2 = 4 \cdot 4^2 = 64$ . Тогда

$$Y \stackrel{d}{=} n\mu + \sqrt{n}\sigma Z = 8 + 8Z.$$

В результате получаем:

$$P(Y < 2) = P(4 + 8Z < 2) = P(Z < -0.25) = 0.4013,$$

$$P(Y < -1) = P(4 + 8Z < -1) = P(Z < -0.625) = 0.266,$$

и

$$P(-1 < Y < 2) = 0.4013 - 0.266 \approx 0.135.$$

**Пример 3.** При тех же предположениях, что и выше, найти  $P(-1 < \bar{X} < 2)$ , где  $\bar{X}$  – выборочное среднее.

**Решение:**

Вспомним, что среднее выборочного среднего  $\mu$  и его стандартное отклонение  $\sigma/\sqrt{n}$ , следовательно  $\bar{X} = 2 + 2Z$ . В результате, получаем

$$P(\bar{X} < 2) = P(2 + 2Z < 2) = P(Z < 0) = 1/2;$$

$$P(\bar{X} < -1) = P(2 + 2Z < -1) = P(Z < -1.5) = 0.0668;$$

$$P(-1 < \bar{X} < 2) = 0.5 - 0.0668 \approx 0.433.$$

## 18.5 Пример 4.

Пусть  $\{X_1, \dots, X_9\}$  – выборка из нормальной генеральной совокупности с  $\mu = 1$  и  $\sigma = 3$ . Найти критические значения для  $\alpha = 0.1$  (a)  $\frac{X-\mu}{\sigma}$ , (b)  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ , (c)  $\frac{\bar{X}-\mu}{s/\sqrt{n}}$  и (d)  $s^2$ .

**Решение.** Критическое значение  $x_\alpha$  для случайной величины  $X$  определяется уравнением  $P(X > x_\alpha) = \alpha$ . Решения этого уравнения табулированы.

(a) Случайная величина  $\frac{X-\mu}{\sigma}$  имеет стандартное нормальное распределение, его критические значения даны в нижней строке таблицы Table A.4 (при  $v = \infty$ ):  $z_{0.1} = 1.282$ .

(b) Эта величина имеет такое же распределение, как и ранее.

(c) Эта величина имеет t-распределение с 8 степенями свободы (см. Table in 17.3), её критическое значение 1.397 (Table A.4, Page 672.)

(d) Эта величина связана с  $\chi^2$  (см. Table in 17.3), тогда

$$P(s^2 > s_\alpha^2) = P\left(\chi^2 > \frac{n-1}{\sigma^2} s_\alpha^2\right) = P(\chi^2 > \chi_\alpha^2) = \alpha$$

где

$$\chi_\alpha^2 = \frac{n-1}{\sigma^2} s_\alpha^2.$$

Из таблицы Table A.5 (Page 675) найдём для  $\alpha = 0.1$  и  $v = n - 1 = 8$   $\chi_{0.1}^2(8) = 13.362$ , откуда

$$s_\alpha^2 = \frac{\sigma^2}{n-1} \chi_{0.1}^2(8) \approx \frac{9}{8} 13.4 \approx 15.$$

## STAT 312 Spring

### Домашняя работа 4

(must be returned on March 29, 12:30 p.m., CLPP 108)

**Задача 1.** Дано бета-распределение с параметрами  $\alpha = 3$ ,  $\beta = 2$ ,

**Найти:**

(a)  $\mu$ , (b)  $\sigma$ , (c)  $\gamma_1$  и (d) моду.

**Задача 2.** Пусть  $X$  – нормальная случайная величина с  $\mu = 1$  и  $\sigma = 2$ .

**Найти:**

(a)  $P(X < 0)$ , (b)  $P(X < -1)$ , (c)  $P(-1 < X < 1)$ , (d)  $P(X^2 > 1)$ .

**Задача 3.** Пусть  $X_1, X_2, X_3$  – три независимые случайные величины распределённые, как и ранее.

**Найти:**

(a)  $P(\bar{X} < 0)$ , (b)  $P(\bar{X} < -1)$ , (c)  $P(-1 < \bar{X} < 1)$ , (d)  $P(\bar{X}^2 > 1)$ .

**Задача 4.** Дана выборка  $\{13, 34, 40, 47, 17, 34, 40, 47, 21, 34\}$ ,

**найти:** (a) моду, (b) медиану, (c) среднее, (d) дисперсия.

**Задача 5.** Пусть  $\{X_1, \dots, X_{17}\}$  – случайная выборка из нормальной генеральной совокупности с  $\sigma^2 = 2$ .

**Найти критические значения для выборочной дисперсии:** (a)  $s_{0.05}^2$ , (b)  $s_{0.1}^2$ , (c)  $s_{0.2}^2$ , (d)  $s_{0.5}^2$ .

**Задача 6.** Пусть  $\{19, 24, 20, 21, 17, 23, 20, 25, 22\}$  – случайная выборка из нормальной генеральной совокупности с  $\mu = 21$ .

**Найти:** (a) выборочное среднее, (b) выборочную дисперсию, (c) выборочную медиану, (d) выборочную моду, (e) критические значения для  $(\bar{X} - \mu)/(s/\sqrt{n})$  и вероятности  $\alpha = 0.01, 0.05, 0.10, 0.20$ .

#### Домашняя работа 4. Решения

**Задача 1.** Дано бета-распределение с параметрами  $\alpha = 3$ ,  $\beta = 2$ ,

**Найти:**

(a)  $\mu$ , (b)  $\sigma$ , (c)  $\gamma_1$  и (d) моду.

**Решение.** Бета-плотность даётся формулой

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1, \text{ with } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (. \text{ 12.2}).$$

Для заданных параметров получаем  $B(3, 2) = \frac{2! \cdot 1!}{4!} = \frac{1}{12}$  и  $f(x) = 12x^2(1-x)$ . Таким образом

(a)

$$\mu = \int_{-\infty}^{\infty} x f(x) dx = 12 \int_0^1 x^3 (1-x) dx = 12 \left( \frac{1}{4} - \frac{1}{5} \right) = \frac{3}{5}.$$

(b)

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = 12 \int_0^1 x^4 (1-x) dx - \mu^2 = \frac{2}{5} - \left( \frac{3}{5} \right)^2 = \frac{1}{25},$$
$$\sigma = \frac{1}{5}.$$

(c)

$$\begin{aligned} \gamma_1 &= \frac{1}{\sigma^3} \int_0^1 (x - \mu)^3 f(x) dx = \frac{1}{\sigma^3} \int_0^1 (x^3 - 3\mu x^2 + 3\mu^2 x - \mu^3) f(x) dx = \\ &= \frac{1}{\sigma^3} \left[ \int_0^1 x^3 f(x) dx - 3\mu \int_0^1 x^2 f(x) dx + 3\mu^2 \int_0^1 x f(x) dx - \mu^3 \int_0^1 f(x) dx \right]. \end{aligned}$$

Из вышеизложенного видно, что

$$\int_0^1 x^2 f(x) dx = \frac{2}{5}, \quad \int_0^1 x f(x) dx = \mu = \frac{3}{5}, \quad \sigma = \frac{1}{5}.$$

Благодаря условию нормировки  $\int_0^1 x f(x) dx = 1$ , нам нужно вычислить только третий момент:

$$\int_0^1 x^3 f(x) dx = 12 \int_0^1 (x^5 - x^6) dx = 12 \left( \frac{1}{6} - \frac{1}{7} \right) = \frac{2}{7}.$$

В результате получаем

$$\gamma_1 = 5^3 \left( \frac{2}{7} - \frac{18}{25} + \frac{81}{125} - \frac{27}{125} \right) = -\frac{2}{7}.$$

(d)

Чтобы найти положение максимуму плотности вычислим производные:

$$f'(x) = 12(2x - 3x^2), \quad f''(x) = 12(2 - 6x).$$

Решение уравнений

$$f'(x) = 0, \quad f''(x) < 0$$

даёт моду  $2/3$ .

**Задача 2.** Пусть  $X$  – нормальная случайная величина с  $\mu = 1$  и  $\sigma = 2$ .

**Найти:**

(a)  $P(X < 0)$ , (b)  $P(X < -1)$ , (c)  $P(-1 < X < 1)$ , (d)  $P(X^2 > 1)$ .

**Решение.** Любая нормальная случайная величина с параметрами  $\mu$  и  $\sigma$  представляется в виде

$$X = \mu + \sigma Z,$$

где  $Z$  – стандартная нормальная случайная величина, представленная в Table A.3 (Page 630). Используя это соотношение, получим

$$P(X < x) = P(\mu + \sigma Z < x) = P\left(Z < \frac{x - \mu}{\sigma}\right) = P(Z < z(x))$$

где  $z_x = (x - \mu)/\sigma$ . Подставляя сюда заданные числа и используя таблицу, получаем:

(a)

$$z(0) = (0 - 1)/2 = -1/2, \quad P(X < 0) = P(Z < -0.5) = 0.3085.$$

(b)

$$z(-1) = (-1 - 1)/2 = -1, \quad P(X < -1) = P(Z < -1) = 0.1587.$$

(c)

$$z(-1) = -1, \quad z(1) = 0,$$

$$P(-1 < X < 1) = P(z(-1) < Z < z(1)) = P(-1 < Z < 0) = P(Z < 0) - P(-1) = 0.5 - 0.1587 = 0.3413.$$

(d)

$$P(X^2 > 1) = P(X > 1) + P(X < -1) = 1 - P(-1 < X < 1) = 1 - 0.3413 = 0.6587.$$

**Задача 3.** Пусть  $X_1, X_2, X_3$  – три независимые случайные величины распределённые, как и ранее.

**Найти:**

(a)  $P(\bar{X} < 0)$ , (b)  $P(\bar{X} < -1)$ , (c)  $P(-1 < \bar{X} < 1)$ , (d)  $P(\bar{X}^2 > 1)$ .

**Решение.** Если  $X_j$  – независимые нормально распределённые случайные величины, тогда их сумма и выборочное среднее также нормально распределены и мы можем записать

$$\bar{X} \equiv \frac{1}{n} \sum_{j=1}^n X_j = \mu_{\bar{X}} + \sigma_{\bar{X}} Z,$$

где  $Z$  – стандартная нормальная величина с распределением вероятностей, представленным в таблице Table A.3 (Page 670). Параметры величины  $\bar{X}$

$$E\bar{X} = E\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n} n\mu = \mu = 1,$$

и

$$\text{Var}\bar{X} = \text{Var}\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n^2} n \sigma^2 = \frac{1}{n} \sigma, \quad \sigma_{\bar{X}} = \frac{1}{\sqrt{n}} \sigma = \frac{2}{\sqrt{3}},$$

следовательно,  $\bar{X} = 1 + (2/\sqrt{3})Z$ . Остальная часть решения получается аналогично.

(a)

$$P(\bar{X} < 0) = P(1 + (2/\sqrt{3})Z < 0) = P(Z < -\sqrt{3}/2) = P(Z < -0.86) = 0.1949.$$

(b)

$$P(\bar{X} < -1) = P(1 + (2/\sqrt{3})Z < -1) = P(Z < -\sqrt{3}) = P(Z < -1.71) = 0.0436.$$

(c)

$$\begin{aligned} P(-1 < \bar{X} < 1) &= P(-1 < 1 + (2/\sqrt{3})Z < 1) = P(-2 < (2/\sqrt{3})Z < 0) = \\ &= P(Z < 0) - P(Z < -1.71) = 0.5 - 0.0436 = 0.4564. \end{aligned}$$

(d)

$$P(\bar{X}^2 > 1) = 1 - P(\bar{X}^2 < 1) = 1 - P(-1 < \bar{X} < 1) = 1 - 0.4564 = 0.5436.$$

**Задача 4.** Дана выборка {13, 34, 40, 47, 17, 34, 40, 47, 21, 34},

**найти:** (a) моду, (b) медиану, (c) среднее, (d) дисперсия.

**Решение.** Дана упорядоченная выборка

$$\{13, 17, 21, 34, 34, 34, 40, 40, 47, 47\}.$$

(a) Выборочная мода равна 34.

(b) Выборочная медиана 34.

(c) Выборочное среднее  $(13 + 17 + \dots + 47)/10 = 32.7$

(d) Выборочная дисперсия

$$s^2 = \frac{1}{9} \left[ (13^2 + 17^2 + \dots + 47^2) - 10 \cdot 32.7^2 \right] =$$

**Задача 5.** Пусть  $\{X_1, \dots, X_{17}\}$  – случайная выборка из нормальной генеральной совокупности с  $\sigma^2 = 2$ .

**Найти критические значения для выборочной дисперсии:** (a)  $s_{0.05}^2$ , (b)  $s_{0.1}^2$ , (c)  $s_{0.2}^2$ , (d)  $s_{0.5}^2$ .

**Решение.** Эта величина связана с  $\chi^2$  (см. Table in **17.3**), тогда

$$P(s^2 > s_{\alpha}^2) = P\left(\chi^2 > \frac{n-1}{\sigma^2} s_{\alpha}^2\right) = P(\chi^2 > \chi_{\alpha}^2) = \alpha,$$

где

$$\chi_{\alpha}^2 = \frac{n-1}{\sigma^2} s_{\alpha}^2.$$

Из таблицы Table **A.5** (Page 675) находим для  $v = n - 1 = 16$  и заданного  $\alpha$  критические значения

$$\chi_{0.05}^2 = 26.296, \chi_{0.10}^2 = 23.542, \chi_{0.20}^2 = 20.465, \chi_{0.50}^2 = 15.338.$$

Теперь, используя формулу

$$s_{\alpha}^2 = \frac{\sigma^2}{n-1} \chi_{\alpha}^2(n-1) = \frac{1}{8} \chi_{\alpha}^2(16)$$

получаем

$$(a) s_{0.05}^2 =, (b) s_{0.10}^2 =, (c) s_{0.20}^2 =, (d) s_{0.50}^2 = .$$

**Задача 6.** Пусть  $\{19, 24, 20, 21, 17, 23, 20, 25, 22\}$  – случайная выборка из нормальной генеральной совокупности с  $\mu = 21$ .

**Найти:** (a) выборочное среднее, (b) выборочную дисперсию, (c) выборочную медиану, (d) выборочную моду, (e) критические значения для  $(\bar{X} - \mu)/(s/\sqrt{n})$  и вероятности  $\alpha = 0.01, 0.05, 0.10, 0.20$ .

**Решение:** (a)  $\bar{X} = 21.2$ , (b)  $s^2 = 6.45$ .

Представляя выборку в упорядоченной форме

$$\{17, 19, 20, 20, 21, 22, 23, 24, 25\}$$

получим (c)  $x_{1/2} = 21$ , (d)  $m = 20$ .

Случайная величина  $(\bar{X} - \mu)/(s/\sqrt{n})$  имеет  $t$ -распределение с  $v = n - 1 = 8$  степенями свободы, и из таблицы **A.4** находим:  $t_{0.01}(8) = 2.896$ ,  $t_{0.05}(8) = 1.860$ ,  $t_{0.10}(8) = 1.397$ ,  $t_{0.20}(8) = 0.889$ .

Значение  $\mu = 21$  не является необходимым для этих вычислений, но будет полезно, если вы хотите знать вероятности  $P(x_1 < \bar{X} < x_2)$ .

## 19 Лекция 18. Интервальные оценки

### 19.1 Статистические выводы, оценки и проверка гипотез

Вспомните понятия *генеральная совокупность*, *выборка*, *объём выборки*. *Статистический вывод* – заключение о свойствах генеральной совокупности, сделанное на основе изучения выборки. Статистические выводы могут быть поделены на два больших класса: *оценивание* (например, оценка  $\mu$ , или  $\sigma$ , или других параметров генеральной совокупности) и *проверка гипотез* (ответ "да-нет").

Широко используются два метода оценки: *точечная оценка* и *интервальная оценка*. Статистики, которые мы рассматривали ранее

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j,$$
$$S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$$

и т.д., принадлежат к первому типу: каждая из них - число, представленное точкой на действительной числовой оси.

Напомним, что точечная оценка  $\hat{\theta} = \Theta(X_1, X_2, \dots, X_n)$  называется *несмещённой* оценкой параметра  $\theta$ , если  $E\hat{\theta} = E\Theta(X_1, X_2, \dots, X_n) = \theta$ .  $\bar{X}$  и  $S^2$  – несмещённые оценки величин  $\mu$  и  $\sigma^2$  соответственно.

Если у нас имеются две оценки  $\hat{\theta}_1$  и  $\hat{\theta}_2$  одного и того же параметра  $\theta$  и  $\sigma_1^2 < \sigma_2^2$  мы говорим, что  $\hat{\theta}_1$  более *эффективна*, чем  $\hat{\theta}_2$ .

Перейдём теперь к интервальным оценкам.

### 19.2 Понятие интервальной оценки

Точечная оценка параметра содержит случайную ошибку. Например,

$$\bar{X} = \mu + \epsilon,$$

где  $\epsilon \stackrel{d}{=} (\sigma/\sqrt{n})Z$  – случайная ошибка. Для нормальной генеральной совокупности это непрерывная случайная величина. Вероятность, что непрерывная случайная величина примет какое-то конкретное значение  $x$  равна 0:

$$P(\epsilon = x) = \lim_{\Delta x \rightarrow 0} P(x < \epsilon < x + \Delta x) = \lim_{\Delta x \rightarrow 0} f_\epsilon(x) \Delta x = f_\epsilon(x) \cdot 0 = 0.$$

Таким образом

$$P(\bar{X} = \mu) = P(\epsilon = 0) = 0.$$

Мы никогда не сможем получить точное значение среднего генеральной совокупности. То, что мы можем получить по выборке – это оценку вероятности

$$P(\epsilon_1 < \epsilon < \epsilon_2) = P(\epsilon_1 < \bar{X} - \mu < \epsilon_2),$$

но нам не известна величина  $\mu$  (мы ищем её!).

Чтобы преодолеть эту трудность, мы зададим числовое значение этой вероятности. Обычно это значение берётся близким к 1 и записывается как  $1 - \alpha$  с достаточно малым  $\alpha$ :

$$P(\epsilon_1 < \bar{X} - \mu < \epsilon_2) = 1 - \alpha.$$



Переставляя члены в аргументе вероятности, мы перепишем равенство в виде

$$P(\mu_{\alpha/2}^L < \mu < \mu_{\alpha/2}^U) = 1 - \alpha \quad (1)$$

где  $\mu_{\alpha/2}^L = \bar{X} - \epsilon_2$  и  $\mu_{\alpha/2}^U = \bar{X} + \epsilon_1$  – нижняя и верхняя границы интервала  $(\mu_{\alpha/2}^L, \mu_{\alpha/2}^U)$ . Индексы обозначают вероятность для  $\mu < \mu_{\alpha/2}^L$  и  $\mu > \mu_{\alpha/2}^U$ . Благодаря равенству этих вероятностей уравнение (1) распадается на пару уравнений

$$P(\mu < \mu_{\alpha/2}^L) = \alpha/2$$

и

$$P(\mu > \mu_{\alpha/2}^U) = \alpha/2.$$

Таким образом, мы имеем два уравнения для двух неизвестных.

Вообще говоря, мы можем ввести для любого параметра генеральной совокупности  $\theta$  интервал  $(\theta_{\alpha/2}^L, \theta_{\alpha/2}^U)$ , подчиняющийся условию

$$P(\theta_{\alpha/2}^L < \theta < \theta_{\alpha/2}^U) = 1 - \alpha$$

с пределами, которые удовлетворяют уравнениям

$$P(\theta < \theta_{\alpha/2}^L) = \alpha/2 \quad (2a)$$

и

$$P(\theta > \theta_{\alpha/2}^U) = \alpha/2. \quad (2b)$$

Если нам удастся решить эти уравнения, мы получим интервал  $(\theta_{\alpha/2}^L, \theta_{\alpha/2}^U)$ , содержащий искомую величину  $\theta$  с вероятностью  $1 - \alpha$ . Такая оценка неизвестного параметра  $\theta$  называется *интервальной оценкой*, этот интервал называется  $(1 - \alpha)100\%$ -*доверительным интервалом*,  $1 - \alpha$  называется *доверительным коэффициентом*??????.

Далее мы рассмотрим эту процедуру более детально.

### 19.3 Доверительные интервалы для $\mu$ ( $\sigma$ известны)

Начнём с простейшего случая, когда стандартное отклонение генеральной совокупности известно, скажем, из предыдущих измерений.

Таким образом, имеется выборка  $X_1, X_2, \dots, X_n$  из нормальной генеральной совокупности с известной дисперсией  $\sigma$ , вычисленной  $\bar{X} = (1/n) \sum_{j=1}^n X_j$  и нужно найти 95%-й доверительный интервал для искомого генерального среднего  $\mu$ . Для того, чтобы достичь этой цели, нужно выразить левую часть уравнения (1) в явном виде и затем решить его. Из **15.2** мы знаем, что  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  распределена как стандартная нормальная величина  $Z$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{d}{=} Z. \quad (3)$$

Из таблицы Table **A.3** (page 670) найдём  $z_{\alpha/2}$ , подчиняющееся уравнению

$$P(Z > z_{\alpha/2}) = \alpha/2. \quad (4)$$

Использование (3) даёт

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > z_{\alpha/2}\right) = P(-\mu > -\bar{X} + (\sigma/\sqrt{n})z_{\alpha/2}) = P(\mu < \bar{X} - (\sigma/\sqrt{n})z_{\alpha/2}) = \alpha/2.$$

Сравнивая с (2а), мы получаем нижнюю границу доверительного интервала:

$$\mu_{\alpha/2}^L = \bar{X} - (\sigma/\sqrt{n})z_{\alpha/2}.$$

Благодаря симметрии

$$P(Z < -z_{\alpha/2}) = \alpha/2$$

и мы получаем

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) = P\left(-\mu < -\bar{X} - (\sigma/\sqrt{n})z_{\alpha/2}\right) = P\left(\mu > \bar{X} + (\sigma/\sqrt{n})z_{\alpha/2}\right) = \alpha/2.$$

Сравнивая это выражение с уравнением (2b), мы получаем верхнюю границу доверительного интервала:

$$\mu_{\alpha/2}^U = \bar{X} + (\sigma/\sqrt{n})z_{\alpha/2}.$$

Следовательно,  $(1 - \alpha)100\%$  доверительный интервал для  $\mu$  при известной генеральной дисперсии  $\sigma$

$$(\mu_{\alpha/2}^L, \mu_{\alpha/2}^U) = (\bar{X} - (\sigma/\sqrt{n})z_{\alpha/2}, \bar{X} + (\sigma/\sqrt{n})z_{\alpha/2}).$$

Пример: Если  $n = 100$ ,  $\sigma = 10$ ,  $\bar{X} = 2$  тогда для  $\alpha = 0.01$  (99%-й доверительный интервал) мы имеем  $(-0.575, 4.575)$  и для  $\alpha = 0.05$  (95%-й доверительный интервал) имеем  $(0.04, 3.96)$ .

## 19.4 Доверительные интервалы для $\mu$ : $\sigma$ неизвестна

В этом случае необходимо использовать выборочное стандартное отклонение  $s$  вместо генерального стандартного отклонения. Если объём выборки велик (скажем, больше чем 30), можно поступать как и прежде. Но если это не так, то нужно учитывать, что статистика  $\frac{\bar{X} - \mu}{s/\sqrt{n}}$  распределена согласно  $t$ -плотности и заменить  $z_{\alpha/2}$  на величины  $t_{\alpha/2}$ , которые могут быть найдены в таблице Table A.4 (page 672) для  $v = n - 1$ . В результате получим

$$\mu_{\alpha/2}^L = \bar{X} - t_{\alpha/2}s/\sqrt{n}, \quad \mu_{\alpha/2}^U = \bar{X} + t_{\alpha/2}s/\sqrt{n}.$$

**Пример:** Если  $n = 9$ ,  $s = 10$ ,  $\bar{X} = 2$  тогда для  $\alpha = 0.01$  (99% доверительный интервал) имеем  $t_{0.005}(v = 8) = 3.355$  и интервал  $(-9.2, 13.2)$  в то время как для  $\alpha = 0.05$  (95% доверительный интервал) имеем  $t_{0.025}(v = 8) = 2.306$  и интервал  $(-5.69, 9.69)$ .

## 19.5 Доверительные интервалы для $\sigma$

Теперь нужно учитывать соотношения (см. формулу (1) из 16.3)

$$s^2 \frac{n-1}{\sigma^2} \stackrel{d}{=} \chi^2(n-1). \quad (4)$$

Решение уравнения

$$P(\chi^2(n-1) > \chi_{\alpha/2}^2) = \alpha/2$$

может быть взято непосредственно из таблицы **A.5.** и мы получим

$$\begin{aligned} P(\chi^2(n-1) > \chi_{\alpha/2}^2) &= P\left(s^2 \frac{n-1}{\sigma^2} > \chi_{\alpha/2}^2(n-1)\right) = P\left(\frac{\sigma^2}{n-1} < \frac{s^2}{\chi_{\alpha/2}^2(n-1)}\right) = \\ &= P\left(\sigma < s \sqrt{\frac{n-1}{\chi_{\alpha/2}^2(n-1)}}\right) = \alpha/2. \end{aligned}$$

Сравнивая его с (2а), видим, что нижняя граница доверительного интервала

$$\sigma_{\alpha/2}^L = s \sqrt{\frac{n-1}{\chi_{\alpha/2}^2(n-1)}}.$$

Чтобы найти верхнюю границу, надо решить уравнение

$$P(\chi^2(n-1) < \chi^2) = \alpha/2. \quad (5)$$

Так как

$$P(\chi^2(n-1) < \chi^2) = 1 - P(\chi^2(n-1) > \chi^2),$$

уравнение (5) эквивалентно

$$P(\chi^2(n-1) > \chi^2) = 1 - \alpha/2.$$

Решение  $\chi^2$  последнего уравнения снова может быть найдено в таблице **A.5**:  $\chi^2 = \chi_{1-\alpha/2}^2(n-1)$ . Возвращаясь к уравнению (5)

$$P(\chi^2(n-1) < \chi_{1-\alpha/2}^2) = \alpha/2$$

и используя (4), получаем:

$$P\left(s^2 \frac{n-1}{\sigma^2} < \chi_{1-\alpha/2}^2\right) = \alpha/2.$$

Повторяя те же вычисления, что и для нижней границы, приходим к уравнению

$$P\left(\sigma > s \sqrt{\frac{n-1}{\chi_{1-\alpha/2}^2(n-1)}}\right) = \alpha/2$$

сравнивая которое с (2b), получаем верхнюю границу доверительного интервала для генерального стандартного отклонения:

$$\sigma_{\alpha/2}^U = s \sqrt{\frac{n-1}{\chi_{1-\alpha/2}^2(n-1)}}.$$

**Пример.** Некоторый изготовитель производит 10-мм болты. Ему известно, что диаметры производимых болтов могут отличаться как от 10 мм, так и друг от друга. Он берёт случайную выборку из 12 болтов и тщательно измеряет их диаметр. Результаты измерений (в миллиметрах): 10.05, 10.00, 10.02, 9.97, 10.07, 10.03, 9.98, 10.10, 9.95, 9.99, 10.00, 10.08.

**Шаг 1.** Выборочное стандартное отклонение диаметров болтов

$$s = \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n-1}} = \sqrt{\frac{1204.83 - (120.24)^2/12}{11}} = 0.047.$$

Изготовитель хочет определить 95% доверительный интервал; так что доверительный уровень равен 0.95 и  $\alpha = 1 - 0.95 = 0.05$ . Так как  $v = 12 - 1 = 11$ , обращаясь к таблице **A.5** он находит, что

**Шаг 2.**

$$\chi_{\alpha/2}^2 = \chi_{0.05/2}^2 = \chi_{0.025}^2 = 21.920$$

и

$$\chi_{1-\alpha/2}^2 = \chi_{1-0.05/2}^2 = \chi_{0.975}^2 = 3.816.$$

**Шаг 3.**

$$\sigma_{0.025}^L = 0.047 \sqrt{(11)/\chi_{0.025}^2} = 0.033, \quad \sigma_{0.025}^U = 0.047 \sqrt{(11)/\chi_{0.975}^2} = 0.080.$$

## 20 Лекция 19. Оценки по двум выборкам

### 20.1 Разность двух средних (дисперсии известны)

Предположим, имеются две независимые нормальные генеральные совокупности с распределениями  $f(x; \mu_1, \sigma_1)$  and  $f(x; \mu_2, \sigma_2)$ , дисперсии которых известны. Требуется оценить разность  $\mu_1 - \mu_2$ , если даны выборки объёмами  $n_1$  и  $n_2$  соответственно. Точечная несмещённая оценка разности  $\mu_1 - \mu_2$  даётся выражением  $\bar{X}_1 - \bar{X}_2$ . Для того, чтобы найти доверительный интервал, рассмотрим разность между выборочными средними:

$$\bar{X}_1 \stackrel{d}{=} \mu_1 + \frac{\sigma_1}{\sqrt{n_1}} Z_1$$

и

$$\bar{X}_2 \stackrel{d}{=} \mu_2 + \frac{\sigma_2}{\sqrt{n_2}} Z_2 :$$

$$\bar{X}_1 - \bar{X}_2 \stackrel{d}{=} \mu_1 - \mu_2 + \frac{\sigma_1}{\sqrt{n_1}} Z_1 - \frac{\sigma_2}{\sqrt{n_2}} Z_2.$$

Благодаря симметрии стандартных нормальных распределений  $-Z_2 \stackrel{d}{=} Z_2$ . Далее,

$$\frac{\sigma_1}{\sqrt{n_1}} Z_1 + \frac{\sigma_2}{\sqrt{n_2}} Z_2 \stackrel{d}{=} N(0, \sigma_1/\sqrt{n_1}) + N(0, \sigma_2/\sqrt{n_2}) \stackrel{d}{=} N\left(0, \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right) \stackrel{d}{=} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} Z.$$

В результате имеем:

$$\bar{X}_1 - \bar{X}_2 \stackrel{d}{=} \mu_1 - \mu_2 + \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} Z.$$

Это следует из того, что Из

$$\mu_1 - \mu_2 \stackrel{d}{=} \bar{X}_1 - \bar{X}_2 - \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} Z$$

следует, что

$$\left( \bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \right)$$

покрывает неизвестное значение разности  $\mu_1 - \mu_2$  с вероятностью  $1 - \alpha$ .

### 20.2 Разность между средними (Дисперсии одинаковы, но неизвестны)

Как следует из вышеприведённого

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \stackrel{d}{=} Z.$$

При  $\sigma_1 = \sigma_2 = \sigma$

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{1/n_1 + 1/n_2}} \stackrel{d}{=} Z.$$

Но  $\sigma$  неизвестно и приходится использовать выборочные дисперсии. Мы можем использовать любую из двух:  $s_1^2$  and  $s_2^2$ , но самое эффективное – использовать обе дисперсии для вычисления так называемой *объединённой pooled* выборочной дисперсии:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

В результате, получаем

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} \stackrel{d}{=} T(n_1 + n_2 - 2).$$

Доверительный интервал для  $\mu_1 - \mu_2$

$$\left( (\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}(n_1 + n_2 - 2)s_p \sqrt{1/n_1 + 1/n_2}, (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}(n_1 + n_2 - 2)s_p \sqrt{1/n_1 + 1/n_2} \right).$$

### 20.3 Разность между средними (дисперсии неизвестны, но не обязательно одинаковы)

Возьмём снова формулу

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \stackrel{d}{=} Z.$$

Если  $\sigma_1 \neq \sigma_2$ , мы не можем использовать объединённую pooled??? выборочную дисперсию и должны заменить  $\sigma_1^2$  и  $\sigma_2^2$  на выборочные дисперсии  $s_1^2$  и  $s_2^2$  соответственно. Результат будет распределяться согласно  $t$ -плотности,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \stackrel{d}{=} T(v),$$

с числом степеней свободы

$$v = \frac{[s_1^2/n_1 + s_2^2/n_2]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Граничные точки доверительного интервала для  $\mu_1 - \mu_2$  определяются числами

$$(\mu_1 - \mu_2)^{U,L} = (\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}(v) \sqrt{s_1^2/n_1 + s_2^2/n_2},$$

округлёнными до ближайшего целого числа.

### 20.4 Отношение дисперсий

Снова имеются две выборки из двух различных генеральных совокупностей и требуется оценить отношение их дисперсий  $\sigma_1/\sigma_2$ . Точечная оценка этого отношения  $s_1^2/s_2^2$ , но нам нужна интервальная оценка. Чтобы сделать это, мы должны рассмотреть распределение отношения  $s_1^2/s_2^2$ . Согласно **17.3**

$$s_1^2/s_2^2 \stackrel{d}{=} (\sigma_1^2/\sigma_2^2)F(n_1 - 1, n_2 - 1).$$

Таблица Table **A.6** (page 676) содержит критическое значение этого  $F$ -распределения:

$$P(F > f_{\alpha}(v_1, v_2)) = \alpha.$$

Чтобы найти доверительный интервал, зададим вероятность

$$P((\sigma_1^2/\sigma_2^2)^L < (\sigma_1^2/\sigma_2^2) < (\sigma_1^2/\sigma_2^2)^U) = 1 - \alpha,$$

вычислим число степеней свободы  $v_1 = n_1 - 1$ ,  $v_2 = n_2 - 1$  и найдём  $f_{\alpha/2}(v_1, v_2)$  и  $f_{1-\alpha/2}(v_1, v_2)$  из таблицы.

Начнём с верхней границы для  $F$  (как будет видно далее, она относится к нижней границе для  $(\sigma_1^2/\sigma_2^2)$ ):

$$P(F > f_{\alpha/2}(v_1, v_2)) = \alpha/2,$$

$$\begin{aligned} P(F > f_{\alpha/2}(v_1, v_2)) &= P((s_1^2/s_2^2)(\sigma_2^2/\sigma_1^2) > f_{\alpha/2}(v_1, v_2)) = P((\sigma_2^2/\sigma_1^2) > (s_2^2/s_1^2)f_{\alpha/2}(v_1, v_2)) = \\ &= P((\sigma_1^2/\sigma_2^2) < x) = \alpha/2, \quad x = (s_1^2/s_2^2)/f_{\alpha/2}(n_1 - 1, n_2 - 1). \end{aligned}$$

Эта формула определяет значение  $x$ , оставляющее небольшую вероятность  $\alpha/2$  слева от доверительного интервала, то есть это *нижняя граница для  $(\sigma_1^2/\sigma_2^2)$* :

$$x = (\sigma_1^2/\sigma_2^2)^L = (s_1^2/s_2^2)/f_{\alpha/2}(n_1 - 1, n_2 - 1).$$

Теперь рассмотрим нижнюю границу для  $F$ :

$$P(F < f_{1-\alpha/2}(v_1, v_2)) = \alpha/2,$$

$$\begin{aligned} P(F < f_{1-\alpha/2}(v_1, v_2)) &= P((s_1^2/s_2^2)(\sigma_2^2/\sigma_1^2) < f_{1-\alpha/2}(v_1, v_2)) = P((\sigma_2^2/\sigma_1^2) < (s_2^2/s_1^2)f_{1-\alpha/2}(v_1, v_2)) = \\ &= P((\sigma_1^2/\sigma_2^2) > x) = \alpha/2, \quad x = (s_1^2/s_2^2)/f_{1-\alpha/2}(n_1 - 1, n_2 - 1). \end{aligned}$$

Эта формула определяет значение  $x$ , оставляющее небольшую вероятность  $\alpha/2$  справа от доверительного интервала, то есть это *верхняя граница для  $(\sigma_1^2/\sigma_2^2)$* :

$$x = (\sigma_1^2/\sigma_2^2)^U = (s_1^2/s_2^2)/f_{1-\alpha/2}(n_1 - 1, n_2 - 1).$$

**Замечание:** Существует теорема (Theorem 8.7, page 225) утверждающая, что

$$f_{1-\alpha/2}(v_1, v_2) = \frac{1}{f_{\alpha/2}}(v_2, v_1)$$

(обратите внимание на взаимозаменяемость аргументов). Таким образом, верхняя граница может быть переписана как

$$(\sigma_1^2/\sigma_2^2)^U = (s_1^2/s_2^2)f_{\alpha/2}(n_2 - 1, n_1 - 1).$$

## 21 Lecture 20. Типичные задачи

### 21.1 Оценка пропорции

Представим, что генеральная совокупность представляет собой набор шаров, занумерованных действительными числами и помещённых в закрытый ящик. Говоря о нормальной совокупности, то есть о генеральной совокупности с плотностью вероятности

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty,$$

мы предполагаем, что на этих шарах представлены все реальные числа. Объём генеральной совокупности в этом случае бесконечно большой, более того, несчётный.

Рассмотрим другую предельную ситуацию: только одно из двух чисел 1 (*успех*) и 0 (*неудача*) представлено на каждом из шаров. Эта генеральная совокупность характеризуется единственным параметром  $p = P(X = 1)$ , часто называемым *пропорцией proportion*?????. Пусть  $n$  – объём случайной выборки, полученной из данной генеральной совокупности. Принимая во внимание, что *число успехов*

$$S_n = \sum_{j=1}^n X_j$$

в  $n$  испытаниях Бернулли является случайной величиной, распределённой по биномиальному закону со средним  $\mu = np$  и дисперсией  $\sigma^2 = np(1-p)$ , получим для выборочного среднего  $\bar{X} = (1/n)S_n$

$$\mu_{\bar{X}} = p, \quad \sigma_{\bar{X}}^2 = \frac{p(1-p)}{n}.$$

Несмотря на неправильный характер генеральной совокупности, если  $n$  достаточно велико ( $n > 30$ ), тогда сумма  $S_n$  и соответственно выборочное среднее  $\bar{X}$  распределены приблизительно по нормальному закону с параметрами  $\mu$  и  $\sigma$ , как показано выше (см. Центральную Предельную Теорему). Следовательно, мы можем использовать для точечной и интервальной оценки генеральной пропорции the population proportion???? следующее выражение:

$$\hat{p} = \frac{1}{n}S_n,$$

$$(p^L, p^U)_\alpha \simeq (\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}).$$

### 21.2 Пример с небольшой генеральной совокупностью

Пусть рассматриваемая генеральная совокупность имеет небольшой объём. Например, имеется пять шаров, на трёх из которых написана цифра 1 и на остальных 0. Генеральное среднее (генеральная пропорция population proportion????) и генеральная дисперсия соответственно

$$\mu = p = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot \frac{2}{5} + 1 \cdot \frac{3}{5} = 0.6,$$

$$\sigma^2 = 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) - \mu^2 = 0.6 - 0.36 = 0.24.$$

но мы не знаем эту пропорцию и должны оценить её, используя случайную выборку объёма  $n = 2$ .

Отметим каждый шар его собственным номером, как в таблице:

Номер шара	1	2	3	4	5
Значение $X$	1	1	1	0	0

Теперь выпишем все возможные выборки с их оценками

$$\hat{p} = \bar{X} = \frac{X_1 + X_2}{2}$$

и

$$s = \sqrt{s^2} = \sqrt{[1/(n-1)][\sum X^2 - n\bar{X}^2]} = \sqrt{\frac{1}{2-1} \left( X_1^2 + X_2^2 - 2 \left( \frac{X_1 + X_2}{2} \right)^2 \right)} = \frac{|X_1 - X_2|}{\sqrt{2}}.$$

Составим таблицу:

Номер выборки	Номера шаров	Случайные величины $X_1, X_2$	Выборочные средние $\bar{X} = \hat{p}$	Выборочное ст. откл. $s$
1	1, 2	1, 1	1	0
2	1, 3	1, 1	1	0
3	1, 4	1, 0	0.5	0.71
4	1, 5	1, 0	0.5	0.71
5	2, 3	1, 1	1	0
6	2, 4	1, 0	0.5	0.71
7	2, 5	1, 0	0.5	0.71
8	3, 4	1, 0	0.5	0.71
9	3, 5	1, 0	0.5	0.71
10	4, 5	0, 0	0	0

Эти суммы равновероятны и мы можем вычислить распределения вероятностей для для выборочной пропорции proportion????:

Возможные значения $x$	0	0.5	1
Вероятность $P(\hat{p} = x)$	0.1	0.6	0.3

и выборочное стандартное отклонение:

Возможные значения $x$	0	0.71
Вероятность $P(s = x)$	0.4	0.6

Имея в своём распоряжении эти распределения, мы можем вычислить любые характеристики этих оценок, например

$$P(\hat{p} < p) = P(\hat{p} < 0.6) = P(\hat{p} = 0) + P(\hat{p} = 0.5) = 0.1 + 0.6 = 0.7$$

или

$$P(s > \sigma) = P(s > 0.49) = P(s = 0.71) = 0.6.$$

## 21.3 Пример со среднеквадратичной ошибкой

Предположим, требуется оценить нормальное генеральное среднее на основе выборки {24.4, 30.6, 26.4, 26.8, 33.5, 32.2, 32.4, 13.9, 27.1}. Делаем следующее:

$$\bar{X} = (1/9)(24.4 + \dots + 27.1) = 27.5.$$



Но хотелось бы знать и возможную ошибку этой оценки.

-Хотите ли вы найти доверительный интервал? - спрашиваете вы.

-Нет, некоторой оценки ошибки будет достаточно,- отвечают вам.

-Хотите ли вы знать генеральное стандартное отклонение? - снова спрашиваете вы.

-Да, хотим. Оно равно 6.9. Нам не известно только генеральное среднее.

-ОК!

Тогда вы берёте формулу

$$\bar{X} \stackrel{d}{=} \mu + \frac{\sigma}{\sqrt{n}}Z,$$

переписываете её в виде

$$\bar{X} - \mu \stackrel{d}{=} \frac{\sigma}{\sqrt{n}}Z,$$

называете правую часть случайно ошибкой, обозначаете её  $\varepsilon$ ,

$$\varepsilon = \frac{\sigma}{\sqrt{n}}Z,$$

и показываете, что

$$\mu_\varepsilon = E\varepsilon = 0.$$

-Означает ли это, что ошибка отсутствует? - спрашивают вас.

- Нет,- отвечаете вы, - это означает, что она распределена симметрично, и оба знака + и - одинаково вероятны.

-Можете ли вы тогда вычислить среднее её абсолютного значения?

-Конечно, я могу, но, может быть, вас удовлетворит СКО ?

-Что это?

-Это среднеквадратичная ошибка: сначала вычисляется  $\varepsilon^2$ , затем эта величина усредняется и после этого находится квадратный корень:

$$MSE = \sqrt{E\varepsilon^2}.$$

Все используют её.

-ОК!

Вы вычисляете:

$$\varepsilon^2 = \frac{\sigma^2}{n}Z^2,$$

$$E\varepsilon^2 = \frac{\sigma^2}{n}EZ^2 = \frac{\sigma^2}{n},$$

так как

$$EZ^2 = E(Z - 0)^2 = E(Z - EZ)^2 = \text{Var}Z = 1.$$

Затем выдаёте результат:

$$CKO = \sqrt{E\varepsilon^2} = \frac{\sigma}{\sqrt{n}} = \frac{6.9}{3} = 2.30.$$

Три недели спустя:

-Вы знаете, наш начальник не мог понять, почему вы так сделали, и просил вычислить среднее абсолютной ошибки.

- Нет проблем!

Вы пишете следующую цепочку формул:

$$E|\varepsilon| = \frac{\sigma}{\sqrt{n}}E|Z|,$$

$$\begin{aligned} E|Z| &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \left( \int_0^{\infty} z e^{-z^2/2} dz + \int_{-\infty}^0 (-z) e^{-z^2/2} dz \right) = \\ &= \frac{1}{\sqrt{2\pi}} \left( \int_0^{\infty} z e^{-z^2/2} dz + \int_0^{\infty} z e^{-z^2/2} dz \right) = 2 \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z e^{-z^2/2} dz = \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-t} dt = \sqrt{\frac{2}{\pi}} = 0.80, \end{aligned}$$

и показываете результат:

$$E|\varepsilon| = \frac{\sigma}{\sqrt{n}} E|Z| = 2.30 \cdot 0.80 = 1.84.$$

- О, она значительно меньше, чем раньше. Как мы сможем объяснить это различие нашему начальнику?

- Скажите ему, что оба результата правильны, просто последний интервал  $(\mu - 1.84, \mu + 1.84)$  содержит меньшую вероятность, чем первый  $(\mu - 2.30, \mu + 2.30)$ , поскольку

$$P(-\sigma_Z < Z < \sigma_Z) =$$

Лучше всего использовать доверительные интервалы.

- Спасибо! До свидания.

Три месяца спустя.

- Извините, это снова мы. Наш начальник забыл, является ли значение 6.9 для генерального стандартного отклонения правильным или нет, и просил бы вас повторить расчёты без использования этого числа. Можете Вы это сделать?

-Конечно, могу. Я закончил ??? I graduated CWRU.

Вы возвращаетесь к той же выборке, вычисляете выборочное стандартное отклонение

$$s = \sqrt{(1/8)[24.4^2 + \dots + 27.1^2 - 9 \cdot 27.5^2]} = 6.0$$

и используете формулу

$$\bar{X} \stackrel{d}{=} \mu + \frac{s}{\sqrt{n}} T(n-1),$$

где  $T(n-1)$  – случайная величина, подчиняющаяся  $t$ -распределению с  $n-1$  степенями свободы. Повторяя расчёты с  $s$  вместо  $\sigma$  и  $T(n-1)$  вместо  $Z$ , получаете:

$$E\varepsilon^2 = \frac{s^2}{n} ET^2.$$

Для  $v = n - 1 = 9 - 1 = 8$   $ET^2 = \text{Var}(T) = \sigma_T^2 = v/(v-2) = \frac{4}{3}$  и отдаёте результат:

$$\text{MSE} = \sqrt{E\varepsilon^2} = \frac{s}{\sqrt{n}} \sigma_{T(n-1)} = \frac{6}{3} \cdot \frac{2}{\sqrt{3}} = 2.31.$$

## 21.4 Пример со стандартным отклонением

Имея случайную выборку объёма  $n$ , полученную из нормальной генеральной совокупности, мы можем получить  $(1 - \alpha)$  доверительный интервал для  $\sigma^2$  с использованием результатов пункта 18.5:

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} < s^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}. \quad (1)$$

Ясно, что этот  $1 - \alpha$  доверительный интервал для  $\sigma^2$  может быть преобразован в соответствующий  $1 - \alpha$  доверительный интервал для  $\sigma$  путём вычисления квадратных корней.

Чтобы проиллюстрировать использование этой формулы, предположим, что в 16 тестовых запусках расход бензина экспериментальным двигателем дал выборочное стандартное отклонение в объеме  $s = 22$  галлонов, и мы хотим определить доверительный интервал для  $\sigma$  в качестве показателя изменчивости расхода бензина этим двигателем. Предположим, далее, что уровень значимости degree of confidence равен 0.99, то есть  $\alpha = 0.01$ . Допуская далее, что рассматриваемые данные можно считать случайной выборкой, полученной из нормальной генеральной совокупности, подставляем  $n = 16$ ,  $s = 2.2$ ,  $\chi^2_{0.005}(15) = 32.8$  и  $\chi^2_{0.995} = 4.60$  в (1), получая

$$\frac{15 \cdot 2.2^2}{32.80} < \sigma^2 < \frac{15 \cdot (2.2)^2}{4.60}$$

и, далее,

$$2.21 < \sigma < 15.78.$$

Вычисляя квадратные корни, находим соответствующие уровню значимости 0.99 границы для  $\sigma$ :

$$1.49 < \sigma < 3.97$$

галлонов. Таким образом, мы можем утверждать с вероятностью  $1 - \alpha$ , что расход бензина может наблюдаться в пределах между 1.49 и 3.97 галлонами.

## STAT 312 Spring

### Домашняя работа 5

(must be returned on Apr 12, 12:30 p.m., CLPP 108)

**Problem 1.** Предположим, генеральная совокупность состоит из значений роста пяти начинающих игроков мужской баскетбольной команды. Рост, в дюймах, – 76, 78, 79, 81, 86. Два значения выбираются случайно, чтобы оценить генеральное среднее.

**Найти:** (а) генеральное среднее  $\mu$  и стандартное отклонение  $\sigma$ ; (b) выборочное среднее  $\bar{X}$  и выборочное стандартное отклонение  $s$  всех возможных выборок объёма 2 (постройте таблицу "выборка – выборочное среднее – выборочное стандартное отклонение"); (с) вероятности событий  $|\bar{X} - \mu| \leq 1$ ,  $|\bar{X} - \mu| \leq 2$ ,  $|\bar{X} - \mu| \leq 3$ ; (d) вероятности событий  $s < 1$ ,  $s < 1.5$ ,  $s < 2$ .

#### Задача 2.

Случайная выборка из 40 новых передвижных домов new mobile homes даёт цены, в тысячах долларов, представленные в таблице:

24.4	30.6	26.4	26.8	33.5	32.2	32.4	13.9
24.4	29.3	26.2	14.1	21.4	20.0	33.0	17.6
24.8	27.0	22.8	18.8	35.1	26.7	22.1	37.2
31.9	24.0	28.4	15.8	29.3	31.4	22.8	8.4
24.7	16.6	31.1	13.9	16.8	29.5	17.0	9.9

Как известно из предыдущих исследований, генеральное стандартное отклонение цен равно 7.2 тысячи долларов.

**Найти:** (а) выборочную среднюю цену  $\bar{X}$ , (b) её среднеквадратичную ошибку  $\sqrt{\epsilon^2}$ , (с) выборочное стандартное отклонение  $s$ , (d) 95% доверительный интервал для  $\mu$ .

**Задача 3.** Чтобы оценить средний срок вынашивания щенков домашними собаками, случайным образом были выбраны 15 собак, за которыми проводились наблюдения в течение беременности. Их сроки вынашивания, в днях, представлены в таблице:

62.0	61.4	59.8	62.2	60.3
60.4	59.4	60.2	60.4	60.8
61.8	59.2	61.1	60.4	60.9

**Найти:** (а) выборочное среднее  $\bar{X}$ , (b) среднеквадратичную ошибку величины  $\bar{X}$ , (с) выборочную дисперсию  $s$ , (d) 95% доверительный интервал для  $\mu$ .

**Задача 4.** Производитель наручных часов утверждает, что недельная ошибка в показаниях часов, которые он производит имеет стандартное отклонение около 1 секунды. Чтобы проверить это утверждение, на 20 случайно отобранных часах было установлено правильное время. После истечения 1 недели, ошибки были зафиксированы. Результаты, в секундах, выглядят следующим образом:

0.6	2.3	2.0	-2.1	-1.4
-0.5	1.5	-0.3	0.4	0.6
0.4	-2.2	0.7	0.5	-1.3
-2.0	2.6	-0.8	1.0	-0.6

**Найти:** (a) выборочное среднее, (b) выборочное стандартное отклонение, (c) 95% доверительный интервал для генерального среднего, (d) 90% доверительный интервал для генерального стандартного отклонения.

**Задача 5.** Компания производит консервированные томаты в банках весом около 14 унций. Веса десяти случайно отобранных банок показаны в следующей таблице.

13.85	13.95	13.90	13.49	14.17
14.33	14.03	13.48	14.27	14.19

**Найти:** (a) выборочное среднее; (b) выборочное стандартное отклонение; (c) 90% доверительный интервал для генерального среднего  $\mu$ , (d) 90% доверительный интервал для генерального стандартного отклонения  $\sigma$ .

**Задача 6.** Данные в таблице показывают потребление протеина, в граммах, за 24-часовой период для независимых случайных выборок из 15 человек с доходами ниже прожиточного минимума и 10 человек с доходами выше прожиточного минимума:

Ниже прожиточного минимума			Выше прожиточного минимума	
51.4	49.7	72.0	86.0	69.0
76.7	65.8	55.0	59.7	80.2
73.7	62.1	79.7	68.6	78.1
66.2	75.8	65.4	98.6	69.8
65.5	62.0	73.3	87.7	77.2

**Найти:** (a) выборочное среднее  $\mu_1$  и стандартное отклонение  $\sigma_1$  для первой выборки, (b) выборочное среднее  $\mu_2$  и стандартное отклонение  $\sigma_2$  для второй выборки, (c) pooled??? объединённое выборочное стандартное отклонение  $s_p$ , (d) 95% доверительный интервал для разности  $\mu_1 - \mu_2$  между средним потреблением протеина всеми людьми с доходами ниже прожиточного минимума и средним потреблением протеина всеми людьми с доходами выше прожиточного минимума.

## HOMEWORK 5. Решения

**Problem 1.** Предположим, генеральная совокупность состоит из значений роста пяти начинающих игроков мужской баскетбольной команды. Рост, в дюймах, – 76, 78, 79, 81, 86. Два значения выбираются случайно, чтобы оценить генеральное среднее.

**Найти:** (a) генеральное среднее  $\mu$  и стандартное отклонение  $\sigma$ ; (b) выборочное среднее  $\bar{X}$  и выборочное стандартное отклонение  $s$  всех возможных выборок объёма 2 (постройте таблицу "выборка – выборочное среднее – выборочное стандартное отклонение"); (c) вероятности событий  $|\bar{X} - \mu| \leq 1$ ,  $|\bar{X} - \mu| \leq 2$ ,  $|\bar{X} - \mu| \leq 3$ ; (d) вероятности событий  $s < 1$ ,  $s < 1.5$ ,  $s < 2$ .

**Решение:**

(a)

$$\mu = (1/5)(76 + 78 + 79 + 81 + 86) = 80$$

$$\sigma = \sqrt{(1/n)\sum x^2 - \mu^2} = \sqrt{(1/5)(76^2 + 78^2 + 79^2 + 81^2 + 86^2) - 80^2} = \sqrt{11.6} \approx 3.406.$$

(b) Для  $n = 2$

$$\bar{X} = (1/2)(X_1 + X_2),$$

и

$$s = \sqrt{[1/(n-1)][\sum X^2 - n\bar{X}^2]} = (1/\sqrt{2})|X_1 - X_2|.$$

Выборка	$\bar{X}$	$s$
76, 78	77.0	1.41
76, 79	77.5	2.12
76, 81	78.5	3.54
76, 86	81.0	7.07
78, 79	78.5	0.71
78, 81	79.5	2.12
78, 86	82.0	5.66
79, 81	80.0	1.41
79, 86	82.5	4.95
81, 86	83.5	3.54

(c)

$$\begin{aligned} P(|\bar{X} - \mu| \leq 1) &= P(\mu - 1 \leq \bar{X} \leq \mu + 1) = P(80 - 1 \leq \bar{X} \leq 80 + 1) = \\ &= P(79 \leq \bar{X} \leq 81) = \frac{n(79 \leq \bar{X} \leq 81)}{n(\text{all})} = \frac{3}{10} = 0.3 \end{aligned}$$

(только три из десяти исходов принадлежат к этому интервалу: 79.5, 80 и 81). Аналогично:

$$P(|\bar{X} - \mu| \leq 2) = 0.6, \quad P(|\bar{X} - \mu| \leq 3) = 0.9.$$

(d) Вероятность

$$P(s < 1) = \frac{n(s < 1)}{n(\text{all})} = 0.1.$$

Аналогично,

$$P(s < 1.5) = 0.3, P(s < 2) = 0.3.$$

### Задача 2.

Случайная выборка из 40 новых передвижных домов new mobile homes даёт цены, в тысячах долларов, представленные в таблице:

24.4	30.6	26.4	26.8	33.5	32.2	32.4	13.9
24.4	29.3	26.2	14.1	21.4	20.0	33.0	17.6
24.8	27.0	22.8	18.8	35.1	26.7	22.1	37.2
31.9	24.0	28.4	15.8	29.3	31.4	22.8	8.4
24.7	16.6	31.1	13.9	16.8	29.5	17.0	9.9

Как известно из предыдущих исследований, генеральное стандартное отклонение цен равно 7.2 тысячи долларов.

**Найти:** (a) выборочную среднюю цену  $\bar{X}$ , (b) её среднеквадратичную ошибку  $\sqrt{\epsilon^2}$ , (c) выборочное стандартное отклонение  $s$ , (d) 95% доверительный интервал для  $\mu$ .

**Решение:**

(a)

$$\bar{X} \approx 24.3.$$

(b)

$$\sqrt{\epsilon^2} = \sqrt{\frac{\sigma^2 Z^2}{n}} = \frac{\sigma}{\sqrt{n}} \sqrt{Z^2} = \frac{\sigma}{\sqrt{n}} \approx 1.14.$$

(c)

$$s^2 = \frac{1}{n-1} [\sum X_j^2 - n\bar{X}] = \frac{1}{39} [(24.4)^2 + \dots + (9.9)^2 - 40(24.3)^2] = 51.8, \quad s = 7.20.$$

(d)

$$\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left( 24.3 - z_{0.05/2} \frac{7.2}{\sqrt{40}}, 24.3 + z_{0.05/2} \frac{7.2}{\sqrt{40}} \right).$$

Из таблицы Table A.4 (нижняя строка) находим  $z_{0.025} = 1.96$  и получаем

$$(\mu^L, \mu^U) = (22.1, 26.5).$$

**Задача 3.** Чтобы оценить средний срок вынашивания щенков домашними собаками, случайным образом были выбраны 15 собак, за которыми проводились наблюдения в течение беременности. Их сроки вынашивания, в днях, представлены в таблице:

62.0	61.4	59.8	62.2	60.3
60.4	59.4	60.2	60.4	60.8
61.8	59.2	61.1	60.4	60.9

**Найти:** (a) выборочное среднее  $\bar{X}$ , (b) среднеквадратичную ошибку величины  $\bar{X}$ , (c) выборочную дисперсию  $s$ , (d) 95% доверительный интервал для  $\mu$ .

**Решение:**

(a)  $\bar{X} = 60.69$ .

(b)  $\sqrt{\epsilon^2} = 0.27$

(c)  $s = 0.90$ .

(d) В этот раз  $\sigma$  не известна и мы должны использовать вместо неё  $s$ . Объём выборки меньше, чем 30, так что мы должны использовать  $t$ -распределение. В результате, имеем  $t_{0.025}(14) = 2.145$ , (60.19, 61.18).

**Задача 4.** Производитель наручных часов утверждает, что недельная ошибка в показаниях часов, которые он производит, имеет стандартное отклонение около 1 секунды. Чтобы проверить это утверждение, на 20 случайно отобранных часах было установлено правильное время. После истечения 1 недели ошибки были зафиксированы. Результаты, в секундах, выглядят следующим образом:

0.6	2.3	2.0	-2.1	-1.4
-0.5	1.5	-0.3	0.4	0.6
0.4	-2.2	0.7	0.5	-1.3
-2.0	2.6	-0.8	1.0	-0.6

**Найти:** (a) выборочное среднее, (b) выборочное стандартное отклонение, (c) 95% доверительный интервал для генерального среднего, (d) 90% доверительный интервал для генерального стандартного отклонения.

**Ответ:** (a)  $\bar{X} = 0.07$ , (b)  $s = 1.44$ , (c)  $(-0.60, 0.74)$ ,  
(d)

$$\left( \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}} \right) = \left( \sqrt{\frac{19 \cdot (1.44)^2}{\chi_{0.05}^2(19)}}, \sqrt{\frac{19 \cdot (1.44)^2}{\chi_{0.95}^2(19)}} \right) = (1.14, 1.97)$$

поскольку  $\chi_{0.05}^2(19) \approx 30.14$  и  $\chi_{0.95}^2(19) \approx 10.12$  (см. таблицу Table A.5).

**Задача 5.** Компания производит консервированные томаты в банках весом около 14 унций (oz). Веса десяти случайно отобранных банок показаны в следующей таблице.

13.85	13.95	13.90	13.49	14.17
14.33	14.03	13.48	14.27	14.19

**Найти:** (a) выборочное среднее; (b) выборочное стандартное отклонение; (c) 90% доверительный интервал для генерального среднего  $\mu$ , (d) 90% доверительный интервал для генерального стандартного отклонения  $\sigma$ .

**Ответ:** (a)  $\bar{X} \approx 14.0$  oz, (b)  $s = 0.298$  oz, (c) (13.8, 14.1) oz. (d) (0.218, 0.491) oz.

**Задача 6.** Данные в таблице показывают потребление протеина, в граммах, за 24-часовой период для независимых случайных выборок из 15 человек с доходами ниже прожиточного минимума и 10 человек с доходами выше прожиточного минимума:

Ниже прожиточного минимума	Выше прожиточного минимума
51.4 49.7 72.0	86.0 69.0
76.7 65.8 55.0	59.7 80.2
73.7 62.1 79.7	68.6 78.1
66.2 75.8 65.4	98.6 69.8
65.5 62.0 73.3	87.7 77.2



**Найти:** (a) выборочное среднее  $\mu_1$  и стандартное отклонение  $\sigma_1$  для первой выборки, (b) выборочное среднее  $\mu_2$  и стандартное отклонение  $\sigma_2$  для второй выборки, (c) pooled??? объединённое выборочное стандартное отклонение  $s_p$ , (d) 95% доверительный интервал для разности  $\mu_1 - \mu_2$  между средним потреблением протеина всеми людьми с доходами ниже прожиточного минимума и средним потреблением протеина всеми людьми с доходами выше прожиточного минимума.

**Решение:**

Ниже прожиточного минимума	Выше прожиточного минимума
$\bar{X}_1 = 66.29 \text{ g}$ $s_1 = 9.17 \text{ g}$ $n_1 = 15$	$\bar{X}_2 = 77.49 \text{ g}$ $s_2 = 11.34 \text{ g}$ $n_2 = 10$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 10.07 \text{ g}$$

Для 95% доверительного интервала  $t_{\alpha/2}(n_1 + n_2 - 2) = t_{0.025}(15 + 10 - 2) = 2.069$ .

Граничные точки этого доверительного интервала

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2) \cdot s_p \sqrt{(1/n_1) + (1/n_2)} = -11.20 \pm 8.51 = (-19.71, -2.69) \text{ g}.$$

Таким образом, мы можем быть на 95% уверенными, что средний человек с доходом выше прожиточного минимума получает в сутки как минимум на 2.69 г больше протеина, чем средний человек с доходом ниже прожиточного минимума.

## 22 Лекция 21. Проверка гипотез

### 22.1 Два типа гипотез

Если кто-то говорит вам, что некоторая генеральная совокупность является нормальной, вы можете поверить ему, но вы также можете пожелать проверить это утверждение, потому что выборка может и не быть нормальной. Утверждения "Генеральная совокупность нормальная" и "Генеральная совокупность ненормальная" взаимноисключающие. Поскольку не ясно, какое из них истинное, оба они называются *гипотезами*. Утверждения "Стандартное отклонение генеральной совокупности равно  $\sigma$ ", "Стандартное отклонение генеральной совокупности не равно  $\sigma$ ", "Стандартное отклонение генеральной совокупности меньше, чем  $\sigma$ ", "Стандартное отклонение генеральной совокупности больше, чем  $\sigma$ " также являются гипотезами.

Мы будем рассматривать проблему выбора между такими двумя гипотезами о генеральной совокупности на основе сделанной из неё случайной выборки. Эта статистическая задача называется *проверкой гипотез*. Одна из этих гипотез называется *нулевой гипотезой* и обозначается  $H_0$ , в то время как другая называется *альтернативной гипотезой* и обозначается  $H_1$  (или  $H_a$ ). В большинстве случаев  $H_0$  обозначает "статус кво", "текущий взгляд" на субъект, состояние "неизменения", тогда как альтернативная гипотеза определяет "новый взгляд", "новую теорию", "новое объяснение состояния вещей". Для того, чтобы доказать, что нулевая гипотеза истинна, надо доказать, что альтернативная гипотеза должна быть отвергнута.

Альтернативная гипотеза может быть разных видов.

Примеры:

$$H_0 : \mu = 68, H_1 : \mu < 68,$$

или

$$H_0 : \mu = 68, H_1' : \mu > 68,$$

или

$$H_0 : \mu = 68, H_1'' : \mu \neq 68.$$

Две первых альтернативных гипотезы – *односторонняя* (или *однохвостная*), а другая *двухсторонняя* (*двуххвостная*).

### 22.2 Два типа ошибок

Имея дело со статистическими явлениями и объектами, такими, как выборки, мы никогда не можем что-то определённо утверждать: каждое утверждение может оказаться истинным или ложным. Каждое утверждение с некоторой вероятностью может оказаться ложным, и разница между двумя утверждениями, одно из которых кажется истинным, а другое кажется ложным, заключается всего лишь в разнице вероятностей соответствующих ошибок.

**Определение 1.** Отклонение нуль-гипотезы, когда она верна называется *ошибкой I-го рода*.

**Определение 2.** Принятие нуль-гипотезы, когда она ложна называется *ошибкой II-го рода*.

Возможные ситуации представлены в таблице:

Решение	$H_0$ истинна	$H_0$ ложна
Принять $H_0$	Правильное решение	Ошибка II-го рода
Отклонить $H_0$	Ошибка I-го рода	правильное решение

Процесс принятия решений должен осуществляться с учётом вероятности ошибочного заключения.

**Определение 3.** Вероятность совершения ошибки I-го рода называется *уровнем значимости* и обозначается греческой буквой  $\alpha$ :

$$P(\text{type I error}) = \alpha.$$

Вероятность ошибки II-го рода обозначается  $\beta$ .

## 22.3 Процедура проверки гипотез (классическая, $\alpha$ -подход)

**Шаг 1.** Сформулировать нуль- и альтернативную гипотезы  $H_0$  и  $H_1$ .

**Шаг 2.** Выбрать уровень значимости  $\alpha$ .

**Шаг 3.** Выбрать статистику и найти её *критические значения* для заданной вероятности  $\alpha$ .

**Шаг 4.** Вычислить значение выборочной тестовой статистики ( $TS$ ).

**Шаг 5.** Если она попадает в область отвергания, тогда отвергаем  $H_0$ . Иначе, принимаем её.

**Шаг 6.** Сформулировать заключение словами.

## 22.4 Процедура проверки гипотез (подход на основе $P$ -значения)

**Шаг 1.** Сформулировать нуль- и альтернативную гипотезы  $H_0$  и  $H_1$ .

**Шаг 2.** Выбрать уровень значимости  $\alpha$ .

**Шаг 3.** Выбрать статистику.

**Шаг 4.** Вычислить выборочное значение тестовой статистики ( $TS_{\text{calc}}$ ) и найти соответствующее  $P$ -значение из таблицы *площади под кривой* для вычисленного  $TS_{\text{calc}}$ .

**Шаг 5.** Если  $P < \alpha$ , тогда  $H_0$  отвергается. Иначе, принимается.

**Шаг 6.** Сформулировать заключение словами.

**Замечание.** Чем меньше  $P$ -значение, тем сильнее свидетельство против нуль-гипотезы.

## 22.5 Пример

Допустим, мы знаем среднюю розничную цену учебников по теории вероятностей 1990 года, например, \$ 35.49, имеем случайную выборку цен текущего года  $\{42, 44, \dots, 28\}$  объёмом 41 и хотим узнать, выросла ли средняя цена этого года по сравнению со средней ценой \$35.44 в 1990 году. Сначала пойдём по классическому пути ( $\alpha$ -подход).

**Шаг 1.**

Формулируем нулевую и альтернативную гипотезы:

$H_0 : \mu = 35.44$  (средняя цена не выросла).

$H_1 : \mu > 35.44$  (средняя цена выросла).

**Шаг 2.** Определим уровень значимости. Пусть он соответствует 99%, то есть  $\alpha = 0.01$ .

**Шаг 3.** Выберем статистику  $TS$  и найдём её критическое значение:

$$TS = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = T \approx Z$$

(поскольку  $n > 30$ ). Здесь мы имеем дело с *односторонним тестом*, таким образом критическое значение удовлетворяет уравнению  $P(Z > z_\alpha) = \alpha$ . Из **Table A.4** для *критических значений* мы видим  $z_{0.01} = 2.326$  (более точное значение  $t_{0.01}(40) = 2.423$ , как можно видеть из той же таблицы). То есть область отвергания гипотезы для  $H_0$  – это  $(2.33 < Z < \infty)$ .

**Шаг 4.** Вычислим *выборочное значение* статистики:

$$TS = \hat{Z} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \approx 2.85.$$

**Шаг 5.** Заметим, что  $\hat{Z} = 2.85 \in (2.33, \infty)$ . Тогда гипотеза  $H_0$  должна быть отвергнута.

**Шаг 6.** При уровне значимости 0.01 можно утверждать, что средняя цена учебников по теории вероятностей выросла по сравнению со средней ценой 1990 года.

А сейчас рассмотрим ту же самую проблему в рамках подхода с использованием  $P$ -значения.

**Шаг 1.** Тот же самый.

**Шаг 2.** Тот же самый.

**Шаг 3.** Выбираем  $TS$  как и раньше, но не ищем её критическое значение.

**Шаг 4.** Вместо этого мы вычисляем выборочное значение для  $TS$  (оно, конечно, 2.85) и находим  $P$ -значение, то есть  $P = P(Z > 2.86)$ . Используя свойство вероятности и таблицу **Table A.3** для *площадей под Нормальной Кривой*, находим

$$P = P(Z > 2.86) = 1 - P(Z < 2.86) = 1 - 0.9979 = 0.0021.$$

**Шаг 5.** Сравним  $P$  с  $\alpha$ . Поскольку  $P = 0.0021 < \alpha = 0.01$ , заключаем, что гипотеза  $H_0$  должна быть отвергнута.

**Шаг 6.** При  $P$ -значении 0.0021, можно утверждать, что средняя цена учебников по теории вероятностей выросла по сравнению со средней ценой 1990 года.

## 23 Lecture 22. Критерий согласия ???Goodness-of-Fit Test

### 23.1 Столбчатые диаграммы

Вернёмся к вопросу, сформулированному в начале 21-ой лекции: как можно доказать, что мы имеем дело с генеральной совокупностью, подчиняющейся распределению  $f(x)$ ? Или, как можно проверить, что распределение именно то, что заявлено? Вся эта лекция посвящена рассмотрению этой проблемы.

Но начнём мы с простейшей задачи: как представить распределение, используя случайную выборку? Предположим, вы знаете, что генеральная совокупность дискретна со значениями  $x_1, x_2, \dots, x_k$  и неизвестными вероятностями  $f(x)$ . Возьмём случайную выборку  $\{X_1, X_2, \dots, X_n\}$  и составим порядковую статистику ???the order statistic. Она будет выглядеть следующим образом:

$$\underbrace{x_1, x_1, \dots, x_1}_{N_1} \underbrace{x_2, x_2, \dots, x_2}_{N_2} \dots \underbrace{x_k, x_k, \dots, x_k}_{N_k}$$

Данные распадаются на  $k$  групп, часто называемых *классами*. Первый класс содержит  $N_1$  элементов, второй содержит  $N_2$  и так далее, так что  $N_1 + N_2 + \dots + N_k = n$ . Конечно, если вы будете повторять составлять выборку, вы получите другие числа (вот почему мы обозначаем их заглавными буквами, как случайные величины), но если они достаточно

велики, их вариация от выборки к выборке будет достаточно мала. Эти числа называются *частотами*, а отношения  $N_j/n$ , называемые *относительными частотами*, могут быть использованы для аппроксимации вероятностей:

$$N_1/n \approx P(X = x_1) = f(x_1), \quad N_2/n \approx P(X = x_2) = f(x_2), \quad \dots \quad N_k/n \approx P(X = x_k) = f(x_k).$$

Графически они представляются *вертикальными столбиками*.

Рассмотрим следующий пример. Имеем случайную выборку

7 1 3 2 2 9 4 3 5 1  
4 2 8 5 5 2 6 2 3 4  
3 4 2 4 6 7 1 5 6 4

Упорядоченная выборка выглядит следующим образом:

1 1 1 2 2 2 2 2 2 3  
3 3 3 4 4 4 4 4 4 5  
5 5 5 6 6 6 7 7 8 9

Составим таблицу:

$x_k$	Частота $N_k$	Относительная частота $N_k/n$
1	3	0.100
2	6	0.200
3	4	0.133
4	6	0.200
5	4	0.133
6	3	0.100
7	2	0.067
8	1	0.033
9	1	0.033
Нормировка	$\sum(\dots) = n = 30$	$\sum(\dots) = 0.999 \approx 1$

Малое отклонение последней суммы от 1 возникает из-за округления относительных частот и может быть устранено при более точных вычислениях.

## 23.2 Гистограммы

Имея дело с непрерывными распределениями, можно поделить упорядоченную выборку на подходящее число классов, обычно на серию интервалов одинаковой длины, а затем работать с ними как было показано ранее. Например, имеется упорядоченная выборка объёма 30 :

57.3 59.1 61.9 66.5 67.2 67.3 67.9 68.0 69.1 71.9  
73.1 73.1 76.6 78.3 78.4 78.8 78.9 80.0 81.5 81.8  
82.7 84.1 87.6 89.4 89.9 90.0 94.1 94.4 95.8 100.0

После того, как данные отсортированы, таблица частот может быть построена так же, как

и в дискретном случае:

Интервал	Частота $N_k$	Относительная частота $N_k/n$	Плотность относительной частоты $N_k/(n\Delta x)$
$50 < x \leq 60$	2	0.067	0.0067
$60 < x \leq 70$	7	0.233	0.0233
$70 < x \leq 80$	9	0.300	0.0300
$80 < x \leq 90$	8	0.267	0.0267
$90 < x \leq 100$	4	0.133	0.0133
Нормировка	$\sum(\dots) = n = 30$	$\sum(\dots) = 1$	$\sum(\dots)\Delta x = 1$

Обратите внимание на *соглашение о граничной точке* (каждая граничная точка принадлежит только к одному интервалу) и разницу в нормировочных условиях. Графически эти данные представляются гистограммой, составленной из соприкасающихся прямоугольников: высота каждого прямоугольника представляет частоту, а основания прямоугольников располагаются между последовательными границами интервалов. Заметим, что подобно столбчатой диаграмме, измеряемые значения располагаются на горизонтальной оси, а частота появления - на вертикальной оси. Однако, между вертикальными столбиками нет пропусков, в отличие от столбчатой диаграммы. Этот факт объясняется непрерывностью измеряемых величин.

Основное решение, с необходимостью принятия которого вы столкнётесь при построении гистограммы – это выбор числа и величины классовых интервалов. Если интервалов слишком много, тогда число точек данных  $n_j$  может быть мало и статистические ошибки могут быть большими. Если интервалов у вас недостаточно, тогда гистограмма будет недостаточно детальна. Заметим, что ширина интервалов не обязательно должна быть одна и та же. В области с высокой плотностью данных можно выбирать более узкие интервалы, а там, где плотность данных мала и изменяется медленно, можно выбирать интервалы более длинными.

### 23.3 Хи-квадрат распределение в выборках с известной $f(x)$

Как следует из вышесказанного, нормированные частоты могут служить оценками для распределения вероятностей (в дискретном случае) и плотности вероятности (в непрерывном случае). Интерпретируя частоты как целочисленные случайные величины  $N_1, N_2, \dots, N_k$ , необходимо знать их совместное распределение.

Рассмотрим очень простой случай, когда наблюдения или измерения попадают в один из двух классов так,  $k = 2$ . Пусть  $p_1$  представляет теоретическую вероятность того, что наблюдение попадёт в класс 1, а  $p_2$  – вероятность того, что наблюдение попадает в класс 2,  $p_1 + p_2 = 1$ . Тогда через  $n$  обозначим объём выборки (полное число наблюдений в выборке), и через  $N_1$  обозначим число наблюдений, попадающих в класс 1. Очевидно,  $N_1$  – биномиальная случайная величина, распределённая с параметрами  $n$  и  $p_1$ . Следует вспомнить три свойства этой случайной величины:

$$\mu_{N_1} = np_1,$$

$$\sigma_{N_1} = \sqrt{np_1(1 - p_1)},$$

и

$$\frac{N_1 - \mu_{N_1}}{\sigma_{N_1}} = \frac{N_1 - np_1}{\sqrt{np_1(1 - p_1)}} \stackrel{d}{=} Z, \quad (1)$$

где  $p$  достаточно велико.

Далее рассмотрим квадрат случайной величины

$$\left( \frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2 = \frac{(N_1 - np_1)^2}{np_1(1-p_1)}.$$

Принимая во внимание соотношения  $p_1 + p_2 = 1$  и  $N_1 + N_2 = n$ , получаем

$$N_1 - np_1 = n - N_2 - n(1-p_2) = -(N_2 - np_2), \quad \frac{1}{np_1(1-p_1)} = \frac{1}{np_1p_2} = \frac{1}{np_1} + \frac{1}{np_2},$$

и

$$\begin{aligned} \frac{(N_1 - np_1)^2}{np_1(1-p_1)} &= (N_1 - np_1)^2 \left( \frac{1}{np_1} + \frac{1}{np_2} \right) = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} = \\ &= \sum_{j=1}^2 \frac{(N_j - np_j)^2}{np_j}. \end{aligned} \quad (2)$$

Как следует из (1) и (2)

$$\sum_{j=1}^2 \frac{(N_j - np_j)^2}{np_j} \stackrel{d}{=} Z^2.$$

В свою очередь,  $Z^2$  распределена с плотностью хи-квадрат с 1 степенью свободы,

$$Z^2 \stackrel{d}{=} \chi^2(1),$$

таким образом

$$\sum_{j=1}^2 \frac{(N_j - np_j)^2}{np_j} \stackrel{d}{=} \chi^2(1). \quad (3)$$

В случае произвольного числа классов  $k$  случайные величины  $N_1, N_2, \dots, N_k$  имеют мультиномиальное multinomial распределение и обобщение формулы (3) выглядит следующим образом:

$$\sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \stackrel{d}{=} \chi^2(k-1). \quad (4)$$

Если вероятности  $p_j$  были вычислены с использованием  $m$  параметров, полученных из той же выборки, уравнение примет вид

$$\sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \stackrel{d}{=} \chi^2(k-m-1). \quad (5)$$

## 23.4 Критерий согласия $\chi^2$

Вспомним, что уравнение (4) имеет место, когда вероятности  $p_j$  вычисляются с использованием распределения генеральной совокупности  $f_0(x)$ , из которой получена случайная выборка: для каждого интервала  $\Delta x_j$

$$p_j \equiv P(X \in \Delta x_j) = \begin{cases} \sum_{x \in \Delta x_j} f_0(x), & \text{в дискретном случае,} \\ \int_{\Delta x_j} f_0(x) dx, & \text{в непрерывном случае.} \end{cases}$$

Если генеральная совокупность имеет какое-то другое распределение  $f(x) \neq f_0(x)$

$$\sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \not\stackrel{d}{=} \chi^2(k-1).$$

Следовательно, левую часть уравнения (4) можно использовать в качестве тестовой статистики для проверки гипотезы о распределении генеральной совокупности. Обозначим её через  $X^2$ :

$$X^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}. \quad (6)$$

Такая проверка называется *критерием согласия*  $\chi^2$ .

**Замечание.** Некоторые авторы используют для обозначения  $TS$  (6) символ  $\chi^2$ . Это не совсем корректно, поскольку  $X^2 \stackrel{d}{=} \chi^2$  только если  $f(x) = f_0(x)$ .

Пусть случайная выборка  $\{X_1, \dots, X_n\}$  получена из некоторой генеральной совокупности с неизвестным распределением  $f(x)$ . Мы хотим проверить, будет ли  $f(x) = f_0(x)$ , где  $f_0(x)$  – некоторое определённое распределение, (называемое *ожидаемым, теоретическим* или *гипотетическим* распределением). Другими словами, мы формулируем нулевую и альтернативную гипотезы:

**Шаг 1.**  $H_0 : f(x) = f_0(x)$  vs  $H_1 : f(x) \neq f_0(x)$ .

Затем выполняем два следующих шага.

**Шаг 2.** Выбираем уровень значимости  $\alpha$ .

**Шаг 3.** Выбираем статистику и определяем её критическое значение. Мы выбираем  $X^2$  в качестве  $TS$ . Затем мы выбираем число интервалов (классов)  $k$ , чтобы найти критическое значение  $\chi_\alpha^2(k-1)$  из таблицы from Table A.5 содержащей решения уравнения  $P(\chi^2 > \chi_\alpha^2(k-1)) = \alpha$ .

**Шаг 4.** Вычисляем  $TS$ . Для того, чтобы сделать это мы должны разбить возможные значения  $X$  на  $k$  взаимно непересекающихся интервалов (как при построении частотной таблицы или гистограммы). Затем вычисляем соответствующие вероятности  $p_j$ ,  $j = 1, 2, \dots, k$  на основе *гипотетического распределения*  $f_0(x)$  (называемые *ожидаемыми, теоретическими* или *гипотетическими* вероятностями). После этого находим число наблюдений  $N_j$ ,  $j = 1, 2, \dots, k$  в выборке из *исследуемой генеральной совокупности*, попадающих в каждый интервал. Используя эти данные, вычисляем тестовую статистику

$$X^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j},$$

где  $np_j$  называются *ожидаемыми, теоретическими* или *гипотетическими частотами*.

**Замечание.** Мы не должны обозначать тестовую статистику через  $\chi^2$ , поскольку реальное распределение  $f(x)$  ещё не известно.

Процедура вычисления этой тестовой статистики представлена в следующей таблице.

Интервалы	Наблюдаемые частоты $N_j$	Гипотетические вероятности $p_j$	Гипотетические частоты $np_j$	$\frac{(N_j - np_j)^2}{np_j}$
1	$N_1$	$p_1$	$np_1$	$\frac{(N_1 - np_1)^2}{np_1}$
2	$N_2$	$p_2$	$np_2$	$\frac{(N_2 - np_2)^2}{np_2}$
...	...	...	...	...
$k$	$N_k$	$p_k$	$np_k$	$\frac{(N_k - np_k)^2}{np_k}$
Total	$n$	1.00	$n$	$X^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}$

**Шаг 5.** Сравнение  $X^2$  с  $\chi_\alpha^2(k-1)$ . Если  $X^2 > \chi_\alpha^2(k-1)$ , тогда нулевая гипотеза отвергается, иначе она принимается.



**Шаг 6.** Сформулируем заключение словами.

**Замечание.** Это классический  $\alpha$ -подход. Чтобы использовать подход на основе  $P$ -значения, мы должны найти  $P = P(\chi^2(k-1) > X^2)$  и сравнить его с  $\alpha$ . Нулевая гипотеза отклоняется, когда  $P > \alpha$  и принимается в противоположном случае.

## 24 Lecture 23. Типичные задачи

### 24.1 Пример 1. Односторонняя проверка гипотез One-sided HT

Диетолог полагает, что средний человек с доходом ниже прожиточного минимума потребляет кальция меньше рекомендуемой ежедневной величины (РЕВ) в 800 мг.

Чтобы проверить свою догадку, он рассмотрел ежедневные дозы потребления кальция для случайной выборки из 35 человек с доходами ниже прожиточного минимума:

879	1096	701	986	828	1077	703
555	422	997	473	702	508	530
513	720	944	673	574	707	864
1199	743	1325	655	1043	599	1008
705	180	287	542	893	1052	473

Будут ли представленные данные при уровне значимости в 5% являться достаточным свидетельством в пользу указанного предположения?

**Решение:**

**Шаг 1.** Сформулируем гипотезы:  $H_0 : \mu = \mu_0 \equiv 800mg$  против  $H_1 : \mu < \mu_0$ . Заметим, что проверка гипотезы - левосторонняя, поскольку знак "меньше" ( $<$ ) появляется в левосторонней гипотезе.

**Шаг 2.** Установим уровень значимости в 5%:  $\alpha = 0.05$

**Шаг 3.** Выберем тестовую статистику (TS)  $\frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ . Поскольку  $n = 35 > 30$ , она распределена как  $Z$ . Соответствующее критическое значение  $-z_{0.05} = -1.645$ .

**Шаг 4.** Вычисляем TS:

$$\bar{X} = \frac{879 + 1096 + \dots + 473}{35} = 747,$$

$$s^2 = \frac{1}{34} [879^2 + 1096^2 + \dots + 473^2 - 35 \cdot (747)^2] = 68749, \quad s = 262,$$

$$Z = \frac{747 - 800}{262/\sqrt{35}} = -1.19.$$

**Объяснение следующего шага.** Мы не знаем, получена ли выборка  $\{879, 1096, \dots, 473\}$  из генеральной совокупности с  $\mu = \mu_0$  или из генеральной совокупности с  $\mu < \mu_0$ . Однако, мы знаем, что в первом случае (когда нулевая гипотеза верна) событие  $\{Z < -1.645\}$  "почти невозможно" (его вероятность равна всего лишь 0.05), в то время как во втором случае (когда нулевая гипотеза ложна) оно должно быть более вероятно (соответствующее среднее значение сдвинуто в сторону значения  $z = -1.645$ ). То есть, если мы наблюдаем "почти невозможное" (для гипотезы  $H_0$ ) событие  $\{Z < -1.645\}$ , мы должны отвергнуть нуль-гипотезу. Уровень значимости  $\alpha$  служит для разделения всех в принципе, возможных событий на две группы: практически возможные и практически невозможные события.

Мы отвергаем нулевую гипотезу, если наблюдаем *практически невозможное* (в рамках этой гипотезы) событие.

**Шаг 5.** Поскольку  $Z = -1.19 > -1.645$ , мы не можем отвергать нулевую гипотезу.

**Шаг 6.** Заключение, сформулированное словами: при уровне значимости 5%, выборка из 35 потреблённых доз кальция не обеспечивает достаточного свидетельства в пользу заключения о том, что среднее суточное потребление кальция  $\mu$  всеми людьми с доходами меньше прожиточного минимума является меньше РЕВ в 800 мг.

## 24.2 Пример 2. Двусторонний тест

Основной акционер общества с ограниченной ответственностью говорит потенциальному инвестору, что средняя месячная арендная плата за трёхкомнатную квартиру в городе составляет \$ 587. Чтобы проверить это заявление инвестор выбирает 32 трёхкомнатные квартиры в городе и узнаёт их ежемесячные арендные платы:

289	560	726	643	586	657	565	676
656	577	663	729	745	597	669	626
450	669	603	545	661	610	604	598
507	675	609	503	589	521	595	472

Предполагают ли эти данные, что заявление основного акционера корректно? Проведём соответствующую проверку гипотез.

**Шаг 1.** Формулируем гипотезы:  $H_0 : \mu = \mu_0 \equiv \$587$  и  $H_1 : \mu \neq \mu_0$ . Отметим, что проверка гипотез двусторонняя.

**Шаг 2.** Определим уровень значимости:  $\alpha = 0.05$ .

**Шаг 3.** Выберем TS. Поскольку  $n = 32 < 30$ ,  $\frac{\bar{X} - \mu_0}{s/\sqrt{n}} = Z$ . Критические значения для двусторонней проверки  $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.96$ .

**Шаг 4.** Вычислим значение TS:

$$\bar{X} = \$599.22, s = \$91.2, Z = \frac{599.22 - 598}{91.2/\sqrt{32}} = 0.76.$$

**Шаг 5.** Результат  $Z = 0.76 \in (-1.96, 1.96)$ , то есть он не попадает в область отвергания  $(-\infty, -1.96) \cup (1.96, \infty)$ . Следовательно, мы не отвергаем  $H_0$ .

**Шаг 6.** При уровне значимости 5%, данные не обеспечивают достаточной уверенности в том, что средняя месячная арендная плата  $\mu$ , отличается от заявленной основным акционером в размере \$587.

В случае, если вас просят использовать **подход на основе  $P$ -значения**: необходимо взять наблюдаемое значение  $Z_{obs} = 0.76$  в качестве граничных точек интервала  $(-0.76, 0.76)$  и найти вероятность за пределами интервала, используя таблицу Table A.3. Это и будет  $P$ -значение:

$$P = P(Z < -0.76) + P(Z > 0.76) = 2P(Z < -0.76) = 2 \cdot 0.224 = 0.448.$$

Это значение больше, чем  $\alpha = 0.05$ , следовательно нулевая гипотеза не может быть отвергнута.

## 24.3 Пример 3. Построение гистограмм

Рассмотрим данные из следующей таблицы, которые показывают время жизни 40 одинаковых автомобильных аккумуляторов, с точностью до десятой доли года.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Процедура построения частотной ( $F = N_j$ ) гистограммы, гистограммы относительных частот ( $R = N_j/n$ ) и гистограммы плотности относительных частот ( $D = N_j/(n\Delta x)$ ) со среднеквадратичными ошибками ( $MSE = \epsilon$ ,  $\epsilon_F = \sqrt{F}$ ,  $\epsilon_R = \epsilon_F/n$ ,  $\epsilon_D = \epsilon_F/(n\Delta x)$ ) представлена в следующей таблице.

Интервалы	Средняя точка	Метки	$F$	$\epsilon_F$	$F \pm \epsilon_F$	$R \pm \epsilon_R$	$D \pm \epsilon_D$
1.5-1.9	1.7	**	2	1.4	$2 \pm 1.4$	$.050 \pm .035$	$.125 \pm .088$
2.0-2.4	2.2	*	1	1	$1 \pm 1$	$.025 \pm .025$	$.062 \pm .062$
2.5-2.9	2.7	****	4	2	$4 \pm 2$	$.100 \pm .050$	$.250 \pm .125$
3.0-3.4	3.2	*****	15	3.9	$15 \pm 3.9$	$.375 \pm .098$	$.938 \pm .244$
3.5-3.9	3.7	*****	10	3.2	$10 \pm 3.2$	$.250 \pm .080$	$.625 \pm .200$
4.0-4.4	4.2	*****	5	2.2	$5 \pm 2.2$	$.125 \pm .055$	$.313 \pm .138$
4.5-4.9	4.7	***	3	1.7	$3 \pm 1.7$	$.075 \pm .042$	$.188 \pm .106$

**Объяснение СКО.** Среднеквадратичная ошибка определяется выражением  $\epsilon = \sqrt{E(N_j - \mu_{N_j})^2} = \sigma_{N_j}$ . Каждая частота  $N_j$  распределена согласно биномиальному распределению со средним значением  $np_j$  и дисперсией  $\sigma^2 = np_j(1 - p_j)$ . Обычно,  $p_j \ll 1$ , тогда  $\sigma^2 = \sqrt{np_j(1 - p_j)} \approx \sqrt{np_j}$ . Но  $np_j = EN_j$ , и мы можем рассматривать  $N_j$  как несмещённую оценку  $np_j$ , и таким образом  $\epsilon = \sigma \approx \sqrt{N_j}$ .

## 24.4 Пример 4. Критерий согласия

Вы хотите проверить, является ли данная игральная кость правильной (честной) или неправильной (нечестной).

**Шаг 1.** Нулевая гипотеза: игральная кость правильная, альтернативная гипотеза: игральная кость неправильная. На языке распределений

$$H_0 : f(x) = \frac{1}{6}, \quad H_1 \neq \frac{1}{6} \quad x = 1, 2, \dots, 6.$$

**Шаг 2.** Уровень значимости  $\alpha = 0.05$ .

**Шаг 3.** Проверочная статистика  $X$ , число классов  $k = 6$ , и критическое значение  $\chi_{0.05}(6 - 1) = 11.07$ .

**Шаг 4.** Берём случайную выборку объёмом, скажем,  $n = 60$ , пересортируем её в порядке возрастания, найдём наблюдаемые частоты, вычислим вероятности  $p_j$  согласно нулевой гипотезе ( $p_j = 1/6$ ), и посчитаем  $X^2$ . Эта процедура представлена в следующей таблице.

Классы	Наблюдаемые частоты $N_j$	Гипотетические вероятности $p_j$	Гипотетические частоты $np_j$	$\frac{(N_j - np_j)^2}{np_j}$
1	11	1/6	10	$\frac{(11-10)^2}{10} = 0.1$
2	6	1/6	10	$\frac{(6-10)^2}{10} = 1.6$
3	9	1/6	10	$\frac{(9-10)^2}{10} = 0.1$
4	14	1/6	10	$\frac{(14-10)^2}{10} = 1.6$
5	7	1/6	10	$\frac{(7-10)^2}{10} = 0.9$
6	13	1/6	10	$\frac{(13-10)^2}{10} = 0.9$
Всего	60	1.00	60	$X^2 = 5.2$

**ОБЪЯСНЕНИЕ СЛЕДУЮЩЕГО ШАГА.** Если наблюдаемые частоты  $N_j$  совпадают с соответствующими ожидаемыми частотами  $np_j$ , тогда  $X^2$  равняется 0. Однако, это событие очень маловероятно. Благодаря естественному рассеиванию наблюдаемых частот, называемому *статистическими флуктуациями*,  $N_j \neq np_j$  и  $X^2$  положительны (как сумма положительных членов  $(N_j - np_j)^2 / np_j$ ). Критическое значение  $\chi^2_\alpha$  показывает верхнюю границу "практически возможных" значений величины  $X^2$  в случае  $EN_j = np_j$  (то есть при нулевой гипотезе). Событие  $\{X^2 > \chi^2_\alpha\}$  "практически невозможно" для  $H_0$ , так что если оно наблюдается, то  $H_0$  должна быть отвергнута.

**Шаг 5.** Сравнивая  $X^2 = 5.2$  с критическим значением  $\chi^2_{0.05}(5) = 11.07$ , видим, что  $X^2 < \chi^2_{0.05}(5)$ . Таким образом, нулевая гипотеза не может быть отвергнута.

**Шаг 6.** С вероятностью 95%, заключение "игральная кость - правильная" справедливо.

## STAT 312 Spring

### Домашняя работа 6

(must be returned on Apr 26, 12:30 p.m., CLPP 108)

**Задача 1.** Несколько десятилетий назад средняя продолжительность жизни в США составляла 70 лет и  $\sigma = 10$ . Полученная недавно случайная выборка объёма  $n = 100$  показывает 72 года. Означает ли это, что средняя продолжительность жизни действительно увеличилась? Используйте односторонний тест с 5% уровнем значимости.

**Задача 2** Имеется выборка  $X_1, \dots, X_{32}$  с  $\bar{X} = 599$  и  $S = 91.2$ . Ранее все считали, что  $\mu = 587$ . Опровергает ли этот результат общепринятое мнение? Используйте двухсторонний тест с 5%-м уровнем значимости.

**Задача 3.** Электрическая фирма, производящая лампы утверждает, что среднее время их безотказной работы  $\mu$  составляет 800 часов и стандартное отклонение  $\sigma = 40$  часов. Проверить гипотезу  $\mu = 800$ ч против  $\mu \neq 800$ ч, если  $n = 30$   $\bar{X} = 788$ . Использовать двухсторонний тест и подход на основе  $P$ -значения.

**Задача 4.** Статья в журнале Технометрика ???Technometrics (Vol. 19, 1977, p.425) представляет следующие данные по октановым числам моторного топлива нескольких марок бензина:

88.5	94.7	84.3	90.1	89.0	89.8	91.6	90.3	90.0
91.5	89.9	98.8	88.3	90.4	91.2	90.6	92.2	87.7
91.1	86.7	93.4	96.1	89.6	90.4	91.6	90.7	88.6
88.3	94.2	85.3	90.1	89.3	91.1	92.2	83.4	91.0
88.2	88.5	93.3	87.4	91.1	90.5	87.6	92.7	87.9
93.0	94.4	90.4	91.2	86.7	94.2	90.8	90.1	91.8
88.4	92.6	93.7	96.5	84.3	93.2	88.6	88.7	92.7
89.3	91.0	87.5	87.8	88.3	89.2	92.3	88.9	89.8
92.7	93.3	86.7	91.0	90.9	89.9	91.8	89.7	92.2

Используя 8 одинаковых интервалов между 80 и 100 построить: (а) таблицы частот, относительных частот и плотностей относительных частот; (b) гистограммы частот, относительных частот и плотностей относительных частот; (с) вычислить СКО для плотности относительных частот, построить таблицу и показать их на соответствующей гистограмме.

**Задача 5.** Критерий согласия часто используется, чтобы установить, является ли последовательность цифр случайной. В среднем, случайная последовательность должна содержать в равных пропорциях все цифры от 0 до 9.

(а) Построить частотную столбчатую диаграмму.

(b) Построить на одном рисунке столбчатые диаграммы относительных частот и равномерной вероятности.

(с) Относительно представленных ниже случайных чисел проверить, может ли быть принята гипотеза о том, что генеральная совокупность подчиняется равномерному распределению вероятностей  $p_j = 0.1$ ,  $j = 0, \dots, 9$  при уровне значимости  $\alpha = 0.1$ .

1	6	0	4	8	8	1	8	9	9
0	4	1	5	3	5	3	3	8	1
7	9	4	0	1	2	1	4	3	8
8	3	0	3	5	9	2	3	5	0

**Задача 6.** Знаменитый исторический пример пуассоновского распределения – число несчастных случаев в Прусской армии, вызванных пинком лошади в период 1875-1884. Эти данные приведены ниже:

Число смертей	0	1	2	3	4	$\geq 5$
Частота	109	65	22	3	1	0

- Построить столбчатую диаграмму частот.
- Построить столбчатые диаграммы относительных частот и пуассоновских вероятностей на одном рисунке.
- Использовать критерий согласия, чтобы оценить заявление, что эти данные подчиняются распределению Пуассона с параметром  $\mu = 0.6$ .

## Домашняя работа 6. Ответы и некоторые решения

**Задача 1.** Несколько десятилетий назад средняя продолжительность жизни в США составляла 70 лет и  $\sigma = 10$ . Полученная недавно случайная выборка объёма  $n = 100$  показывает 72 года. Означает ли это, что средняя продолжительность жизни действительно увеличилась? Используйте односторонний тест с 5% уровнем значимости.

**Решение.**

**Шаг 1.**  $H_0 : \mu = \mu_0 = 70$ ;  $H_1 : \mu > 70$  (односторонняя гипотеза).

**Шаг 2.**  $\alpha = 0.05$ .

**Шаг 3.** Поскольку  $n = 100 > 30$  мы используем нормальное распределение для  $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ . Из таблицы Table A.3 (page 670)  $z_\alpha = z_{0.05} = 1.645$ .

**Шаг 4.** Значение тестовой статистики:  $Z = \frac{72-70}{10/\sqrt{100}} = 2$ .

**Шаг 5.** Решение:  $Z = 2 > 1.645$ , оно попадает в область отвергания и мы отклоняем  $H_0$ . При уровне значимости 0.05 мы видим, что средняя продолжительность жизни увеличилась.

**Задача 2** Имеется выборка  $X_1, \dots, X_{32}$  с  $\bar{X} = 599$  и  $S = 91.2$ . Ранее все считали, что  $\mu = 587$ . Опровергает ли этот результат общепринятое мнение? Используйте двухсторонний тест с 5%-м уровнем значимости.

**Решение.**

**Шаг 1.**  $H_0 : \mu = \mu_0 = 587$ ;  $H_1 : \mu \neq 587$  (двусторонняя гипотеза).

**Шаг 2.**  $\alpha = 0.05$ .

**Шаг 3.** Из нормального распределения  $z_{\alpha/2} = z_{0.025} = 1.96$ .

**Шаг 4.** Значение тестовой статистики:  $Z = \frac{599-587}{91.2/\sqrt{32}} = 0.76$ .

**Шаг 5.** Решение:  $Z = 0.76 < 1.96$ , оно не попадает в область отвергания, следовательно мы не отклоняем общее мнение  $H_0$ .

**Задача 3.** Электрическая фирма, производящая лампы утверждает, что среднее время их безотказной работы  $\mu$  составляет 800 часов и стандартное отклонение  $\sigma = 40$  часов. Проверить гипотезу  $\mu = 800$ ч против  $\mu \neq 800$ ч, если  $n = 30$   $\bar{X} = 788$ . Использовать двухсторонний тест и подход на основе  $P$ -значения.

**Решение.** Используем способ из пункта 20.4.

**Шаг 1.**  $H_0 : \mu = \mu_0 = 800$ ;  $H_1 : \mu \neq 800$  (двухсторонняя гипотеза).

**Шаг 2.** Тестовая статистика  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{788-800}{40/\sqrt{30}} = -1.65$ .

**Шаг 3.**  $P$ -значение (из нормального распределения)

$$P = 2P(Z < -1.65) = 2 \cdot 0.0495 \approx 0.10.$$

**Шаг 4.** Решение:  $\mu = 800$  с  $P$ -значением 10%.

**Задача 4.** Статья в журнале Технометрика ???Technometrics (Vol. 19, 1977, p.425) представляет следующие данные по октановым числам моторного топлива нескольких марок бензина:

88.5	94.7	84.3	90.1	89.0	89.8	91.6	90.3	90.0
91.5	89.9	98.8	88.3	90.4	91.2	90.6	92.2	87.7
91.1	86.7	93.4	96.1	89.6	90.4	91.6	90.7	88.6
88.3	94.2	85.3	90.1	89.3	91.1	92.2	83.4	91.0
88.2	88.5	93.3	87.4	91.1	90.5	87.6	92.7	87.9
93.0	94.4	90.4	91.2	86.7	94.2	90.8	90.1	91.8
88.4	92.6	93.7	96.5	84.3	93.2	88.6	88.7	92.7
89.3	91.0	87.5	87.8	88.3	89.2	92.3	88.9	89.8
92.7	93.3	86.7	91.0	90.9	89.9	91.8	89.7	92.2

Используя 8 одинаковых интервалов между 80 и 100 построить: (а) таблицы частот, относительных частот и плотностей относительных частот; (b) гистограммы частот, относительных частот и плотностей относительных частот; (с) вычислить СКО для плотности относительных частот, построить таблицу и показать их на соответствующей гистограмме.

**Answer:**

**Задача 5.** Критерий согласия часто используется, чтобы установить, является ли последовательность цифр случайной. В среднем, случайная последовательность должна содержать в равных пропорциях все цифры от 0 до 9.

(а) Построить частотную столбчатую диаграмму.

(b) Построить на одном рисунке столбчатые диаграммы относительных частот и равномерной вероятности.

(с) Относительно представленных ниже случайных чисел проверить, может ли быть принята гипотеза о том, что генеральная совокупность подчиняется равномерному распределению вероятностей  $p_j = 0.1$ ,  $j = 0, \dots, 9$  при уровне значимости  $\alpha = 0.1$ .

1	6	0	4	8	8	1	8	9	9
0	4	1	5	3	5	3	3	8	1
7	9	4	0	1	2	1	4	3	8
8	3	0	3	5	9	2	3	5	0

**Задача 6.** Знаменитый исторический пример пуассоновского распределения – число несчастных случаев в Прусской армии, вызванных пинком лошади в период 1875-1884. Эти данные приведены ниже:

Число смертей	0	1	2	3	4	$\geq 5$
Частота	109	65	22	3	1	0

(а) Построить столбчатую диаграмму частот.

(b) Построить столбчатые диаграммы относительных частот и пуассоновских вероятностей на одном рисунке.

(с) Использовать критерий согласия, чтобы оценить заявление, что эти данные подчиняются распределению Пуассона с параметром  $\mu = 0.6$ .



## 25 Lecture 24. Корреляционный анализ

### 25.1 Двухмерные распределения и корреляции

Часто возникает заинтересованность в исследовании возможной связи между некоторыми переменными, представленными случайными величинами, скажем,  $X$  и  $Y$ . Например, мы хотим знать, есть ли какая-нибудь зависимость между температурой печи и твёрдостью керамики. Существует ли взаимосвязь между продажной и стартовой ценой ценной бумаги? ??? Is there a relationship between the stock's closing value and the opening price? Показывают ли женатые студенты на экзамене результаты лучше, чем неженатые?

В теории, способ ответа на эти вопросы основан на совместных распределениях  $f(x, y)$ . Графически, двумерная совместная плотность представляется как поверхность над координатной плоскостью  $xOy$ , которая описывается уравнением  $z = f(x, y)$ .

Маргинальные плотности

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

могут рассматриваться как некоторый вид проекций этой фигуры на плоскости  $yOz$  и  $xOz$  соответственно.

Условные плотности

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} \Big|_{y=\text{const}}, \quad f(y|x) = \frac{f(x, y)}{f_X(x)} \Big|_{x=\text{const}}$$

представляются, как сечение поверхности  $z = f(x, y)$  вертикальными плоскостями  $y = \text{const}$  и  $x = \text{const}$  соответственно, и удовлетворяют условию нормировки 1.

Если случайные величины  $X$  и  $Y$  независимы, то есть

$$f(y|x) = f_Y(y) \text{ for any } x, \text{ or, equally, } f(x|y) = f_X(x) \text{ for any } y,$$

то

$$f(x, y) = f_X(x)f_Y(y).$$

Существует более удобный критерий статистической зависимости: *линейный коэффициент корреляции*  $\rho_{X,Y}$ . Он определяется через дисперсии  $\sigma_X^2 = E(X - \mu_X)^2$ ,  $\sigma_Y^2 = E(Y - \mu_Y)^2$  и ковариацию  $\sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)]$  следующим образом:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

Если  $\rho_{X,Y} = 0$ , то случайные величины называются *линейно некоррелированными*. При  $\rho_{X,Y}$  близком к  $+1$  ( $-1$ ), говорят, что существует сильная положительная (отрицательная) линейная корреляция между случайными величинами  $X$  и  $Y$ .

Заметим, что для любого совместного распределения, если  $X$  и  $Y$  независимы, тогда они некоррелированы:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{E(X - \mu_X)E(Y - \mu_Y)}{\sigma_X \sigma_Y} = 0,$$

поскольку  $E(X - \mu_X) = EX - \mu_X = \mu_X - \mu_X = 0$ . Обратное утверждение: "если  $X$  и  $Y$  некоррелированы, тогда они независимы" является верным *только при условии, что совместное распределение  $X$  и  $Y$  нормально*.

## 25.2 Двухмерное нормальное распределение

Будем рассматривать две центрированные нормально распределённые случайные величины как координаты случайных точек  $(X, Y)$  с дисперсиями  $\sigma_X^2$  и  $\sigma_Y^2$  соответственно. Если эти координаты независимы,

$$\begin{aligned} f(x, y) &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{x^2}{2\sigma_X^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{y^2}{2\sigma_Y^2}\right) = \\ &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2}\right)\right]. \end{aligned}$$

Легко видеть, что кривая постоянных величин  $f(x, y)$  представляет собой эллипс с главными осями, лежащими на координатных осях  $x$  и  $y$ :

$$\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} = \text{const.} \quad (1)$$

Повернём систему координат вокруг её центра на произвольный угол и обозначим новые координаты через  $x_1, x_2$  (мы можем также представить, что мы вращаем плоскость с распределениями относительно координатных осей, такое преобразование также ведёт к изменению координат каждой точки  $(x, y) \mapsto (x_1, x_2)$ ). Уравнение (1) примет вид

$$\left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1}{\sigma_1}\right)\left(\frac{x_2}{\sigma_2}\right) + \left(\frac{x_2}{\sigma_2}\right)^2 = \text{const},$$

где  $\sigma_1 > 0$ ,  $\sigma_2 > 0$  и  $\rho \in [-1, 1]$  – новые постоянные, определяемые начальными константами  $\sigma_X, \sigma_Y$  и углом поворота. Соответствующая плотность имеет вид

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1}{\sigma_1}\right)\left(\frac{x_2}{\sigma_2}\right) + \left(\frac{x_2}{\sigma_2}\right)^2\right]\right\}, \quad (1)$$

где множители  $1/\sqrt{1-\rho^2}$  перед экспоненциальной функцией и  $1/(1-\rho^2)$  в её аргументе обеспечивает выполнение нормировки:

$$\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 f(x_1, x_2) = 1.$$

Уравнение (1) представляет двухмерное нормальное распределение специального *центрированного* вида. Более общее представление

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

где  $\mu_1$  и  $\mu_2$  – математические ожидания величин  $X_1$  и  $X_2$  соответственно:

$$\begin{aligned} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 f(x_1, x_2) &= \mu_1, \\ \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2 f(x_1, x_2) &= \mu_2. \end{aligned}$$

## 25.3 Корреляции в нормальных распределениях

Вернёмся к центрированной двумерной плотности (1). Она удовлетворяет следующим соотношениям:

$$\begin{aligned} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 f(x_1, x_2) &= 0, & \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2 f(x_1, x_2) &= 0, \\ \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1^2 f(x_1, x_2) &= \sigma_1^2, & \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2^2 f(x_1, x_2) &= \sigma_2^2, \\ \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 x_2 f(x_1, x_2) &= \sigma_{12} = \rho \sigma_1 \sigma_2. \end{aligned}$$

Последнее утверждение может быть легко доказано с помощью производящей функции моментов

$$M_{X_1, X_2}(t_1, t_2) \equiv \mathbb{E} \exp\{t_1 X_1 + t_2 X_2\} = \exp\{[(\sigma_1 t_1)^2 + 2\rho(\sigma_1 t_1)(\sigma_2 t_2) + (\sigma_2 t_2)^2]/2\}. \quad (2)$$

Обозначая экспоненту  $\{\dots\}$  через  $\Omega(t_1, t_2)$  и учитывая, что  $\Omega(0, 0) = 0$ ,  $\partial\Omega/\partial t_1|_{t_1=0, t_2=0} = \partial\Omega/\partial t_2|_{t_1=0, t_2=0} = 0$ , получаем:

$$\sigma_{12} = \left. \frac{\partial^2 M_{X_1, X_2}(t_1, t_2)}{\partial t_1 \partial t_2} \right|_{t_1=0, t_2=0} = \left\{ \frac{\partial^2 \Omega}{\partial t_2 \partial t_1} e^{\Omega} + \frac{\partial \Omega}{\partial t_2} \frac{\partial \Omega}{\partial t_1} e^{\Omega} \right\}_{t_1=0, t_2=0} = \rho \sigma_1 \sigma_2.$$

Таким образом, константа, которую мы обозначаем  $\rho$  представляет собой ничто иное, как *коэффициент корреляции* (смотри ...). Из этого выражения следует важная теорема:

**Теорема.** *Если две нормальные случайные величины некоррелированы, то они независимы.*

## 25.4 Оценка коэффициента корреляции

Экспериментальные исследования статистической взаимосвязи между случайными величинами  $X$  и  $Y$  основаны на парных измерениях их реализаций  $(X_1, Y_1)$ ,  $(X_2, Y_2)$ , ...,  $(X_n, Y_n)$ . Результаты можно собрать в таблицу или ввести в компьютер, но сначала представим их на графике рассеяния. График рассеяния scatterplot – это график на координатной плоскости, который показывает  $x - y$  координату каждой пары  $X_j, Y_j$ . Используя график рассеяния, можно построить двумерную гистограмму путём разбиения плоскости  $x - y$  на малые прямоугольники  $\Delta x \times \Delta y$  и вычисляя соответствующие частоты. Но графики рассеяния сами по себе представляют хорошую наглядную картину наличия или отсутствия зависимости или взаимосвязи.

Самая популярная мера степени зависимости основана на выборочном коэффициенте корреляции  $\hat{\rho}_{X,Y}$ . Он конструируется по аналогии с генеральным коэффициентом корреляции

$$\begin{aligned} \rho_{X,Y} &= \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}, & \sigma_{X,Y}^2 &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \equiv \mathbb{E}(XY) - \mu_X \mu_Y, \\ \sigma_X &= \sqrt{\mathbb{E}X^2 - \mu_X^2}, & \sigma_Y &= \sqrt{\mathbb{E}Y^2 - \mu_Y^2} \end{aligned}$$

и имеет вид:

$$\hat{\rho}_{X,Y} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y}, \quad \hat{\sigma}_{X,Y} = \frac{1}{n} \sum_{j=1}^n [(X_j - \bar{X})(Y_j - \bar{Y})] = \frac{1}{n} \sum_{j=1}^n X_j Y_j - \left( \frac{1}{n} \sum_{j=1}^n X_j \right) \left( \frac{1}{n} \sum_{j=1}^n Y_j \right),$$

$$\hat{\sigma}_X^2 = \frac{1}{n} \sum_{j=1}^n X_j^2 - \left( \frac{1}{n} \sum_{j=1}^n X_j \right)^2, \quad \hat{\sigma}_Y^2 = \frac{1}{n} \sum_{j=1}^n Y_j^2 - \left( \frac{1}{n} \sum_{j=1}^n Y_j \right)^2.$$

В результате, мы имеем

$$\hat{\rho}_{X,Y} = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}. \quad (1)$$

## 25.5 Пример

Данные по возрасту и цене для выборки из 11 автомобилей Nissan Zs представлены в первых двух колонках следующей таблицы.

Возраст (годы) $X$	Цена (\$ 100s $Y$ )	$XY$	$X^2$	$Y^2$
5	85	425	25	7,225
4	103	412	16	10,609
6	70	420	36	4,900
5	82	410	25	6,724
5	98	445	25	7,921
5	98	490	25	9,604
6	66	396	36	4,356
6	95	570	36	9,025
2	169	338	4	28,561
7	70	490	49	4,900
7	48	336	49	2,301
$\sum X = 58$	$\sum Y = 975$	$\sum XY = 4732$	$\sum X^2 = 326$	$\sum Y^2 = 96,129$

При подстановке сумм из последней строки таблицы в формулу (1), мы получаем

$$\hat{\rho} = \frac{11 \cdot 4732 - 58 \cdot 975}{\sqrt{11 \cdot 326 - (58)^2} \sqrt{11 \cdot 96,129 - (975)^2}} = -0.924.$$

Закключение: существует сильная *отрицательная* линейная корреляция между возрастом и ценой Nissan Zs.

## 26 Лекция 25. Регрессионный анализ

### 26.1 Линейная регрессионная модель

Вспомним, что полное описание двух независимых нормальных случайных величин  $X$  и  $Y$  требует знания четырёх параметров:  $\mu_X, \sigma_X, \mu_Y$  и  $\sigma_Y$ , в то время как для зависимых нормальных случайных величин мы должны знать пятый параметр  $\rho_{X,Y}$ . Вычисляя условную плотность

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y} \exp \left\{ -\frac{[y - (\mu_X + \rho(\sigma_Y/\sigma_X)(x - \mu_X))]^2}{2(1-\rho^2)\sigma_Y^2} \right\}. \quad (1)$$

Как можно видеть из этого выражения, условное среднее случайной величины  $Y$  является линейной функцией данного значения  $x$  другой величины  $X$

$$EY|x = \mu_X + \rho(\sigma_Y/\sigma_X)(x - \mu_X).$$

Это означает, что условная случайная величина  $Y|x$  может быть представлена как

$$Y|x = \alpha + \beta x + \epsilon, \quad (2)$$

где

$$\alpha = (1 - \rho(\sigma_Y/\sigma_X))\mu_X, \quad \beta = \rho(\sigma_Y/\sigma_X)$$

и  $\epsilon$  – нормальная случайная величина с нулевым средним и стандартным отклонением

$$\sigma_\epsilon = \sqrt{1 - \rho^2}\sigma_Y. \quad (3)$$

Очевидно, стандартное отклонение  $Y$  при заданном  $X = x$  меньше, чем у безусловной величины  $Y$  и этот факт можно использовать для предсказания одной из двух случайных величин, скажем  $Y$ , зная другую величину  $X$ .

Когда совместное распределение величин  $X$  и  $Y$  не является нормальным, уравнение (1) может не выполняться, но если оно выполняется, мы можем использовать уравнение (2) для предсказания значения  $Y$  обладая информацией о  $X$ . Для того, чтобы сделать это, представим выборочные данные парами  $(x_j, Y_j)$ ,  $j = 1, \dots, n$ , рассматривая первую переменную как неслучайное число и подбирая коэффициенты  $\alpha$  и  $\beta$  таким образом, чтобы прямая

$$y = \alpha + \beta x \quad (4)$$

проходила как можно ближе ко всем точкам на плоскости  $xy$ . Эта схема называется *линейной регрессионной моделью*, уравнение (4) называется *регрессионным уравнением*, а числа  $\alpha$   $\beta$  называются *коэффициентами регрессии*.

### 26.2 Два примера подгонки ???Fitting

Рассмотрим задачу подгонки прямой линии к четырём данным точкам, представленным в таблице:

$x$	1	1	2	4
$y$	1	2	2	6

Эти точки не лежат на одной прямой линии, то есть провести через них прямую можно только приближённо. Существует бесконечно много линейных аппроксимаций этих данных. Рассмотрим три из них:

$$y = 0.50 + 1.25x, \quad (A)$$

$$y = -0.40 + 1.60x, \quad (B)$$

и

$$y = -0.25 + 1.50x. \quad (C)$$

Возьмём точку  $x = 2$  с заданным значением  $y = 2$ . Значения  $y$  предсказанные линиями  $A$ ,  $B$ , и  $C$  для этой точки равны

$$y_A = 0.50 + 1.25 \cdot 2 = 3.00, \quad y_B = -0.40 + 1.60 \cdot 2 = 2.80, \quad y_C = -0.25 + 1.50 \cdot 2 = 2.75.$$

Соответствующие ошибки

$$\epsilon_A = y - y_A = 2 - 3 = -1, \quad \epsilon_B = y - y_B = 2 - 2.80 = -0.80, \quad \epsilon_C = y - y_C = 2 - 2.75 = -0.75.$$

таким образом,  $C$ -аппроксимация более точна в этой точке, чем аппроксимации  $A$  и  $B$ , а  $B$ -аппроксимация более точна, чем  $A$ .

Рассмотрим теперь другую точку, скажем, точку  $x = 4$ . Здесь  $y = 6$  при

$$y_A = 5.50, \quad y_B = 6.00, \quad y_C = 5.75$$

с соответствующими ошибками

$$\epsilon_A = 0.50, \quad \epsilon_B = 0.00, \quad \epsilon_C = 0.25.$$

В этой точке, лучшая аппроксимация даётся линией  $B$ , а худшую подгонку обеспечивает  $A$ .

## 26.3 Метод наименьших квадратов

Это один из методов для получения регрессионных коэффициентов с использованием данных из выборки. Эти коэффициенты следует выбирать таким образом, чтобы *сумма квадратов ошибок*

$$SSE \equiv \sum_{j=1}^n \epsilon_j^2 \quad (5)$$

принимала наименьшее значение (это и дало название методу). Находим из уравнения (2) ошибки

$$SSE = \sum_{j=1}^n (Y_j - \alpha - \beta x_j)^2.$$

Координаты  $\alpha, \beta$  точки, где  $SSE$  достигает своего минимального значения удовлетворяет уравнениям

$$\frac{\partial SSE}{\partial \alpha} = 0, \quad \frac{\partial SSE}{\partial \beta} = 0.$$

Подставляя сюда уравнение (5) и выполняя дифференцирование, получаем:

$$\frac{\partial \sum_{j=1}^n (Y_j - \alpha - \beta x_j)^2}{\partial \alpha} = -2 \sum_{j=1}^n (Y_j - \alpha - \beta x_j) = 0$$

и

$$\frac{\partial \sum_{j=1}^n (Y_j - \alpha - \beta x_j)^2}{\partial \beta} = -2 \sum_{j=1}^n (Y_j - \alpha - \beta x_j) x_j = 0.$$

Вводя обозначения

$$\bar{x} = \frac{1}{n} \sum x, \quad \overline{x^2} = \frac{1}{n} \sum x^2, \quad \bar{Y} = \frac{1}{n} \sum Y, \quad \overline{xY} = \frac{1}{n} \sum xY,$$

мы представляем решение системы в виде:

$$\alpha = \bar{Y} - \beta \bar{x}, \quad \beta = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2}.$$

## 26.4 Некоторые другие формы второго коэффициента

$$\beta = \frac{n \sum xY - (\sum x)(\sum Y)}{n \sum x^2 - (\sum x)^2}.$$

$$\beta = \frac{\sum (x_j - \bar{x})(Y_j - \bar{Y})}{\sum (x_j - \bar{x})^2}.$$

## 26.5 Пример

Данные выборки представлены в таблице:

$x$	$Y$
0	0
0	1
2	1
2	2

Вычисление  $\alpha$  и  $\beta$  продемонстрировано в следующей таблице:

$x$	$Y$	$xY$	$x^2$
0	0	0	0
0	1	0	0
2	1	2	4
2	2	4	4
$\sum = 4$	$\sum = 4$	$\sum = 6$	$\sum = 8$

В результате имеем:

$$\beta = \frac{n \sum (xY) - (\sum x)(\sum Y)}{n \sum x^2 - (\sum x)^2} = \frac{4 \cdot 6 - 4 \cdot 4}{4 \cdot 8 - 4^2} = \frac{24 - 16}{32 - 16} = \frac{8}{16} = \frac{1}{2},$$

$$\alpha = \frac{\sum Y - \beta \sum x}{n} = \frac{4 - 4/2}{4} = \frac{1}{2}.$$

Уравнение регрессии принимает вид:

$$y = \frac{x + 1}{2}.$$

## 27 Лекция 26. Вариационный анализ ???Analysis of Variance (ANOVA)

### 27.1 Основная идея ANOVA

Причина, по которой слово *вариация* присутствует в ANOVA заключается в том, что процедура сравнения средних использует анализ вариации в выборочных данных. Чтобы увидеть, как это работает, предположим, что независимые случайные величины получены из двух генеральных совокупностей с неизвестными средними  $\mu_1$  and  $\mu_2$ . Далее предположим, что средние двух выборок равны  $\bar{X}_1 = 20$  и  $\bar{X}_2 = 25$ . Можем ли мы уверенно утверждать, что  $\mu_1 \neq \mu_2$ ? Нет, мы не можем. Чтобы ответить на этот вопрос, мы должны знать вариацию внутри выборок.

Если разница между выборочными средними не велика по отношению к вариации внутри выборок, мы не можем утверждать, что  $\mu_1 \neq \mu_2$ . В этом случае эти две выборки перекрывают друг друга частично или полностью и не ясно, вызвана ли разница между выборочными средними разницей между генеральными средними или вариацией внутри генеральных совокупностей.

Рассмотрим противоположный случай, когда части выборочных данных концентрируются вокруг соответствующих выборочных средних, так что не только выборки не перекрываются, но даже наблюдается некоторый пустой промежуток между выборками из разных генеральных совокупностей. В этот раз мы можем сделать вывод, что  $\mu_1 \neq \mu_2$  поскольку кажется ясным, что разница между выборочными средними возникает благодаря разнице между генеральными средними, а не из-за вариации внутри генеральных совокупностей. Другими словами, поскольку вариация между выборочными средними велика по отношению к вариациям внутри выборок, мы можем заключить, что  $\mu_1 \neq \mu_2$ .

В общем случае, мы можем работать с более чем двумя генеральными совокупностями. Тогда основная идея ANOVA может быть сформулирована в следующем виде.

- (1) Возьмём независимые случайные выборки из заданных генеральных совокупностей.
- (2) Вычислим выборочные средние.
- (3) Если вариация между выборочными средними велика по отношению к вариации внутри выборок, заключаем что средние генеральных совокупностей *не равны*.

Чтобы сделать эту процедуру точной нам необходима количественная мера для вариации.

### 27.2 Меры вариации

Обозначим через  $k$  число генеральных совокупностей, через  $n_1, \dots, n_k$  – соответствующие объёмы выборок, тогда  $n = n_1 + \dots + n_k$  – полный объём выборки,  $\bar{X}_1, \dots, \bar{X}_k$  – выборочные средние,  $s_1^2, \dots, s_k^2$  – выборочные дисперсии и  $\bar{X}$  полное выборочное среднее.

**Мера вариации среди выборочных средних:** ???treatment mean square *MSTR*:

$$MSTR = \frac{n_1(\bar{X}_1 - \bar{X})^2 + \dots + n_k(\bar{X}_k - \bar{X})^2}{k - 1}.$$

Это похоже на выборочную дисперсию. В специальном случае  $n_1 = \dots = n_k = 1$  we get  $k = n$ ,

$$\bar{X}_1 = \frac{1}{n}(X_1 + \dots + X_{n_1}) = X_1, \dots, \bar{X}_n = X_n$$



и, как результат, обычная формула для выборочной дисперсии

$$MSTR = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2.$$

**Мера вариации внутри выборок:** средний квадрат ошибки ???the error mean square *MSE*:

$$MSE = \frac{(n_1 - 1)s_1^2 + \dots + (n_k - 1)s_k^2}{n - k}.$$

Заметим, что для  $k = 1$ ,  $n_1 = n$  и  $MSE = \frac{n-1}{n-1}s_1^2 = s_1^2$ .

## 27.3 Тестовая статистика для ANOVA

Предположим, независимые случайные выборки объёмов  $n_1, \dots, n_k$  получены из  $k$  нормальных генеральных совокупностей со средними  $\mu_1, \dots, \mu_k$  соответственно. Предположим также, что стандартные отклонения этих  $k$  генеральных совокупностей равны между собой:  $\sigma_1 = \dots = \sigma_k$ . Есть теорема, утверждающая, что при этих условиях, если  $\mu_1 = \dots = \mu_k$ , тогда

$$\hat{F} \equiv \frac{MSTR}{MSE} \stackrel{d}{=} F(k-1, n-k),$$

где  $k-1$  и  $n-k$  – степени свободы  $F$ -распределения.

Когда тестовая статистика  $\hat{F}$  превышает некоторое критическое значение  $f_\alpha$  из  $F$ -распределения, это означает, что равенство не выполняется и по меньшей мере две из генеральных совокупностей имеют разные средние.

## 27.4 Процедура проверки ANOVA

**Шаг 1.** Сформулируем нулевую и альтернативную гипотезы:

$$H_0 : \mu_1 = \dots = \mu_k \quad \text{vs} \quad H_1 : \text{не все } \mu_j \text{ одинаковы.}$$

**Шаг 2.** Определим уровень значимости  $\alpha$  (обычно, 0.05).

**Шаг 3.** Найдём критическое значение  $f_\alpha(k-1, n-k)$ .

**Шаг 4.** Вычислим тестовую статистику  $\hat{F} = MSTR/MSE$ .

**Шаг 5.** Если  $\hat{F}$  попадает в область отвергания, то есть, если  $\hat{F} > f_\alpha(k-1, n-k)$ , тогда отклоняем  $H_0$ , иначе, не отклоняем  $H_0$ .

**Шаг 6.** Формулируем заключение словами.

## 27.5 Пример.

Независимые случайные выборки домовладений ???households в четырёх регионах США дали следующие данные по прошлогоднему потреблению электроэнергии.

Северо-восток	Средний запад	Юг	Запад
15	17	11	10
10	12	7	12
13	18	9	8
14	13	13	7
13	15		9
	12		

Есть ли какая-нибудь разница в потреблении энергии в четырёх регионах?

**Решение.**

**Шаг 1.** Формулируем нулевую и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_1 : \text{не все } \mu_j \text{ одинаковы.}$$

**Шаг 2.** Определяем уровень значимости:  $\alpha = 0.05$ .

**Шаг 3.** Находим критическое значение  $f_{\alpha}(k-1, n-k) = f_{0.05}(3, 16) = 3.24$ .

**Шаг 4.** Вычисляем тестовую статистику  $\hat{F} = MSTR/MSE = 6.32$ .

**Шаг 5.** Поскольку  $\hat{F} = 6.32 > f_{0.05}(3, 16) = 3.24$ , то она попадает в область отвергания и мы отклоняем  $H_0$ .

**Шаг 6.** При уровне значимости 5%, данные обеспечивают достаточное свидетельство в пользу того, что среднее потребление энергии в прошлом году в этих четырёх регионах не равны: по меньшей мере в двух из них среднее потребление отличается.

## 28 The LAST Final exam problems (Fall 2005)

### 28.1 Задача 1

Случайная величина  $X$  задана распределением вероятностей

$x$	0	1	2
$f(x)$	1/3	1/3	1/3

и  $X_j, j = 1, 2, \dots$  – независимые копии величины  $X$ .

**Найти:** а)  $\mu$  и  $\sigma$  для  $X$ ; б)  $\mu$  и  $\sigma$  для  $\sum_{j=1}^{96} X_j$ ; в)  $\mu$  и  $\sigma$  для  $\bar{X} \equiv \frac{1}{96} \sum_{j=1}^{96} X_j$ ;  
д)  $P(\bar{X} < 5/6)$ ; е)  $P(\bar{X} < 7/6)$ ;  $P(11/12 < \bar{X} < 13/12)$ .

**Решение.**

(Будьте внимательны: в таблице представлено распределение вероятностей, а не выборка! Таким образом, вам следует использовать формулы теории вероятностей, а не статистическое оценивание.)

(а)

$$\mu_X = EX = \sum_{x=0}^2 xf(x) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1,$$

$$\sigma_X^2 = \sum_{x=0}^2 x^2 f(x) - \mu_X^2 = 0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} + 2^2 \cdot \frac{1}{3} - 1^2 = \frac{5}{3} - 1 = \frac{2}{3},$$

$$\sigma = \sqrt{2/3} \approx 0.82.$$

(б)

$$\mu_{X_1+\dots+X_n} = E\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n EX_j = \sum_{j=1}^n \mu_X = n\mu_X = 96 \cdot 1 = 96,$$

$$\sigma_{X_1+\dots+X_n}^2 = \sum_{j=1}^n \sigma_X^2 = n\sigma_X^2 = 96 \cdot \frac{2}{3} = 64, \quad \sigma_{X_1+\dots+X_n} = 8.$$

(в)

$$\mu_{\bar{X}} = E\left(\frac{1}{n} \sum_{j=1}^n X_j\right) = \frac{1}{n} \mu_{X_1+\dots+X_n} = \frac{1}{96} \cdot 96 = 1,$$

$$\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \sqrt{\frac{1}{n^2} \sigma_{X_1+\dots+X_n}^2} = \frac{\sigma_{X_1+\dots+X_n}}{n} = \frac{8}{96} \approx 0.083.$$

(d) здесь и далее, используйте центральную предельную теорему (согласно которой  $\bar{X} \stackrel{d}{=} \mu_{\bar{X}} + \sigma_{\bar{X}}Z$ , где  $Z$  – стандартная нормальная случайная величина) и таблицу Table A.3 для площади под нормальной кривой.

$$\begin{aligned} P(\bar{X} < 5/6) &= P(\mu_{\bar{X}} + \sigma_{\bar{X}}Z < 5/6) = P\left(Z < \frac{5/6 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = \\ &= P\left(Z < \frac{5/6 - 1}{8/96}\right) = P(Z < -2) = 0.0228. \end{aligned}$$

(e)

$$P(\bar{X} < 7/6) = P\left(Z < \frac{7/6 - 1}{8/96}\right) = P(Z < 2) = 0.9772.$$

(f)

$$\begin{aligned} P(11/12 < \bar{X} < 13/12) &= P\left(\frac{11/12 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{13/12 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) = \\ P(-1 < Z < 1) &= P(Z < 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826. \end{aligned}$$

## 28.2 Задача 2.

Средняя концентрация загрязнений в реке, определённая по выборке из 36 измерений, проведённых в различных местах, оказалась равна 0.300 от предельно допустимого уровня. Предполагая, что стандартное отклонение равно 0.030,

**найти:** а) 95%-й доверительный интервал для средней генеральной концентрации; б) 99%-й доверительный интервал для средней генеральной концентрации.

**Решение.** Объём выборки  $n = 36$  превышает 30 и генеральное стандартное отклонение  $\sigma_X = 0.03$  известно, тогда

$$\frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \stackrel{d}{=} Z \text{ (standard normal random variable)}. \quad (1)$$

Доверительный интервал для  $Z$  равен  $(-z_{\alpha/2}, z_{\alpha/2})$ , где  $z_{0.025} = 1.960$  и  $z_{0.005} = 2.576$  (см. нижний ряд таблицы Table A.4). После подстановки этих значений и  $\bar{X} = 0.30$  в уравнение (1) мы получаем:

$$\mu_{\bar{X}}^{(L,U)} = \bar{X} \pm (\sigma_X/\sqrt{n})z_{\alpha/2} = 0.300 \pm 0.005z_{\alpha/2}.$$

Используя критическое значение, мы приходим к результату: (a) 95%-й доверительный интервал для  $\mu_X$  равен (0.290, 0.310), (b) 99%-й доверительный интервал для  $\mu_X$  равен (0.287, 0.313).

## 28.3 Задача 3.

Результаты очень сложных независимых астрономических наблюдений одной и той же величины представлены в следующей выборке: 10.2; 10.4; 9.6; 9.8; 10.0.

**Найти:** а) 95%-й доверительный интервал для *истинного значения* (в смысле среднего значения нормальной генеральной совокупности); б) 99%-й доверительный интервал для того же самого.

**Решение.** Объём выборки  $n = 5$  меньше, чем 30, и  $\sigma$  неизвестна, таким образом, следует вычислить выборочное среднее и стандартное отклонение

$$\bar{X} = \frac{1}{5}(\sum X) = \frac{50}{5} = 10,$$

$$s_X^2 = \frac{1}{n-1} \sum (X_j - \bar{X})^2 = \frac{(0.2)^2 + (0.4)^2 + (0.4)^2 + (0.2)^2}{4} = \frac{0.40}{4} = 10,$$

и использовать  $T$ -статистику

$$\frac{\bar{X} - \mu_X}{s_X/\sqrt{n}} \stackrel{d}{=} T(n-1) :$$

$$\mu_{\bar{X}}^{(L,U)} = \bar{X} \pm (s_X/\sqrt{n})t_{\alpha/2}(n-1).$$

Критические значения  $t_{0.025}(4) = 2.776$  и  $t_{0.005}(4) = 4.604$  могут быть найдены в таблице Table A.4, и мы приходим к результату: (а) 95%-й доверительный интервал для  $\mu_X$  is (9.61, 10.39), 99%-й доверительный интервал для  $\mu_X$  is (9.35, 10.65).

## 28.4 Задача 4.

В начальной школе среднее число детей, которые употребляли "горячий обед" каждый день на протяжении нескольких последних лет было равно  $\mu_0 = 125$  детей в день со стандартным отклонением  $\sigma = 4$ . Недавняя случайная выборка за *пять дней* (то есть  $n = 5$ ) дала среднее число  $\bar{X} = 129$  детей в день. Из-за изменений условий жизни возникает вопрос, остаётся ли генеральное среднее по-прежнему таким же ( $H_0 : \mu = \mu_0$ ).

**Протестировать** эту гипотезу с альтернативой  $H_1 : \mu \neq \mu_0$ .

**Решение.**

**Шаг 1.** Гипотезы:  $H_0 : \mu = \mu_0 = 125$  vs  $H_1 : \mu \neq \mu_0$ .

**Шаг 2.** Уровень значимости:  $\alpha = 0.05$ .

**Шаг 3.** Выбираем  $Z$  в качестве тестовой статистики (хотя  $n = 5 < 30$ , но мы знаем  $\sigma$ ) и найдём критическое значение  $z_{\alpha/2} = z_{0.025} = 1.96$  (двухсторонний тест).

**Шаг 4.** Вычисляем  $\hat{Z} \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{129 - 125}{4/\sqrt{5}} = 2.24$ .

**Шаг 5.**  $\hat{Z}$  попадает в область отвергания, поэтому мы отклоняем  $H_0$ .

**Шаг 6.** При 5%-м уровне значимости, данные подтверждают, что генеральное среднее  $\mu$  не то же самое, что несколько лет тому назад.

## 28.5 Задача 5.

Предположим, разработчик нового изоляционного материала хочет знать, на какую величину сожмётся образец материала толщиной 2 дюйма, при воздействии на него давлением различной величины. Пять кусочков материала были отобраны и протестированы при различных давлениях. Данные представлены в таблице:

Давление $x$	1	2	3	4	5
Сжатие $y$	1	1.5	2	2	3

**Найти** методом наименьших квадратов уравнение регрессии, которое описывает взаимосвязь между сжатием и давлением.

**Решение.**

$$\begin{aligned}\sum x &= 15, \quad \sum x^2 = 55, \quad \sum xY = 33, \quad \sum Y = 9.5, \\ \beta &= \frac{n \sum xY - (\sum x)(\sum Y)}{n \sum x^2 - (\sum x)^2} = \frac{5 \cdot 33 - 15 \cdot 9.5}{5 \cdot 55 - (15)^2} = 0.45, \\ \alpha &= (1/n) \sum Y - (\beta/n) \sum x = 1.90 - 1.35 = 0.55.\end{aligned}$$

Таким образом, уравнение линейной регрессии имеет вид:

$$y = 0.55 + 0.45x.$$

Name.....

## Final exam STAT 312 Spring 2005 (May 10)

**Задача 1(14 баллов).** Случайная выборка объёма 64 получается из нормальной генеральной совокупности с  $\mu = 51.4$  и  $\sigma = 6.8$

**Найти:** Вероятности (a)  $P(\sum X < 3290)$ , (b)  $P(\sum X > 3290)$

**Задача 2(12 баллов).** Средняя концентрация загрязнений в реке, определённая по выборке из 36 измерений, проведённых в различных местах, оказалась равна 0.300 от предельно допустимого уровня. Предполагая, что стандартное отклонение равно 0.030,

**найти:** а) 95%-й доверительный интервал для средней генеральной концентрации; б) 99%-й доверительный интервал для средней генеральной концентрации.

**Задача 3(12 баллов).** Результаты очень сложных независимых астрономических наблюдений одной и той же величины представлены в следующей выборке: 10.2; 10.4; 9.6; 9.8; 10.0. **Найти:** а) 95%-й доверительный интервал для *истинного значения* (в смысле среднего значения нормальной генеральной совокупности); б) 99%-й доверительный интервал для того же самого.



**Задача 4(12 баллов).** В начальной школе среднее число детей, которые употребляли "горячий обед" каждый день на протяжении нескольких последних лет было равно  $\mu_0 = 125$  детей в день со стандартным отклонением  $\sigma = 4$ . Недавняя случайная выборка за *пять дней* (то есть  $n = 5$ ) дала среднее число  $\bar{X} = 129$  детей в день. Из-за изменений условий жизни возникает вопрос, остаётся ли генеральное среднее по-прежнему таким же ( $H_0 : \mu = \mu_0$ ).

**Протестировать** эту гипотезу с альтернативой  $H_1 : \mu \neq \mu_0$ .

**Задача 5(10 баллов).**Предположим, разработчик нового изоляционного материала хочет знать, на какую величину сожмётся образец материала толщиной 2 дюйма, при воздействии на него давлением различной величины. Пять кусочков материала были отобраны и протестированы при различных давлениях. Данные представлены в таблице:

Давление $x$	1	2	3	4	5
Сжатие $y$	1	1.5	2	2	3

**Найти** методом наименьших квадратов уравнение регрессии, которое описывает взаимосвязь между сжатием и давлением.

# РЕШЕНИЯ ДОМАШНИХ ЗАДАНИЙ HOMEWORK 1 - SOLUTION

1. Пусть  $S = \{0, 1, 2, 3, 4, 5\}$ ,  $A = \{1, 3, 5\}$ ,  $B = \{0, 2, 4\}$ ,  $C = \{2, 3, 4\}$ ,  $D = \{1, 4, 5\}$ .  
Выпишем исходы событий:

$$(a) A \cup B; \quad (b) A \cap B; \quad (c) C'; \quad (d) (C' \cap D) \cup B; \quad (e) (S \cap C)'; \quad (f) A \cap C \cap D'$$

**Ответ:**

$$(a) A \cup B = S; \quad (b) A \cap B = \emptyset; \quad (c) C' = \{0, 1, 5\}; \quad (d) (C' \cap D) \cup B = \{0, 1, 2, 4, 5\};$$

$$(e) (S \cap C)' = \{0, 1, 5\}; \quad (f) A \cap C \cap D' = \{3\}.$$

2. Дано  $S = \{x : 0 \leq x \leq 3\}$ ,  $A = \{x : 0 \leq x \leq 2\}$ ,  $B = \{x : 1 \leq x \leq 3\}$ , найти

$$(a) A'; \quad (b) B'; \quad (c) A \cup B; \quad (d) A \cap B; \quad (e) A \cap B'; \quad (f) A' \cap B'.$$

**Ответ:** (a)  $2 < x \leq 3$ ; (b)  $0 \leq x < 1$ ; (c)  $S$ ; (d)  $1 \leq x \leq 2$ ; (e)  $0 \leq x < 1$ ; (f)  $\emptyset$ .

3. Монета подбрасывается пять раз. Перечислить все исходы. Найти вероятности событий

(a)  $A$  : все испытания дают  $T$ ;

(b)  $B$  : одно из испытаний даёт  $H$ , остальные дают  $T$ ;

(c)  $C$  : два из испытаний дают  $H$ , остальные —  $T$ ;

(d)  $D$  : первое испытание даёт  $H$ , следующее испытание —  $H$  ;

(e)  $E$  : первые три испытания дают  $H$ .

**Ответ:** (a)  $1/32$ ; (b)  $5/32$ ; (c)  $5/16$ ; (d)  $1/4$ ; (e)  $1/8$ .

4. Бросается пара игральных костей (красная и зелёная) и результат записывается как  $X$  (для красной) и  $Y$  (для зелёной). Найти вероятности событий:

$$A_k : X + Y = k, \quad k = 2, 3, \dots, 12.$$

**Ответ:**  $1/36$ ;  $1/18$ ;  $1/12$ ;  $1/9$ ;  $5/36$ ;  $1/6$ ;  $5/36$ ;  $1/9$ ;  $1/12$ ;  $1/18$ ;  $1/36$ .

5. Бросается пара игральных костей (красная и зелёная) и результат записывается как  $X$  (для красной) и  $Y$  (для зелёной). Найти вероятности событий:

$$(a) A : X - Y = 3; \quad (b) B : |X - Y| = 3; \quad (c) C : X < Y; \quad (d) D : X \leq Y;$$

$$(e) E : X + Y < 6; \quad (f) F : X > 3; \quad (g) G : Y^2 = X.$$

**Ответ:** (a)  $1/12$ ; (b)  $1/6$ ; (c)  $5/12$ ; (d)  $7/12$ ; (e)  $5/18$ ; (f)  $1/2$ ; (g)  $1/18$ .

6. Дано  $P(A \cap B') = 0.3$ ,  $P(A' \cap B) = 0.2$ ,  $P((A \cap B)') = 0.8$  найти:

$$(1) P(A \cap B); \quad (2) P(A); \quad (3) P(B); \quad (4) P(A \cup B).$$

**Решение:** (1)  $P(A \cap B) = 1 - P((A \cap B)') = 1 - 0.8 = 0.2$ ; (2)  $P(A) = P(A \cap B) + P(A \cap B') = 0.2 + 0.3 = 0.5$ ; (3)  $P(B) = P(A \cap B) + P(A' \cap B) = 0.2 + 0.2 = 0.4$ ;  $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.4 - 0.2 = 0.7$ .

## 29 Лекция 27. Типичные задачи

### 29.1 Пример

Количество химического компонента  $y$ , который растворён в 100 граммах воды при разных температурах  $x$  представлено в таблице

$x_j(^{\circ}C)$	$y$ (граммы)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

а) Найти уравнение регрессии.

$$y = 5.8254 + 0.5676x.$$

б) Нарисовать линию на диаграмме рассеяния.

с) Оценить количество химиката, которое растворится в 100 граммах воды при  $50^{\circ} C$ .

$$y(50) = 34.205.$$

## 29.2 Тест на корреляцию

Проверка на корреляцию, то есть гипотезы  $H_0 : \rho = 0$  против альтернативной  $H_1 : \rho \neq 0$  выглядит следующим образом.

- 1) Формулируем  $H_0 : \rho = 0$ ,  $H_1 : \rho \neq 0$ .
- 2) Выбираем  $\alpha$ .
- 3) Находим критические значения  $\pm t_{\alpha/2}$ .
- 4) Вычисляем статистику  $t = \frac{b}{s/\sqrt{S_{xx}}}$ , где  $b = \frac{S_{xy}}{S_{xx}}$ ,  $s = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$ .
- 5) Заключение: если  $t < -t_{\alpha/2}$  или  $t > t_{\alpha/2}$ , то  $H_0$  отклоняется.

**Замечание.** Статистика может быть представлена в виде:  $t = \frac{r\sqrt{n-2}}{\sqrt{1-\rho^2}}$

## 29.3 Пример (???page 394).

- 1)  $H_0 : \rho = 0$ ,  $H_1 \neq 0$ .
- 2)  $\alpha = 0.05$
- 3)  $n = 29$ ,  $t_{0.025}(27) = 2.052$  (Table A4, page 672.)
- 4)  $t = \frac{0.943\sqrt{27}}{1-(0.943)^2} = 14.79$
- 5) Заключение: поскольку  $14.79 > 2.052$  нуль-гипотеза  $H_0$  отвергается.

## 30 Лекция 25. Корреляция в многомерных распределениях

### 30.1 Биномиальное распределение

Начнём с хорошо известного биномиального распределения. Формально, выражение для для его вероятностей может быть получено из биномиальной формулы

$$1 = 1^n = (p_1 + p_2)^n = \sum_{\{n_1+n_2=n\}} \frac{n!}{n_1!n_2!} p_1^{n_1} p_2^{n_2} = \sum_{\{x_1+x_2=n\}} f(x_1, x_2),$$

где

$$f(x_1, x_2) = \frac{n!}{x_1!x_2!} p_1^{x_1} p_2^{x_2}, \quad p_1 + p_2 = 1, \quad x_1 + x_2 = n$$

– совместная плотность вероятности двух случайных величин целого порядка  $X_1$  и  $X_2$  со средними  $\mu_{X_1} = np_1$ ,  $\mu_{X_2} = np_2$  и дисперсиями  $\sigma_{X_1}^2 = \sigma_{X_2}^2 = np_1p_2$ . Вычислим их ковариацию

$$\text{Cov}(X_1, X_2) = E(X_1X_2) - (EX_1)(EX_2) \equiv \sigma_{X_1, X_2},$$

используя соответствующую двумерную производящую функцию моментов:

$$M_{X_1, X_2}(t_1, t_2) \equiv E \exp\{t_1X_1 + t_2X_2\} = (p_1e^{t_1} + p_2e^{t_2})^n.$$

Вывести это выражение нетрудно, но мы лишь проверим его:

$$M_{X_1, X_2}(0, 0) = 1,$$

$$\left. \frac{\partial M_{X_1, X_2}(t_1, t_2)}{\partial t_1} \right|_{t_1=0, t_2=0} = np_1,$$

$$\left. \frac{\partial^2 M_{X_1, X_2}(t_1, t_2)}{\partial t_1^2} \right|_{t_1=0, t_2=0} - \mu_{X_1}^2 = np_1 p_2 = np_1(1 - p_2) = \sigma_{X_1}^2.$$

Эти результаты правильны. Теперь мы вычислим ковариацию:

$$\sigma_{X_1, X_2} = \left. \frac{\partial^2 M_{X_1, X_2}(t_1, t_2)}{\partial t_2 \partial t_1} \right|_{t_1=0, t_2=0} - \mu_{X_1} \mu_{X_2} = n(n-1)p_1 p_2 - n^2 p_1 p_2 = -np_1 p_2.$$

После деления его на произведение стандартных отклонений, мы получим коэффициент корреляции:

$$\rho = \frac{\sigma_{X_1, X_2}}{\sigma_{X_1} \sigma_{X_2}} = \frac{-np_1 p_2}{\sqrt{np_1 p_2} \sqrt{np_1 p_2}} = -1.$$

Этот результат понятен, поскольку случайные величины связаны соотношением  $X_2 = n - X_1$ , таким образом, они совершенно антикоррелируют.

## 30.2 Триномиальное распределение

Триномиальное распределение получается из уравнения

$$1 = 1^n = (p_1 + p_2 + p_3)^n = \sum_{\{n_1+n_2+n_3=n\}} \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3} = \sum_{\{x_1+x_2+x_3=n\}} f(x_1, x_2, x_3),$$

где

$$f(x_1, x_2, x_3) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}, \quad p_1 + p_2 + p_3 = 1, \quad x_1 + x_2 + x_3 = n.$$

Используя, как и ранее, соответствующую совместную производящую функцию моментов

$$M_{X_1, X_2, X_3}(t_1, t_2, t_3) = (p_1 e^{t_1} + p_2 e^{t_2} + p_3 e^{t_3})^n,$$

получаем:

$$\sigma_{X_1, X_2} = \left. \frac{\partial^2 M_{X_1, X_2, X_3}(t_1, t_2, t_3)}{\partial t_2 \partial t_1} \right|_{t_1=0, t_2=0, t_3=0} - \mu_{X_1} \mu_{X_2} = n(n-1)p_1 p_2 - n^2 p_1 p_2 = -np_1 p_2.$$

Но в этот раз  $\rho \neq -1$  так как  $\sigma_{X_1}^2 = np_1(1 - p_1)$  и  $\sigma_{X_2}^2 = np_2(1 - p_2)$ . В результате имеем

$$\rho_{12} = \frac{-np_1 p_2}{np_1(1 - p_1)np_2(1 - p_2)} = -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}.$$

Отметим: чем меньше  $p_1$  и  $p_2$  тем меньше корреляции.

## 30.3 Двухмерное нормальное распределение

Будем рассматривать две центрированные нормально распределённые случайные величины как координаты случайных точек  $(X, Y)$  с дисперсиями  $\sigma_X^2$  и  $\sigma_Y^2$  соответственно. Если эти координаты независимы,

$$\begin{aligned} f(x, y) &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{x^2}{2\sigma_X^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(-\frac{y^2}{2\sigma_Y^2}\right) = \\ &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left\{-\frac{1}{2}\left[\left(\frac{x}{\sigma_X}\right)^2 + \left(\frac{y}{\sigma_Y}\right)^2\right]\right\}. \end{aligned}$$

Легко видеть, что кривая постоянных величин  $f(x, y)$  представляет собой эллипс с главными осями, лежащими на координатных осях  $x$  и  $y$ :

$$\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2} = \text{const.} \quad (1)$$

Повернём систему координат вокруг её центра на произвольный угол и обозначим новые координаты через  $x_1, x_2$  (мы можем также представить, что мы вращаем плоскость с распределениями относительно координатных осей, такое преобразование также ведёт к изменению координат каждой точки  $(x, y) \mapsto (x_1, x_2)$ ). Уравнение (1) примет вид

$$\left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1}{\sigma_1}\right) \left(\frac{x_2}{\sigma_2}\right) + \left(\frac{x_2}{\sigma_2}\right)^2 = \text{const},$$

где  $\sigma_1 > 0$ ,  $\sigma_2 > 0$  и  $\rho \in [-1, 1]$  – новые постоянные, определяемые начальными константами  $\sigma_X, \sigma_Y$  и углом поворота. Соответствующая плотность имеет вид

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1}{\sigma_1}\right) \left(\frac{x_2}{\sigma_2}\right) + \left(\frac{x_2}{\sigma_2}\right)^2 \right] \right\}, \quad (1)$$

где множители  $1/\sqrt{1-\rho^2}$  перед экспоненциальной функцией и  $1/(1-\rho^2)$  в её аргументе обеспечивает выполнение нормировки:

$$\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 f(x_1, x_2) = 1.$$

Уравнение (1) представляет двумерное нормальное распределение специального *центрированного* вида. Более общее представление

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right] \right\},$$

где  $\mu_1$  и  $\mu_2$  – математические ожидания величин  $X_1$  и  $X_2$  соответственно:

$$\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 f(x_1, x_2) = \mu_1,$$

$$\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2 f(x_1, x_2) = \mu_2.$$

## 30.4 Корреляции в нормальных распределениях

Вернёмся к центрированной двумерной плотности (1). Она удовлетворяет следующим соотношениям:

$$\begin{aligned} \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 f(x_1, x_2) &= 0, & \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2 f(x_1, x_2) &= 0, \\ \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1^2 f(x_1, x_2) &= \sigma_1^2, & \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_2^2 f(x_1, x_2) &= \sigma_2^2, \end{aligned}$$

$$\int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 x_1 x_2 f(x_1, x_2) = \sigma_{12} = \rho \sigma_1 \sigma_2.$$

Последнее утверждение может быть легко доказано с помощью производящей функции моментов

$$M_{X_1, X_2}(t_1, t_2) \equiv \mathbb{E} \exp\{t_1 X_1 + t_2 X_2\} = \exp\{[(\sigma_1 t_1)^2 + 2\rho(\sigma_1 t_1)(\sigma_2 t_2) + (\sigma_2 t_2)^2]/2\}. \quad (2)$$

Обозначая экспоненту  $\{\dots\}$  через  $\Omega(t_1, t_2)$  и учитывая, что  $\Omega(0, 0) = 0$ ,  $\partial\Omega/\partial t_1|_{t_1=0, t_2=0} = \partial\Omega/\partial t_2|_{t_1=0, t_2=0} = 0$ , получаем:

$$\sigma_{12} = \left. \frac{\partial^2 M_{X_1, X_2}(t_1, t_2)}{\partial t_1 \partial t_2} \right|_{t_1=0, t_2=0} = \left\{ \frac{\partial^2 \Omega}{\partial t_2 \partial t_1} e^{\Omega} + \frac{\partial \Omega}{\partial t_2} \frac{\partial \Omega}{\partial t_1} e^{\Omega} \right\}_{t_1=0, t_2=0} = \rho \sigma_1 \sigma_2.$$

Таким образом, константа, которую мы обозначаем  $\rho$  представляет собой ничто иное, как *коэффициент корреляции* (смотри ...). Из этого выражения следует важная теорема:

**Теорема.** *Если две нормальные случайные величины некоррелированы, то они независимы.*

## 30.5 Регрессия в нормальном распределении

Учитывая, что из

$$M_{X_1, X_2}(t_1, t_2) = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \exp\{t_1 x_1 + t_2 x_2\} f_{X_1, X_2}(x_1, x_2)$$

и

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} dx_2 f_{X_1, X_2}(x_1, x_2)$$

следует, что

$$M_{X_1, X_2}(t_1, 0) = \int_{-\infty}^{\infty} \exp\{t_1 x_1\} f_{X_1}(x_1) = M_{X_1}(t_1),$$

и полагая в (2)  $t_1 = 0$ , приходим к следующим эквивалентным утверждениям:

$$M_{X_1, X_2}(0, t_2) = \exp\{(\sigma_2 t_2)^2/2\}$$

и

$$f_2(x_2) = \int_{-\infty}^{\infty} dx_1 f(x_1, x_2) = \frac{\exp\{-x_2^2/2\sigma_2^2\}}{\sqrt{2\pi}\sigma_2}.$$

Говорят, что оба маргинальных распределения  $f_1(x)$  и  $f_2(x_2)$  центрированного двумерного нормального распределения  $f(x_1, x_2)$  также являются нормальными распределениями.

Найдём условную плотность вероятности:

$$f_{X_2}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2} \exp\left\{-\frac{[x_2 - \rho(\sigma_2/\sigma_1)x_1]^2}{2(1-\rho^2)\sigma_2^2}\right\}.$$

Таким образом, условное распределение является нормальным, но не центрированным со средним значением

$$\mu_{X_2|x_1} = \rho(\sigma_2/\sigma_1)x_1 \quad (1)$$



и дисперсией

$$\sigma_{X_2|x_1} = (1 - \rho^2)\sigma_2^2.$$

Чем ближе  $\rho$  к  $+1$  или к  $-1$ , тем меньше рассеяние  $Y|x$  около её условного  $\rho(\sigma_2/\sigma_1)x_1$  или  $-\rho(\sigma_2/\sigma_1)x_1$  соответственно. В предельных случаях:

$$X_2|_{x_1} = \pm(\sigma_2/\sigma_1)x_1, \quad \rho = \pm 1.$$

Все эти выводы остаются справедливыми и после замены  $X \leftrightarrow Y$ .

Вспомним, что мы работаем с центрированными случайными величинами. Для нецентрированных величин

$$\mu_{X_2|x_1} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1).$$

Это уравнение - *уравнение регрессии*.

## 31 Лекция 25. Регрессионный анализ

### 31.1 Модель линейной регрессии

Допустим, вы измеряете некоторую переменную  $y$  как функцию другой переменной  $x$  в заданных (неслучайных) точках  $x = x_1, x_2, \dots, x_n$  и наблюдаете нерегулярные отклонения измеренных значений  $y_1, y_2, \dots, y_n$  около линейной зависимости  $y = \alpha + \beta x$ . Вы можете предположить, что наблюдаемые значения представляют некоторые случайные величины  $Y_1, Y_2, \dots, Y_n$  и линейная зависимость имеет место для их математических ожиданий:

$$EY_j = \alpha + \beta x_j, \quad j = 1, 2, \dots, n.$$

Разность  $Y_j - (\alpha + \beta x_j)$  может рассматриваться, как случайный компонент этого процесса, или случайная ошибка  $\epsilon_j$

$$\epsilon_j = Y_j - (\alpha + \beta x_j).$$

Эта модель называется *моделью линейной регрессии*. Она позволяет сформулировать рассматриваемую проблему в терминах статистической зависимости *случайной величины*  $Y$  от *неслучайной величины*  $x$ . Заметим, что в корреляционном анализе обе величины  $Y$  и  $X$  являются случайными.

Основная идея - это оценка параметров  $\alpha$  и  $\beta$  с использованием соответствующей математической процедуры.

### 31.2 Метод наименьших квадратов

Наиболее простая оценка отклонения линейной функции  $y = \alpha + \beta x$  от системы точек  $(x_j, y_j)$  как целого - это *сумма квадратов ошибок* (*sum of squares of the errors*) ( $SSE$ ):

$$SSE \equiv \sum_{j=1}^n \epsilon_j^2 = \sum_{j=1}^n [Y_j - (\alpha + \beta x_j)]^2.$$

Рассматривая  $SSE$  как функцию двух переменных  $(\alpha, \beta)$  можно найти такую точку  $(a, b)$ , что  $SSE$  достигает в ней наименьшего значения. Координаты этой точки  $a$  и  $b$  подчиняются уравнениям:

$$\begin{aligned} \frac{\partial SSE}{\partial a} &= - \sum_{j=1}^n 2(Y_j - a - bx_j) = 0, \\ \frac{\partial SSE}{\partial b} &= - \sum_{j=1}^n 2(Y_j - a - bx_j)x_j = 0. \end{aligned}$$

Эта система, называемая системой *нормальных уравнений*, имеет решение:

$$\begin{aligned} a &= \bar{Y} - b\bar{x}, \\ b &= \frac{n \sum x_j Y_j - (\sum x_j)(\sum Y_j)}{n \sum x_j^2 - (\sum x_j)^2} = \frac{\overline{xY} - \bar{x} \cdot \bar{Y}}{\overline{x^2} - \bar{x}^2}, \end{aligned} \quad (1)$$

где

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j, \quad \overline{x^2} = \frac{1}{n} \sum_{j=1}^n x_j^2, \quad \overline{xY} = \frac{1}{n} \sum_{j=1}^n x_j Y_j.$$

### 31.3 Второе представление для $b$

**Теорема.** Коэффициент  $b$  может быть также представлен в виде

$$b = \frac{\sum(x_j - \bar{x})(y_j - \bar{y})}{\sum(x_j - \bar{x})^2}. \quad (1)$$

**Доказательство.** Покажем, что из (23.1.1) следует (22.2.1). Числитель (23.1.1):

$$\begin{aligned} \sum(x_j - \bar{x})(y_j - \bar{y}) &= \sum x_j y_j - \bar{x} \sum y_j - \bar{y} \sum x_j + \bar{x} \cdot \bar{y} \sum 1 = \\ &= n(\bar{x}\bar{y} - \bar{x} \cdot \bar{y} - \bar{x} \cdot \bar{y} + \bar{x} \cdot \bar{y}) = n(\bar{x}\bar{y} - \bar{x} \cdot \bar{y}). \end{aligned} \quad (2)$$

Знаменатель (23.1.1):

$$\sum(x_j - \bar{x})^2 = \sum x_j^2 - 2\bar{x} \sum x_j + \bar{x}^2 \sum 1 = n(\bar{x}^2 - 2\bar{x}^2 + \bar{x}^2) = n(\bar{x}^2 - \bar{x}^2). \quad (3)$$

Подставляя (2) и (3) в (1), получаем (22.2.1).

**Замечание.** В учебниках можно найти другие обозначения:

$$S_{xy} = \sum(x_j - \bar{x})(y_j - \bar{y}), \quad S_{xx} = \sum(x_j - \bar{x})^2, \quad S_{yy} = \sum(y_j - \bar{y})^2.$$

В этих обозначениях

$$b = S_{xy}/S_{xx}$$

и

$$\begin{aligned} SSE &= \sum[y_j - (a + bx_j)]^2 = (\text{substituting } a = \bar{y} - b\bar{x}) = \sum[y_j - \bar{y} + b\bar{x} - bx_j]^2 = \\ &= \sum[(y_j - \bar{y}) - b(x_j - \bar{x})]^2 = \sum(y_j - \bar{y})^2 - 2b \sum(x_j - \bar{x})(y_j - \bar{y}) + b^2 \sum(x_j - \bar{x})^2 = \\ &= S_{yy} - 2bS_{xy} + b^2S_{xx} = (\text{substituting } bS_{xx} = S_{xy}) = S_{yy} - 2bS_{xy} + bS_{xy} = S_{yy} - bS_{xy}. \end{aligned}$$

### 31.4 Доверительный интервал для $\beta$

**Теорема 1.** Несмещённой оценкой дисперсии  $\sigma_{Y_j}^2 \equiv \sigma^2$  является

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}.$$

Вспомним, что мы считаем  $b$  оценкой для  $\beta$ .

**Теорема 2.** А  $(1 - \alpha)100\%$ -confidence interval for the parameter  $\beta$  in the regression line  $y = \alpha + \beta x$  is

$$b - \frac{t_{\alpha/2}(n-2)s}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2}(n-2)s}{\sqrt{S_{xx}}}, \quad b = \frac{S_{xy}}{S_{xx}}.$$

### 31.5 Коэффициент корреляции

Чтобы получить выборочную оценку коэффициента корреляции  $\rho$  вернёмся к сумме квадратов ошибок  $SSE = S_{YY} - bS_{XY}$



### 31.6 Мультиномиальные коэффициенты

В некоторых ситуациях эксперимент может дать исходы, которые неразличимы от других исходов. Предположим, мы хотим знать, сколько существует различных расположений букв из слова *onion*. Поскольку здесь 5 букв, то существует  $5! = 120$  перестановок этих букв. Однако, некоторые выглядят одинаково. Например, нам интересно знать, сколько перестановок дают *oonni*. Обозначая буквы индексами  $o_1 n_1 i o_2 n_2$ , мы делаем их различными и видим, что

$$o_1 o_2 n_1 n_2 i, \quad o_2 o_1 n_1 n_2 i, \quad o_1 o_2 n_2 n_1 i, \quad o_2 o_1 n_2 n_1 i$$

– перестановки исходных 5 букв, которые дают одну и ту же буквенную последовательность. Действительно, при любом заданном положении двух  $o$  существует  $2!$  их возможных порядков размещения, которые выглядят идентично. Аналогично, существует  $2!$  одинаковых способов расположить буквы  $n$ . Таким образом, из  $5!$  возможных размещений букв каждое отличное от других размещение появляется  $2!2!=4$  раза. Это означает, что только  $5!/2!2! = 30$  расстановок отличаются. Эта идея обобщается следующим образом.

**Теорема.** Предположим, что набор  $n$  объектов составлен таким образом, что в нём содержится  $k$  различных объектов. Пусть  $n_i$  – число объектов типа  $i$ ,  $i = 1, 2, \dots, k$  в наборе, и  $n_1 + n_2 + \dots + n_k = n$ . Тогда существует

$$\frac{n!}{n_1! n_2! \dots n_k!}$$

различных перестановок  $n$  объектов. Это число называется *мультиномиальным коэффициентом* и обозначается

$$\binom{n}{n_1 \ n_2 \ \dots \ n_k} \equiv \frac{n!}{n_1! n_2! \dots n_k!}, \quad n_1 + n_2 + \dots + n_k = n.$$

Заметим, что для  $k = 1$  он равен 1, а для  $k = 2$  совпадает с биномиальным коэффициентом:

$$\binom{n}{n_1 \ n_2} = \frac{n!}{n_1! (n - n_1)!} \equiv \binom{n}{n_1}.$$

### 31.7 Мультиномиальное распределение

Как известно, биномиальные коэффициенты играют важную роль в формулировке биномиального распределения, возникающего в эксперименте, называемом испытаниями Бернулли. Самый простой способ вывода этого распределения основан на использовании бинома Ньютона:

$$\sum_{n_1=0}^n \binom{n}{n_1} a^{n_1} b^{n-n_1} = (a+b)^n.$$

Далее, при подстановке сюда  $a = p_1$  и  $b = 1 - p_1$  мы получаем

$$\sum_{n_1=0}^n \binom{n}{n_1} p_1^{n_1} (1 - p_1)^{n-n_1} = 1^n = 1.$$

Поскольку слагаемые в сумме неотрицательны и сумма равна 1, они могут рассматриваться как распределение вероятностей

$$P(X = n_1) = \binom{n}{n_1} p_1^{n_1} (1 - p_1)^{n-n_1}.$$

С использованием мультиномиальных коэффициентов результат может быть представлен более симметричным образом:

$$P(X_1 = n_1, X_2 = n_2) = \binom{n}{n_1 \ n_2} p_1^{n_1} p_2^{n_2}, \quad n_1 + n_2 + \dots + n_k, \quad p_1 + p_2 + \dots + p_k = 1.$$

с условием нормировки

$$\sum_{\{n_1+n_2=n\}} \binom{n}{n_1 \ n_2} p_1^{n_1} p_2^{n_2} = 1.$$

Обобщение этих формул на случай произвольного числа переменных,  $k$ , основан на мультиномиальной формуле

$$\sum_{n_1+\dots+n_k=1} \binom{n}{n_1 \ \dots \ n_k} a_1^{n_1} \dots a_k^{n_k} = (a_1 + \dots + a_k)^n,$$

тогда

$$P(X_1 = n_1, \dots, X_k = n_k) = \binom{n}{n_1 \ \dots \ n_k} p_1^{n_1} \dots p_k^{n_k}, \quad n_1 + \dots + n_k = 1, \quad p_1 + \dots + p_k = 1.$$

## 32 Лекция 26. Двухмерное нормальное распределение

### 32.1 Двухмерное нормальное распределение

Теперь выполним другой вид измерений, отличный от (22.1): измерим  $x_j$  и  $y_j$  и будем рассматривать их как реализацию двух случайных величин  $(X, Y)$ . Эта пара случайных величин описывается совместной плотностью  $f(x, y)$ . Будем рассматривать  $X$  и  $Y$  как центрированные нормально распределённые случайные величины с дисперсиями  $\sigma_1$  и  $\sigma_2$  соответственно. Если они независимы,

$$\begin{aligned} f(x, y) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{y^2}{2\sigma_2^2}\right) = \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x}{\sigma_1}\right)^2 + \left(\frac{y}{\sigma_2}\right)^2\right]\right\}. \end{aligned}$$

Легко видеть, что кривая постоянных значений  $f(x, y)$  представляет собой эллипс с главными осями, лежащими вдоль координатных осей  $x$  и  $y$ :

$$\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = \text{const.}$$

Если мы повернём эллипс вокруг его центра на произвольный угол, то его уравнение примет вид

$$\left(\frac{x}{\sigma_X}\right)^2 - 2\rho\left(\frac{x}{\sigma_X}\right)\left(\frac{y}{\sigma_Y}\right) + \left(\frac{y}{\sigma_Y}\right)^2 = \text{const},$$

где  $\sigma_X > 0$ ,  $\sigma_Y > 0$  и  $\rho \in [-1, 1]$  – новые постоянные, определяемые начальными константами  $\sigma_1, \sigma_2$  и углом поворота. Соответствующая плотность имеет вид

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x}{\sigma_X}\right)^2 - 2\rho\left(\frac{x}{\sigma_X}\right)\left(\frac{y}{\sigma_Y}\right) + \left(\frac{y}{\sigma_Y}\right)^2\right]\right\}, \quad (1)$$

где множители  $1/\sqrt{1-\rho^2}$  перед экспонентой и  $1/(1-\rho^2)$  в её аргументе обеспечивает нормировку:

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) = 1.$$

Уравнение (1) представляет собой двухмерное нормальное распределение специального – центрированного – вида. Наиболее общее представление имеет вид

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

где  $\mu_X$  и  $\mu_Y$  – математические ожидания величин  $X$  и  $Y$  соответственно:

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy x f(x, y) = \mu_X,$$

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy y f(x, y) = \mu_Y.$$

Вернёмся к центрированной двумерной плотности (1). Можно проверить следующие формулы:

$$\begin{aligned}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, x f(x, y) &= 0, \\ \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, y f(x, y) &= 0, \\ \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, x^2 f(x, y) &= \sigma_X^2, \\ \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, y^2 f(x, y) &= \sigma_Y^2, \\ \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \, xy f(x, y) &= \text{Cov}(X, Y) = \rho \sigma_X \sigma_Y.\end{aligned}$$

Таким образом, постоянная, обозначенная  $\rho$  в действительности ничто иное, как коэффициент корреляции (см ...). Производящая функция моментов двумерного распределения

$$M_{X,Y}(t_X, t_Y) \equiv \mathbb{E} \exp\{t_X X + t_Y Y\} = \exp\{[(\sigma_X t_X)^2 + 2\rho(\sigma_X t_X)(\sigma_Y t_Y) + (\sigma_Y t_Y)^2]/2\}. \quad (2)$$

Из этого выражения следует важная теорема.

**Теорема.** Если две нормальные случайные величины некоррелированы, то они независимы.

Учитывая, что из

$$M_{X_1, X_2}(t_1, t_2) = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \exp\{t_1 x_1 + t_2 x_2\} f_{X_1, X_2}(x_1, x_2)$$

и

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} dx_2 f_{X_1, X_2}(x_1, x_2)$$

следует, что

$$M_{X_1, X_2}(t_1, 0) = \int_{-\infty}^{\infty} \exp\{t_1 x_1\} f_{X_1}(x_1) = M_{X_1}(t_1),$$

и полагая в (2)  $t_Y = 0$ , приходим к следующим эквивалентным утверждениям:

$$M_{X,Y}(t_X, 0) = \exp\{(\sigma_X t_X)^2/2\}$$

и

$$\int_{-\infty}^{\infty} dy \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x}{\sigma_X}\right)^2 - 2\rho\left(\frac{x}{\sigma_X}\right)\left(\frac{y}{\sigma_Y}\right) + \left(\frac{y}{\sigma_Y}\right)^2\right]\right\} = \frac{\exp\{-x^2/2\sigma_X^2\}}{\sqrt{2\pi}\sigma_X}.$$

Говорят, что оба маргинальных распределения  $f_X(x)$  и  $f_Y(y)$  центрированного двумерного нормального распределения  $f_{X,Y}(x, y)$  также центрированные нормальные распределения.



Найдём условную плотность вероятности:

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_X} \exp \left\{ -\frac{[x - \rho(\sigma_X/\sigma_Y)y]^2}{2(1-\rho^2)\sigma_X^2} \right\}.$$

Таким образом, условное распределение является нормальным, но не центрированным, оно имеет среднее значение

$$\mu_{X|y} = \rho(\sigma_X/\sigma_Y)y$$

и дисперсию

$$\sigma_{X|y} = (1 - \rho^2)\sigma_X^2.$$

Чем ближе  $\rho$  к  $+1$  или к  $-1$ , тем меньше рассеяние величины  $X|y$  вокруг её условного среднего  $\rho(\sigma_X/\sigma_Y)y$  или  $-\rho(\sigma_X/\sigma_Y)y$  соответственно. В предельных случаях имеем:

$$X|y = \pm(\sigma_X/\sigma_Y)y, \quad \rho = \pm 1. \quad (1)$$

Все эти выводы остаются справедливыми после замены  $X \leftrightarrow Y$ .

Вспомним, что мы имеем дело с центрированными случайными величинами, которые могут быть обозначены  $\overset{\circ}{X}$  и  $\overset{\circ}{Y}$ , то есть уравнение (1) означает

$$\overset{\circ}{Y}|_x = \pm(\sigma_{\overset{\circ}{Y}}/\sigma_{\overset{\circ}{X}})x$$

## 33 Лекция 16. Выборочное распределение

### 33.1 Выборочное среднее

Предположим, что выборка имеет вид  $\{X_1, \dots, X_n\}$ . *Выборочное среднее*

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{j=1}^n X_j.$$

**Пример.** Бросая игральную кость, мы получили:  $\{4, 6, 1, 3, 1\}$ . Объём выборки (число элементов) равен 5. Выборочное среднее:

$$\bar{X} = \frac{4 + 6 + 1 + 3 + 1}{5} = 3.$$

### 33.2 Свойства выборочного среднего

**Свойство 1.** Если  $\{X'_1, \dots, X'_n\} = \{X_1 + a, \dots, X_n + a\}$ , тогда  $\bar{X}' = \bar{X} + a$ .

**Пример:** вам нужно вычислить выборочное среднее  $\bar{X}'$  для  $\{6, 8, 3, 5, 3\}$ , но вы замечаете, что  $X'_j = X_j + 2$ , где  $\{X_j\}$  - выборка, рассмотренная выше с  $\bar{X} = 3$ . Следовательно, можно записать  $\bar{X}' = \bar{X} + 2 = 3 + 2 = 5$  без дополнительных вычислений.

**Свойство 2.** Если  $\{X'_1, \dots, X'_n\} = \{aX_1, \dots, aX_n\}$ , тогда  $\bar{X}' = a\bar{X}$ .

**Пример:** нужно вычислить выборочное среднее  $\bar{X}'$  для  $\{12, 18, 3, 9, 3\}$ , но вы замечаете, что  $X'_j = 3X_j$ , где  $\{X_j\}$  - выборка, рассмотренная в 17.1 и  $\bar{X} = 3$ . Следовательно, можно записать  $\bar{X}' = 3\bar{X} = 3 \cdot 3 = 9$  без дополнительных вычислений.

### 33.3 Выборочная медиана и мода

Пусть все элементы  $\{X_1, \dots, X_n\}$  различны. Расположим их в возрастающем порядке:

$$\{X_1, \dots, X_n\} \Rightarrow \{X'_1, \dots, X'_n\}, \quad X'_1 < X'_2 < \dots < X'_n.$$

**Например,**

$$\{2.7; 2.2; 3.9; 1.9; 2.5\} \Rightarrow \{1.9; 2.2; 2.5; 2.7; 3.9\}$$

Если  $n$  нечётное (как здесь), то центральное число называется *выборочной медианой* и обозначается  $\tilde{X}$ :  $\tilde{X} = X'_{(n+1)/2}$ . **In the example**,  $\tilde{X} = 2.5$ .

Если  $n$  чётно, **например**, если

$$\{2.7; 2.2; 3.9; 1.9\} \Rightarrow \{1.9; 2.2; 2.5; 2.7\}$$

тогда выборочная медиана определяется как арифметическое среднее двух срединных чисел:  $\tilde{X} = [X'_{n/2} + X'_{(n+1)/2}]/2$ . **В примере**,  $\tilde{X} = 2.35$ .

Если выборка содержит элементы, некоторые из которых совпадают между собой, и количества совпадающих элементов различны, мы можем выбрать элемент, который встречается наиболее часто. Этот элемент называется *выборочной модой* и обозначается  $m$ . **Например**, для последней выборки  $\{2.7; 2.2; 3.9; 1.9\}$  мы не можем определить  $m$ , но для выборки  $\{4, 6, 1, 3, 1\}$  мы можем это сделать:  $m = 1$ .

### 33.4 Размах выборки, выборочная дисперсия и выборочное стандартное отклонение

Размах выборки  $R$  для  $\{X_1, \dots, X_n\}$  определяется как  $R = X_{\max} - X_{\min}$ . Например, для  $\{4, 6, 1, 3, 1\}$  имеем:  $R = 6 - 1 = 5$ .

Выборочная дисперсия определяется как

$$S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2, \quad \bar{X} = \frac{1}{n} \sum_{j=1}^n X_j.$$

Число  $n-1$  называется *числом степеней свободы*. В примере с бросанием игральной кости

$$\bar{X} = 3, \quad \sum_{j=1}^n (X_j - \bar{X})^2 = 18, \quad S^2 = 4\frac{1}{2}.$$

Выборочное стандартное отклонение определяется как  $S = \sqrt{S^2}$ . В нашем примере,  $S = \sqrt{4\frac{1}{2}} = \sqrt{2}$ .

### 33.5 Основные свойства выборочной дисперсии

**Свойство 1.** Если в выборке  $\{X_1, \dots, X_n\}$  выборочная дисперсия  $S^2$ , тогда для выборки  $\{X_1 + a, \dots, X_n + a\}$  будет такая же выборочная дисперсия  $S^2$ .

**Свойство 2.** Если выборка  $\{X_1, \dots, X_n\}$  обладает выборочной дисперсией  $S^2$ , тогда выборка  $\{aX_1, \dots, aX_n\}$  обладает выборочной дисперсией  $a^2 S^2$ .

**Свойство 3.** Существует более удобная формула для вычисления выборочных дисперсий:

$$S^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{j=1}^n X_j^2 - \bar{X}^2 \right).$$

Доказательство.

Для дальнейшего чтения: разделы 1.4; 1.5; 8.1, 8.2 из учебника.

Лекция 18. Некоторые важные статистические распределения

### 33.6 Нормальное распределение

Начнём с рассмотрения случайной выборки  $X_1, X_2, \dots, X_n$ , полученной из генеральной совокупности с произвольным распределением  $f(x)$  со средним  $\mu$  и дисперсией  $\sigma^2$ . Из центральной предельной теоремы сразу следует, что для большого объёма выборки  $n$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \stackrel{d}{\approx} N.$$

Другая важная для практики теорема связана с ней.

**Теорема.** Если независимые случайные выборки объёмов  $n_1$  и  $n_2$  получены из двух разных генеральных совокупностей с параметрами  $(\mu_1, \sigma_1^2)$  и  $(\mu_2, \sigma_2^2)$  соответственно, тогда нормированная выборочная разность

$$\frac{(\bar{X}_{n_1} - \bar{X}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \stackrel{d}{\approx} N.$$

Обычно, как показывает практика, если  $n$  около 30 или больше, чем 30, нормальная аппроксимация для  $\bar{X}_n$  очень хороша. Часто нормальное приближение используется и

для значений  $n$  меньших 30, за исключением случаев, когда генеральная совокупность определённо ненормальна.

Далее мы ограничимся рассмотрением только *нормальных генеральных совокупностей*.

### 33.7 Хи-квадрат распределение

Рассмотрим случайную величину  $\chi_1^2 = \overset{\circ}{N}$  :

$$F_{\chi_1^2}(x) = P(\overset{\circ}{N} < x) = P(-\sqrt{x} < \overset{\circ}{N} < \sqrt{x}) = 2P(0 < \overset{\circ}{N} < \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-y^2/2} dy.$$

Дифференцирование по  $x$  даёт:

$$f_{\chi_1^2}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}, \quad x > 0.$$

Это распределение – член специального подсемейства гамма-плотностей, называемых  $\chi^2$ -плотностями. Общий член подсемейства  $f_{\chi_n^2}(x)$  описывает сумму квадратов  $n$  независимых стандартных нормальных величин:

$$\chi_n^2 = \overset{\circ}{N}_1 + \overset{\circ}{N}_2 + \dots + \overset{\circ}{N}_n.$$

Число  $n$  называется числом *степеней свободы*.

$\chi_n^2$ -плотность
Функция $[2^{n/2}\Gamma(n/2)]^{-1}x^{n/2-1}e^{-x/2}$
Значения $x > 0$
Параметр $n = 1, 2, 3, \dots$
Среднее $n$
Дисперсия $2n$

### 33.8 $t$ -распределение

Предположим,  $\overset{\circ}{N}$  и  $\chi_n^2$  – независимые случайные величины. Их совместная плотность

$$f_{\overset{\circ}{N}, \chi_n^2}(x, y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, \quad -\infty < x < \infty, \quad y > 0.$$

Найдём плотность вероятности  $f_{T_n}(x)$  случайной величины

$$T_n = \frac{\overset{\circ}{N}}{\sqrt{\chi_n^2/n}}.$$

Накладывая условие на  $\chi_n^2$

$$f_{T_n}(x) = \int_0^\infty f_{T_n}(x|\chi_n^2 = y) f_{\chi_n^2}(y) dy$$

и используя соответствующее свойство плотности

$$f_{T_n}(x|\chi_n^2 = y) = f_{\frac{\overset{\circ}{N}}{\sqrt{y/n}}}(x) = \sqrt{y/n} f_{\overset{\circ}{N}}(\sqrt{y/n} x)$$

получаем:

$$\begin{aligned} f_{T_n}(x) &= \int_0^\infty \sqrt{y/n} f_N(\sqrt{y/n} x) f_{\chi_n^2}(y) dy = \\ &= \frac{1}{\sqrt{2\pi} 2^{1/2} \Gamma(n/2)} \int_0^\infty \exp\{-(1/2)[1 + z^2/n]y\} y^{(n-1)/2} dy. \end{aligned}$$

Используя замену переменных, приходим к результату – получаем  $t_n$ -плотность распределения или плотность распределения Стьюдента с  $n$  степенями свободы:

$$f_{T_n}(x) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n} \Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad -\infty < x < \infty.$$

### 33.9 $F$ -распределение

Рассмотрим отношение двух независимых хи-квадратов, умноженных на некоторый нормировочный множитель:

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}, \quad m, n = 1, 2, 3, \dots$$

Поступая также, как и ранее, получаем:

$$\begin{aligned} f_F(x) &= \int_0^\infty f_F(x|\chi_m^2 = y) f_{\chi_m^2}(y) dy = \int_0^\infty (ny/m) f_{\chi_n^2}((ny/m)x) f_{\chi_m^2}(y) dy = \\ &= \frac{(n/m)^{n/2} x^{n/2-1}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} \int_0^\infty y^{(m+n)/2-1} \exp\{(1/2)[1 + (n/m)x]y\} dy = \\ &= \frac{(n/m)^{n/2}}{\Gamma(m/2, n/2)} \frac{x^{n/2-1}}{[1 + (n/m)x]^{(m+n)/2}}, \quad x > 0. \end{aligned}$$

Это  $F_{n,m}$ -плотность вероятности или распределение Фишера с  $n$  и  $m$  степенями свободы ( $n$  называется числителем степеней свободы, а  $m$  – знаменателем степеней свободы).

### 33.10 Приложения статистических распределений

Пусть  $X_1, X_2, \dots, X_n$  случайная выборка из нормальной генеральной совокупности со средним  $\mu$  и дисперсией  $\sigma^2$ ,  $\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$  и  $S_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2$ . Знак  $\stackrel{d}{=}$  обозначает равенство распределений, как и раньше. Имеют место следующие взаимосвязи:

$$\begin{aligned} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} &\stackrel{d}{=} N; \\ \frac{\sum_{j=1}^n (X_j - \mu)^2}{\sigma^2} &\stackrel{d}{=} \chi_n^2; \\ \frac{\sum_{j=1}^n (X_j - \bar{X}_n)^2}{\sigma^2} &= \frac{(n-1)S_n^2}{\sigma^2} \stackrel{d}{=} \chi_{n-1}^2; \end{aligned}$$

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \stackrel{d}{=} T_{n-1}.$$

Пусть  $\bar{X}_{n_1}$  и  $\bar{X}_{n_2}$  – выборочные средние двух независимых случайных выборок объёма  $n_1$  и  $n_2$  полученных из двух различных генеральных совокупностей с параметрами  $(\mu_1, \sigma_1^2)$  и  $(\mu_2, \sigma_2^2)$  соответственно, тогда нормированная выборочная *разность*

$$\frac{(\bar{X}_{n_1} - \bar{X}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \stackrel{d}{=} \overset{\circ}{N}.$$

Пусть  $S_{n_1}^2$  и  $S_{n_2}^2$  – выборочные дисперсии независимых случайных выборок объёма  $n_1$  и  $n_2$ , полученных из разных нормальных генеральных совокупностей с дисперсиями  $\sigma_1^2$  и  $\sigma_2^2$  соответственно, тогда

$$\frac{S_{n_1}^2/\sigma_1^2}{S_{n_2}^2/\sigma_2^2} \stackrel{d}{=} F_{n_1-1, n_2-1}.$$

**Теорема.** Статистики  $\bar{X}_n$  и  $S_n^2$ , определённые на одной и той же выборке, полученной из нормальной генеральной совокупности, являются независимыми случайными величинами.

Лекция 18. Выборочные распределения

### 33.11 Выборочное распределение среднего значения нормальной генеральной совокупности

Генеральная совокупность с  $f(x) = f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/2\sigma^2)$  называется *нормальной генеральной совокупностью*, случайные величины  $X$  связаны со стандартной нормальной величиной  $Z \equiv \overset{\circ}{N}$  с помощью соотношения  $X = \mu + \sigma Z$ .

Легко доказать, что  $\sum_{j=1}^n NX_j$  вновь является нормальной величиной со средним  $n\mu$  и дисперсией  $n\sigma^2$ :

$$\sum_{j=1}^n X_j \stackrel{d}{=} \mu n + \sigma Z \sqrt{n}.$$

Отсюда следует, что

$$\bar{X}_n \equiv \frac{1}{n} \sum_{j=1}^n X_j \stackrel{d}{=} \mu + (\sigma/\sqrt{n})Z.$$

**Теорема.** Выборочное среднее нормальной генеральной совокупности со средним  $\mu$  и стандартным отклонением  $\sigma$  имеет нормальное распределение со средним  $\mu$  и стандартным отклонением  $\sigma/\sqrt{n}$ .

Замечание: это соотношение часто записывают в виде

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \stackrel{d}{=} Z.$$

## 34 Лекция 18. Выборочные распределения. [Учебник: Разделы 8.4; 8.5; 8.6; 8.7; 8.8]

### 34.1 Выборочные распределения среднего: $\sigma$ известно

Пусть  $X_1, X_2, \dots, X_n$  – выборка из нормальной генеральной совокупности с известным средним значением  $\mu$  и дисперсией  $\sigma^2$ , то есть

$$X_j \stackrel{d}{=} \mu + \sigma Z,$$

где  $Z = \overset{\circ}{N}$  – стандартная нормальная величина (см. 13.2). Сумма  $\sum_{j=1}^n X_j$  снова является нормальной величиной со средним  $n\mu$  и дисперсией  $n\sigma^2$ :

$$\sum_{j=1}^n X_j \stackrel{d}{=} n\mu + \sqrt{n}\sigma Z.$$

Таким образом, выборочное среднее

$$\bar{X} = (1/n) \sum_{j=1}^n X_j \stackrel{d}{=} \mu + \sigma/\sqrt{n} Z$$

и

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{d}{=} Z.$$

Это означает, что

$$P(x_1 < \bar{X} < x_2) = P\left(\frac{x_1 - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{x_2 - \mu}{\sigma/\sqrt{n}}\right) = P(z_1 < Z < z_2), \quad z_{1,2} = \frac{x_{1,2} - \mu}{\sigma/\sqrt{n}}.$$

Вероятность  $P(z_1 < Z < z_2)$  представляется как разность

$$P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1),$$

где функция  $F_Z(z) \equiv P(Z < z)$  представляет *площадь под нормальной кривой* и может быть найдена из таблицы Table A.3 (pages 670-671 of the textbook).

**Например,**

$$P(-0.28 < Z < 1.22) = P(Z < 1.22) - P(Z < -0.28) = 0.8888 - 0.3897 = 0.4991.$$

## 34.2 Выборочное распределение величины $S^2$

Сейчас мы рассмотрим распределение выборочной дисперсии для нормальной генеральной совокупности

$$S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2.$$

Перепишем её в виде:

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{j=1}^n \frac{(X_j - \bar{X})^2}{\sigma^2}.$$

Можно доказать, что

$$\sum_{j=1}^n \frac{(X_j - \bar{X})^2}{\sigma^2} \stackrel{d}{=} \sum_{j=1}^{n-1} Z_j^2.$$

Сумма из  $\nu$  квадратов независимых стандартных нормальных величин обозначается через  $\chi^2(\nu)$  и её распределение называется *хи-квадрат распределением с  $\nu$  степенями свободы*. Таким образом

$$\frac{(n-1)S^2}{\sigma^2} \stackrel{d}{=} \chi^2(n-1).$$

Это распределение концентрируется на положительной полуоси. Величины  $\chi_\alpha^2(\nu)$ , подчиняющиеся условию

$$P(\chi^2(\nu) > \chi_\alpha^2(\nu)) = \alpha$$

называются *критическими значениями хи-квадрат распределения*. Эти значения представлены в таблице Table A.5 (see pages 674-675). Используя эту таблицу, можно найти такое значение величины  $s_\alpha^2$ , что  $S^2$  может превысить его с заданной вероятностью  $\alpha$ :

$$P(S^2 > s_\alpha^2) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)s_\alpha^2}{\sigma^2}\right) = P(\chi^2(n-1) > \chi_\alpha^2(n-1)),$$

где  $\chi_\alpha^2(n-1) = (n-1)s_\alpha^2/\sigma^2$ .

**Пример:** если  $\sigma = 1$ ,  $\alpha = 0.01$  и  $n = 21$ , тогда  $\chi_{0.01}^2(20) = 37.566$  и следовательно  $s_{0.01}^2 = \chi_\alpha^2(n-1)\sigma^2/(n-1) = 37.566/20 \approx 1.88$ .

### 34.3 Выборочное распределение среднего: $\sigma$ неизвестно

В этом случае мы должны использовать  $S$  вместо  $\sigma$  и принять во внимание, что полученная случайная величина подчиняется  $t$ -распределению (см. таблицу Table A4, pages 672-673):

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \stackrel{d}{=} T(n-1).$$

Вспомним, что  $t$ -распределение зависит от числа степеней свободы  $\nu = n-1$  и для больших  $n$  оно становится близким к стандартному нормальному распределению.

### 34.4 Разница между средними

Если вы оцениваете разность  $\mu_1 - \mu_2$  между средними двух генеральных совокупностей объёмов  $n_1$  и  $n_2$  с известными дисперсиями  $\sigma_1$  и  $\sigma_2$ , вы должны учесть следующую формулу:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \stackrel{d}{=} Z.$$

### 34.5 Отношение двух дисперсий

Если вы оцениваете отношение дисперсий двух генеральных совокупностей объёмов  $n_1$  и  $n_2$ , вы используете следующую формулу:

$$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \stackrel{d}{=} F_{n_1-1, n_2-1}.$$

### 34.6 Центральная предельная теорема

Распределения, которые мы рассматривали выше были получены при условии, что генеральная совокупность нормальна, то есть каждая из величин  $X_1, X_2, X_3, \dots$  имеет плотность

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

В этом случае выборочная средняя статистика является *строго нормальной для любого объёма выборки  $n$* :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{d}{=} N, \quad n = 1, 2, 3, \dots \quad (1)$$

Тем не менее мы можем использовать вышеприведённые формулы, даже если генеральная совокупность не является нормальной. Это возможно в случае, когда  $f(x)$  имеет конечную дисперсию и  $n$  достаточно велико (обычно,  $n > 30$ ).



**Теорема.** Если  $X_1, X_2, \dots, X_n$  – независимые копии случайной величины  $X$  с конечной дисперсией  $\sigma^2$  и средним  $\mu$  и  $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ , тогда

$$F_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(x) \rightarrow F_{\overset{\circ}{N}}(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

as  $n \rightarrow \infty$ . Этот факт иногда записывается следующим образом:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \overset{\circ}{N}, \quad n \rightarrow \infty \quad (2)$$

(сравните с (1)).

**Доказательство.** Мы должны доказать, что  $M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) \rightarrow e^{t^2/2}$ . Представляя эту случайную величину в виде суммы

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n \overset{\circ}{X}_j,$$

где  $\overset{\circ}{X}_j = X_j - \mu$  – центрированные копии величин  $X_j$ , получаем:

$$M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) = M_{\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n \overset{\circ}{X}_j}(t) = M_{\sum_{j=1}^n \overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right) = \left[M_{\overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right)\right]^n.$$

При  $n \rightarrow \infty$

$$M_{\overset{\circ}{X}_j}\left(\frac{1}{\sigma\sqrt{n}}t\right) \sim 1 + \frac{\mu'_2 t^2}{2\sigma^2 n}.$$

Поскольку  $\mu = 0$ ,  $\mu'_2 = \sigma^2$ . Используя асимптотическое соотношение  $1 + \varepsilon \sim e^\varepsilon$ ,  $\varepsilon \rightarrow 0$ , получаем желаемый результат:

$$M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) \rightarrow e^{t^2/2}, \quad n \rightarrow \infty.$$

## 34.7 Применение ЦПТ к аппроксимации распределений

Из ЦПТ следует, что при выполнении её условий

$$\sum_{j=1}^n X_j \xrightarrow{d} n\mu + \sigma\sqrt{n} \overset{\circ}{N}, \quad n \rightarrow \infty.$$

Это означает, что если некоторая функция  $f(x; n)$  может быть интерпретирована как плотность вероятности распределения суммы  $n$  независимых одинаково распределённых случайных величин со средним  $\mu$  и дисперсией  $\sigma$ , тогда

$$f(x; n) \approx f_{n\mu + \sigma\sqrt{n} \overset{\circ}{N}}(x) = \frac{1}{\sqrt{2\pi n\sigma}} e^{-\frac{(x-n\mu)^2}{2n\sigma^2}}$$

при достаточно большом числе  $n$ . Примеры: гамма-распределение

$$\frac{1}{\Gamma(n)\mu} \left(\frac{x}{\mu}\right)^{n-1} e^{-x/\mu} \approx \frac{1}{\sqrt{2\pi n\mu}} e^{-\frac{(x-n\mu)^2}{2n\mu^2}},$$

$\chi^2$ -плотность

$$f_{\chi^2(n)}(x) \approx \frac{1}{\sqrt{8\pi n}} e^{-\frac{(x-2n)^2}{8n}}.$$

Примеры.

## 34.8 Многомерное нормальное распределение

## 34.9 Симметричные многомерные распределения

## 34.10 Изотропные многомерные распределения

Ниже мы воспользуемся следующей терминологией. Функция  $f(x)$ ,  $x \in \mathbb{R}^N$ , называется *радиальной функцией*, если она зависит только от расстояния  $|x|$ . Случайный вектор  $X \in \mathbb{R}^N$  и его плотность вероятности  $p_N(x)$  будем называть *сферически симметричными*, если  $p_N(x)$  является радиальной функцией, т.е.

$$p_N(x) = \rho_N(|x|), \quad (4.1)$$

где  $\rho_N(r)$  – функция, заданная на полуоси  $r \geq 0$ . Характеристическая функция  $f_N(k)$  сферически симметричного вектора  $X \in \mathbb{R}^N$  является также радиальной функцией,

$$f_N(k) = \phi_N(|k|). \quad (4.2)$$

Очевидно, эти функции удовлетворяют соотношениям

$$\int_{\mathbb{R}^N} \rho_N(|x|) dx = \frac{2\pi^{N/2}}{\Gamma(N/2)} \int_0^\infty \rho_N(r) r^{N-1} dr = 1$$

и

$$\phi_N(0) = 1.$$

Из формул преобразования Фурье

$$f_N(k) = \int_{\mathbb{R}^N} e^{i(k,x)} p_N(x) dx,$$
$$p_N(x) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} e^{-i(k,x)} f_N(k) dk$$

вытекают следующие соотношения для соответствующих радиальных функций [?, ?]:

$$\phi_N(t) = (2\pi)^{N/2} t^{1-N/2} \int_0^\infty \rho_N(r) J_{N/2-1}(tr) r^{N/2} dr \quad (4.3)$$

и

$$\rho_N(r) = (2\pi)^{-N/2} r^{1-N/2} \int_0^\infty \phi_N(t) J_{N/2-1}(rt) t^{N/2} dt. \quad (4.4)$$

Рассмотрим последовательность распределений  $p_1(x), p_2(x_1, x_2), p_3(x_1, x_2, x_3), \dots$  имеющих общую характеристическую радиальную функцию  $\phi(t)$  и таким образом являющуюся сферически симметричной.

**Теорема 1.** Пусть  $X^N = \{X_1, \dots, X_N\}$  – случайный  $N$ -мерный вектор со сферически симметричной плотностью

$$p_N(x) = \rho_N(|x|), \quad x \in \mathbb{R}^N.$$

Тогда его проекция на  $n$ -мерное подпространство  $\{X_1, \dots, X_n\}$  имеет плотность

$$p_n(x) = \rho_n(x), \quad x \in \mathbb{R}^n,$$

с той же самой характеристической радиальной функцией  $\phi(t)$ .

Доказательство. Это довольно очевидно, что характеристическая функция

$$\begin{aligned} f_n(k_1, \dots, k_n) &= f_N(k_1, \dots, k_n, 0, \dots, 0) = \\ &= \phi(\sqrt{k_1^2 + \dots + k_n^2}), \end{aligned}$$

что и подразумевает утверждение теоремы.

Теперь, пусть  $\rho_1(r), \rho_2(r), \rho_3(r), \dots$  – радиальные функции сферически симметричных плотностей и  $\theta_i(s) = \rho_i(\sqrt{s})$ ,  $i = 1, 2, 3, \dots$

Теорема 2. Следующие соотношения выполняются для любых  $N > 1$ :

$$\theta_N(s) = \frac{1}{\sqrt{\pi}} \left( D_-^{1/2} \theta_{N-1} \right) (s) = \pi^{(1-N)/2} \left( D_-^{(N-1)/2} \theta_1 \right) (s), \quad (4.5)$$

где  $D_-^\nu$  – дробная производная (см. А.11) В частности, для любых  $N > 2$

$$\theta_N(s) = -\frac{1}{\pi} \theta'_{N-2}(s), \quad (4.5)$$

или

$$\rho_N(r) = -\frac{1}{2\pi r} \rho'_{N-2}(r). \quad (4.6)$$

Доказательство. Поскольку характеристическая функция  $f_N(k)$ ,  $k \in R^N$  зависит только от  $|k|$ , достаточно рассмотреть

$$f_N(k_1, 0, \dots, 0) = \phi(t), \quad t = |k_1| \geq 0 :$$

$$\begin{aligned} \phi(t) &= \int_{-\infty}^{\infty} dx_1 e^{itx_1} \int_{R^{N-1}} \rho_N(|x|) dx_2 \dots dx_n = \\ &= \int_{-\infty}^{\infty} dx_1 e^{itx_1} \int_{R^{N-1}} \theta_N(x^2) dx_2 \dots dx_n. \end{aligned}$$

Переходя к полярным координатам во внутреннем интеграле, имеем

$$\begin{aligned} \phi(t) &= \int_{-\infty}^{\infty} dx_1 e^{itx_1} \Omega_{N-1} \int_0^{\infty} \theta_N(r^2 + x_1^2) r^{N-2} dr = \\ &= \int_{-\infty}^{\infty} dx_1 e^{itx_1} (\Omega_{N-1}/2) \int_{x_1^2}^{\infty} \theta_N(\sigma) (\sigma - x_1^2)^{(N-3)/2} d\sigma, \quad \Omega_N = 2\pi^{N/2}/\Gamma(N/2). \end{aligned}$$

Это означает, что

$$\theta_1(s) = \frac{\pi^{(N-1)/2}}{\Gamma((N-1)/2)} \int_s^{\infty} \frac{\theta_N(\sigma) d\sigma}{(\sigma - s)^{1-(N-1)/2}} = \pi^{(N-1)/2} (I_-^{(N-1)/2} \theta_N)(s)$$

и после обратной подстановки приходим к (4.5).

В качестве примера применения уравнения (4.5) мы выведем многомерную плотность Коши из одномерной плотности:

$$\rho_1(\sqrt{s}) = \theta_1(s) = \frac{1}{\pi(1+s)}.$$

Согласно (4.5) и (A.11)

$$\theta_N(s) = \pi^{-(N+1)/2} \frac{(-1)^n}{\Gamma(n - (N-1)/2)} \frac{d^n I(s)}{ds^n},$$

где  $n = [(N+1)/2]$  – целая часть  $(N+1)/2$  и

$$I(s) = \int_s^\infty \frac{d\sigma}{(1+\sigma)(\sigma-s)^{(N+1)/2-n}\mu} (s+1)^{-\mu} \int_0^\infty \frac{dx}{x^\mu(1+x)} = \frac{\pi}{(s+1)^\mu \sin(\pi\mu)}$$

с

$$\mu = (N+1)/2 - n < 1.$$

Так как

$$\frac{d^n}{ds^n} (s+1)^{-\mu} = (-1)^n \mu(\mu+1) \dots (\mu+n-1) (s+1)^{-\mu-n},$$

окончательно имеем

$$\theta_N(s) = \frac{\Gamma((N+1)/2)}{[\pi(1+s)]^{(N+1)/2}}$$

и соответственно

$$q_N(x; 1) = \frac{\Gamma((N+1)/2)}{[\pi(1+x^2)]^{(N+1)/2}}$$

для всех  $N$ .

### 34.11 Многомерные устойчивые распределения

Чтобы найти характеристические функции многомерных сферически симметричных устойчивых законов, воспользуемся  $N$ -мерным аналогом симметричного распределения Ципфа-Парето:

$$\mathbf{P}(|X| > r) = \begin{cases} Ar^{-\alpha}, & r > \epsilon = A^{1/\alpha}, \\ 1, & r < \epsilon. \end{cases}$$

Радиальная функция плотности такого распределения имеет вид

$$\rho_N(r) = -\frac{1}{S_{N-1}} \frac{d\mathbf{P}(|X| > r)}{dr} = \frac{\alpha A \Gamma(N/2)}{2\pi^{N/2}} r^{-\alpha-N}, \quad r > \epsilon,$$

а радиальная характеристическая функция

$$\phi_N(t) = 2^{N/2-1} \alpha A \Gamma(N/2) t^\alpha \int_{\epsilon t}^\infty \tau^{-\alpha-N/2} J_{N/2-1}(\tau) d\tau.$$

Выполняя интегрирование по частям с учетом соотношения

$$\frac{d}{d\tau} [\tau^{-(N/2-1)} J_{N/2-1}(\tau)] = -\tau^{-N/2+1} J_{N/2}(\tau),$$

получим

$$\begin{aligned} \phi_N(t) = 2^{N/2-1} \alpha A \Gamma(N/2) t^\alpha \{ (r_0 t)^{-N/2-\alpha+1} J_{N/2-1}(r_0 t) - \\ - \int_{\epsilon t}^\infty \tau^{-N/2-\alpha+1} J_{N/2}(\tau) d\tau \}. \end{aligned}$$

При  $t \rightarrow 0$

$$(\dots) \sim (\epsilon t)^{-N/2-\alpha+1} [(\epsilon t/2)^{N/2-1} / \Gamma(N/2) - \dots]$$

$$-\int_0^{\infty} \tau^{-N/2-\alpha+1} J_{N/2}(\tau) d\tau$$

и в результате имеем

$$1 - \phi_N(t) \sim \frac{A\Gamma(N/2)\Gamma(1-\alpha/2)}{\Gamma((N+\alpha)/2)} (t/2)^\alpha.$$

Пусть теперь

$$Z_n = (X_1 + \dots + X_n)/b_n. \quad (5.1)$$

и  $f_N(k)$  – характеристическая функция нормированной векторной суммы  $Z_n$  независимых слагаемых  $X_i$ :

$$\begin{aligned} f_N^{(n)}(k) &= \mathbb{E} \exp\{ik \sum_{j=1}^n X_j/b_n\} = \\ &= \phi_N^n(|k/b_n|). \end{aligned}$$

При  $n \rightarrow \infty$

$$\phi_N^n(k/b_n) \sim (1 - \frac{A\Gamma(N/2)\Gamma(1-\alpha/2)}{\Gamma((N+\alpha)/2)} |k/(2b_n)|^\alpha)^n.$$

Полагая

$$b_n = b_1 n^{1/\alpha}. \quad (5.2)$$

находим, что

$$f_N^{(n)}(k) \rightarrow g_N(k; \alpha) = e^{-|k|^\alpha}, \quad n \rightarrow \infty. \quad (5.3)$$

при

$$b_1 = \frac{1}{2} \left[ \frac{A\Gamma(N/2)\Gamma(1-\alpha/2)}{\Gamma((N+\alpha)/2)} \right]^{1/\alpha}. \quad (5.4)$$

Заметим, что в одномерном случае коэффициенты (5.2), (5.3) совпадают с приведенными в Табл. 2.1 (необходимо учесть, что  $A = 2c$ ).

Таким образом,  $N$ -мерные плотности сферически симметричных устойчивых распределений записываются в виде

$$q_N(x; \alpha) = \frac{1}{(2\pi)^N} \int_{\mathbb{R}^N} e^{-i(k,x)-|k|^\alpha} dk,$$

а соответствующие радиальные функции, согласно (4.4) имеют вид

$$\rho_N(r; \alpha) = (2\pi)^{-N/2} r^{1-N/2} \int_0^{\infty} e^{-t^\alpha} J_{N/2-1}(rt) t^{N/2} dt. \quad (5.5)$$

Разлагая в ряд экспоненту или функцию Бесселя, получим два разложения для радиальных функций сферически симметричных устойчивых плотностей

$$\rho_N(r; \alpha) = \frac{1}{\pi(r\sqrt{\pi})^N} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n!} \Gamma((n\alpha + N)/2) \Gamma(n\alpha/2 + 1) \sin(\alpha n\pi/2) (r/2)^{-n\alpha} \quad (5.6)$$

и

$$\rho_N(r; \alpha) = \frac{2}{\alpha(2\sqrt{\pi})^N} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{\Gamma((2n + N)/\alpha)}{\Gamma(n + N/2)} (r/2)^{2n}. \quad (5.7)$$

Пользуясь свойствами гамма-функции легко показать, что при  $N = 1$  найденные разложения (5.6) и (5.7) переходят в разложения симметричных одномерных распределений

(4.1.4) и (4.1.3) соответственно. Как и в одномерном случае ряд (5.6) сходится при  $\alpha < 1$  и является асимптотическим при  $\alpha \geq 1$ , а ряд (5.7) – наоборот, сходится при  $\alpha \geq 1$  и является асимптотическим, если  $\alpha < 1$ .

Умножая (5.5) на  $r^s dr$  и интегрируя по полуоси нетрудно найти трансформанту Меллина радиальной функции

$$\bar{\rho}_N(s; \alpha) = \frac{2^{1+s}}{\alpha(4\pi)^{N/2}} \frac{\Gamma((N-s-1)/\alpha)\Gamma((1+s)/2)}{\Gamma((N-s-1)/2)} \quad (5.8)$$

Используя это выражение или разложения (5.6) – (5.7), можно выразить радиальную функцию через функцию Фокса (см. А.9). Формула (5.8) позволяет записать в явном виде абсолютный момент случайного вектора  $Y(\alpha)$  с устойчивым сферически симметричным распределением:

$$\mathbb{E}|Y(\alpha)|^s = \Omega_N \bar{\rho}_N(s + N - 1; \alpha) = 2^s \frac{\Gamma(1 - s/\alpha)\Gamma((s + N)/2)}{\Gamma(1 - s/2)\Gamma(N/2)}. \quad (5.9)$$

Полученное выражение может рассматриваться как аналитическая функция в плоскости  $s$ , исключая такие точки, как  $s = k\alpha$  и  $s = -N - k + 1$  ( $k = 1, 2, \dots$ ), в которых мы имеем простые полюсы функции. Отсюда, в частности, следует, что  $\mathbb{E}|Y|^s$  допускает разложение в ряд Тейлора по степеням  $s$ , в круге  $|s| < \min(N, \alpha)$ .

## 34.12 Примеры

Линии постоянных плотностей и диаграммы рассеяния для  $\rho = 0$ ,  $0 < \rho < 1$ ,  $\rho = 1$ ,  $-1 < \rho < 0$ ,  $\rho = -1$ .

## 34.13 Выборочная оценка коэффициента корреляции

Вспомним, что

$$SSE = S_{yy} - bS_{xy}.$$

Деля обе части на  $S_{yy}$  и заменяя  $S_{xy}$  на  $bS_{xx}$  приходим к соотношению

$$\frac{b^2 S_{xx}}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}.$$

Последнее слагаемое  $SSE/S_{yy} \geq 0$ , таким образом

$$\frac{b^2 S_{xx}}{S_{yy}} \leq 1$$

и

$$-1 \leq b\sqrt{S_{xx}/S_{yy}} \leq 1.$$

Величина  $b\sqrt{S_{xx}/S_{yy}}$ , обозначенная  $r$  – *выборочный коэффициент корреляции*:

$$r = b \frac{S_{xx}}{S_{yy}} = \frac{S_{xy}}{S_{xx}} \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum (x_j - \bar{x})^2 \sum (y_j - \bar{y})^2}}.$$

### 34.14 Тест на корреляцию

Коэффициент корреляции представляет собой силу линейных отношений между двумя величинами: если такая связь отсутствует, то  $\rho = 0$ , если они сильно линейно связаны, то  $\rho = \pm 1$ . Таким образом, цель простейшего теста – выяснить, являются ли эти величины независимыми или нет.

**Проверка** гипотезы  $\rho = 0$  с альтернативой  $\rho \neq 0$  выглядит следующим образом.

- 1) Формулируем  $H_0 : \rho = 0$ ,  $H_1 : \rho \neq 0$ .
- 2) Выбираем  $\alpha$ .
- 3) Находим критические значения  $\pm t_{\alpha/2}$ .
- 4) Вычисляем статистику  $t = \frac{b}{s/\sqrt{S_{xx}}}$ , где  $b = \frac{S_{xy}}{S_{xx}}$ ,  $s = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$ .
- 5) Заключение: если  $t < -t_{\alpha/2}$  или  $t > t_{\alpha/2}$ , то  $H_0$  отвергается.

**Замечание.** Статистика может быть представлена в виде:  $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

### 34.15 Пример (стр. 394).

- 1)  $H_0 : \rho = 0$ ,  $H_1 \neq 0$ .
- 2)  $\alpha = 0.05$
- 3)  $n = 29$ ,  $t_{0.025}(27) = 2.052$  (Table A4, page 672.)
- 4)  $t = \frac{0.943\sqrt{27}}{1-(0.9435)^2} = 14.79$
- 5) Заключение: поскольку  $14.79 > 2.052$ , нуль-гипотеза  $H_0$  отвергается.

**CORZINA**

## 34.16 Взаимосвязь между нормальными и стандартными нормальными случайными величинами

Пусть  $N$  и  $\overset{\circ}{N}(x)$  распределены с плотностью  $n(x; \mu, \sigma)$  и  $\overset{\circ}{n}(x)$  соответственно. Тогда

$$F_N(x) = P(N < x) = \int_{-\infty}^x n(x'; \mu, \sigma) dx' = \dots = \int_{-\infty}^{(x-\mu)/\sigma} \overset{\circ}{n}(x') dx' = F_{\mu+\sigma\overset{\circ}{N}}(x).$$

Следовательно, случайные величины  $N$  и  $\mu + \sigma \overset{\circ}{N}$  одинаково распределены:

$$N \stackrel{d}{=} \mu + \sigma \overset{\circ}{N}.$$

Для дальнейшего чтения: sections 6.2, from the main textbook.

## 35 Лекция 14. Преобразования случайных величин

### 35.1 Теорема преобразования

**Теорема 1.** Пусть  $X$  – случайная величина с непрерывной кумулятивной функцией распределения  $F_X(x)$  и  $Y$  – другая величина, связанная с  $X$  через монотонно возрастающую функцию  $g: Y = g(X)$ . Тогда  $F_X(x) = F_Y(g(x))$ .

**Доказательство.**

**Следствие.** При условиях предыдущей теоремы

$$f_X(x) = f_Y(g(x))g'(x).$$

**Замечание.** Если  $g(x)$  – монотонно убывающая функция, тогда

$$f_X(x) = f_Y(g(x))|g'(x)|.$$

### 35.2 Преобразование стандартных равномерных случайных величин

**Теорема.** Пусть  $U$  – стандартная равномерная (то есть равномерно распределённая на  $(0,1)$ ) случайная величина и  $y=F(x)$  – некоторая кумулятивная функция распределения с обратной функцией  $F^{-1}(y)$ , тогда случайная величина  $X = F^{-1}(U)$  распределена согласно кумулятивной функции  $F(x)$ .

**Доказательство.**

### 35.3 Пример: преобразование стандартной равномерной с.в. в экспоненциальную с.в.

Дана плотность:

$$f_X(x) = 1/\mu e^{-x/\mu}, x > 0.$$

Первый шаг – найти кумулятивную функцию:  $F_X(x) = \int_0^x f_X(x') dx' = 1 - e^{-x/\mu}$ .

Второй шаг – находим обратную функцию:  $F_X^{-1}(y) = -\mu \ln(1 - y)$ .

Третий шаг – записать алгоритм:  $X = -\mu \ln(1 - U) \stackrel{d}{=} -\mu \ln U$ .

**Проверка результата.**



### 35.4 Пример: преобразование экспоненциальной с.в. в гамма-распределённую с.в.

Для  $f_Y(y) = \frac{1}{(n-1)!\mu} \left(\frac{x}{\mu}\right)^{n-1} e^{-x/\mu}$ .

$$Y = X_1 + X_2 + \dots + X_n,$$

где  $X_j$  – независимые экспоненциально распределённые случайные величины со средним  $\mu$ .

### 35.5 Плотность Вейбулла

Применяя следствие 4.1 к экспоненциальной плотности  $f_Y(y) = \alpha e^{-\alpha y}$  и к степенной функции  $g(x) = x^\beta$ ,  $\alpha > 0$ ,  $\beta > 0$  мы получаем *Плотность Вейбулла*.

Плотность Вейбулла
Функция $f(x) = \alpha\beta e^{-\alpha x^\beta} x^{\beta-1}$
Значения $x > 0$
Параметры $\alpha > 0; \beta > 0$
Среднее значение $\alpha^{-1/\beta} \Gamma(1 + \frac{1}{\beta})$
Дисперсия $\alpha^{-2/\beta} \left\{ \Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2 \right\}$

### 35.6 Логнормальная плотность

Применяя следствие 4.1 к нормальной плотности  $n(x; \mu, \sigma)$  и функции  $g(x) = \ln x$ , мы получим *логнормальную плотность*.

Логнормальная плотность
Функция $f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}$
Значения $x > 0$
Параметры $-\infty < \mu < \infty; \sigma > 0$
Среднее значение $\exp(\mu + \sigma^2/2)$
Дисперсия $\exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$

**For further reading :** sections 7.1, 7.2, from the main textbook.

## 36 Лекция 15. Производящая функция моментов

### 36.1 Производящая функция моментов

**Определение.** Математическое ожидание функции  $e^{tX}$  действительной переменной  $t$  называется *производящей функцией моментов случайной величины  $X$  (ПФМ)* и обозначается  $M_X(t)$ :

$$M_X(t) = Ee^{tX}.$$

Если продифференцировать эту функцию  $n$  раз в точке  $t = 0$ , то мы получим  $n$ -й момент случайной величины  $X$ :

$$M_X^{(n)}(0) \equiv \frac{d^n Ee^{tX}}{dt^n} \Big|_{t=0} = E \frac{d^n e^{tX}}{dt^n} \Big|_{t=0} = EX^n = \mu'_n.$$

Совместная ПФМ случайных величин  $X_1, X_2, \dots$  определяется как

$$M_{X_1, X_2, \dots}(t) = Ee^{(t_1 X_1 + t_2 X_2 + \dots)}.$$

### 36.2 Пример: $M_X(t)$ для распределения Пуассона

$$M_X(t) = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} e^{tx} = e^{\mu(e^t-1)}, \quad M'_X(0) = \mu, \quad M''_X(0) = \mu + \mu^2, \quad \sigma_X^2 = \mu.$$

### 36.3 Пример: $M_X(t)$ для нормального распределения

### 36.4 Основные свойства ПФМ

1)  $M_X(0) = 1$  для любой случайной величины  $X$ .

2)  $M_c(t) = e^{tc}$  для любой неслучайной величины  $c$ .

3)  $M_X^{(n)}(0) = \mu'_n$ ,  $n = 1, 2, 3, \dots$  если момент существует.

4)  $M_X(t) = 1 + \mu t + \frac{\mu'_2 t^2}{2} + \dots$ , если моменты существуют.

5)  $M_{a+X} = e^{at} M_X(t)$  для любой неслучайной  $a$ .

6)  $M_X(bX(t)) = M_X(bt)$  для любой неслучайной  $b$ .

7)  $M_{a+bX}(T) = e^{aT} M_X(bT)$  для любых неслучайных  $a$  и  $b$ .

8)  $M_{X+Y}(t) = M_X(t) M_Y(t)$  для любых независимых  $X$  и  $Y$ ,  $M_{X_1+\dots+X_n}(t) = [M_X(t)]^n$  для любых независимых копий  $X$ .

9) При определённых условиях существует взаимно однозначное соответствие между ПФМ и распределениями.

## 36.5 Дискретные распределения

Название Параметры	$f_X(x)$ Значения	Среднее $\mu$	Дисперсия $\sigma^2$	ПФМ $M_X(t)$
Равномерное $n = 2, 3, \dots$	$\frac{1}{n}$ $1, 2, \dots, n$	$(n+1)/2$	$(n^2-1)/12$	$(1/n) \sum_{j=1}^n e^{jt}$
Бернулли $0 < p < 1$	$p^x(1-p)^{1-x}$ $x = 0, 1$	$p$	$p(1-p)$	$(pe^t + 1 - p)$
Биномиальное $n = 1, 2, 3, \dots$ $0 < p < 1$	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \dots, n$	$np$	$np(1-p)$	$(pe^t + 1 - p)^n$
Отрицательное биномиальное $m = 2, 3, \dots$ $0 < p < 1$ Геометрическое, $m = 1$	$\binom{x-1}{m-1} (1-p)^{x-m} p^m$ $x = m, m+1, \dots$	$m/p$	$m(1-p)/p^2$	$\left[ \frac{pe^t}{1-(1-p)e^t} \right]^m$
Гипергеометрическое $N = 1, 2, 3, \dots$ $K = 0, 1, \dots, N$ $n = 1, 2, \dots, n$	$\frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}}$ $x = 0, 1, \dots, \min\{n, K\}$	$n \frac{K}{N}$	$n \frac{K}{N} \frac{N-n}{N-1} \left(1 - \frac{K}{N}\right)$	
Пуассона $\mu > 0$	$(\mu^x/x!) e^{-\mu}$ $x = 0, 1, 2, \dots$	$\mu$	$\mu$	$\exp[\mu(e^t - 1)]$

## 36.6 Непрерывные распределения

Распределение	$f_X(x)$	Среднее $\mu$	Дисперсия $\sigma^2$	ПФМ $M_X(t)$
Биномиальное	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \dots, n$	$np$	$np(1-p)$	$(pe^t + 1 - p)^n$
Геометрическое	$p(1-p)^{x-1}$ $x = 1, 2, 3, \dots$	$1/p$	$(1-p)/p^2$	$pe^t/(1 - (1-p)e^t)$
Пуассоновское	$(\mu^x/x!)e^{-\mu}, x = 0, 1, 2, \dots$	$\mu$	$\mu$	$\exp(\mu(e^t - 1))$
Экспоненциальное	$(1/\mu)e^{-x/\mu}, x \geq 0$	$\mu$	$\mu^2$	$(1 - \mu t)^{-1}$
Гамма	$\frac{1}{\Gamma(\alpha)\mu} \left(\frac{x}{\mu}\right)^{\alpha-1} e^{-x/\mu}, x > 0.$	$\alpha\mu$	$\alpha\mu^2$	$(1 - \mu t)^{-\alpha}$
Нормальное	$\frac{1}{\sqrt{2\pi}\sigma} \exp(-(x - \mu)^2/2\sigma^2), -\infty < x < \infty$	$\mu$	$\sigma^2$	$\exp(\mu t + \sigma^2 t^2/2)$
Хи-квадрат	$[2^{n/2}\Gamma(n/2)]^{-1} x^{n/2-1} e^{-x/2}, x > 0$	$n$	$2n$	$(1 - 2t)^{-n/2}$

**Определение.1** Пусть  $X_1, \dots, X_n$  – независимые одинаково распределённые случайные величины с общим распределением  $f(x)$ . Мы называем  $\{X_1, \dots, X_n\}$  случайной выборкой объёма  $n$  из генеральной совокупности  $f(x)$  и записываем их совместное распределение вероятностей как  $f(x_1, \dots, x_n) = f(x_1) \dots f(x_n)$

**Определение 2.** Любая функция случайных величин  $X_1, \dots, X_n$ , составляющих случайную выборку, называется *статистикой*.

Важные примеры.

**Замечание:** Любая статистика является случайной величиной, и мы говорим, что статистика  $h(X_1, \dots, X_n)$  является *несмещённой оценкой* некоторого параметра  $a$  генеральной совокупности  $f(x; a)$ , если  $Eh(X_1, \dots, X_n) = a$ .

## 36.7 Закон больших чисел

**Теорема.** Если  $EX = \mu$  существует, тогда  $\bar{X} \xrightarrow{d} \mu$  при  $n \rightarrow \infty$ .

**Доказательство.** Утверждение означает, что мы должны доказать, что  $M_{\bar{X}}(t) \rightarrow \mu$ .

$$\begin{aligned}
 M_{\bar{X}}(t) &= M_{\frac{1}{n} \sum_{j=1}^n X_j}(t) = (\text{свойство 6 из 15.4}) = M_{\sum_{j=1}^n X_j}\left(\frac{1}{n}t\right) = (\text{свойство 8}) = \\
 &= \left[M_{X_j}\left(\frac{1}{n}t\right)\right]^n = (\text{свойство 4}) \sim \left[1 + \frac{1}{n}t + \dots\right]^n \sim (\text{используем } e^\varepsilon \sim 1 + \varepsilon, \varepsilon \rightarrow 0) \\
 &= [e^{(\mu/n)t}]^n = e^{\mu t}, \quad n \rightarrow \infty.
 \end{aligned}$$

## 36.8 Устойчивость нормальных распределений

Центральная предельная теорема имеет асимптотический ( $n \rightarrow \infty$ ) смысл: приближённое равенство может быть достигнуто при малых или при больших  $n$ , это зависит от генерального распределения  $f(x)$ . Но есть случай, когда вышеприведённая взаимосвязь является точной и справедливой при любом  $n$ . Это случай *нормальной генеральной совокупности*, когда  $f(x)$  является нормальной плотностью. Легко доказать, что сумма двух независимых нормально распределённых случайных величин является нормально распределённой и это свойство, называемое *устойчивостью распределения*, справедливо для любого числа слагаемых.

Мы покажем это, используя производящую функцию моментов. Пусть  $N_1$  и  $N_2$  являются независимыми нормальными случайными величинами с параметрами  $(\mu_1, \sigma_1^2)$  и  $(\mu_2, \sigma_2^2)$  соответственно, тогда

$$M_{N_1+N_2}(t) = M_{N_1}(t)M_{N_2}(t) = e^{\mu_1 t + \sigma_1^2 t^2/2} e^{\mu_2 t + \sigma_2^2 t^2/2} = e^{(\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2},$$

то есть их сумма является нормальной величиной со средним  $\mu_1 + \mu_2$  и дисперсией  $\sigma_1^2 + \sigma_2^2$ .

## 37 FOR INSERTING

### Chapter 1.Вероятность

## 38 Парадокс игры в кости и его разрешение.

Две игральные кости бросаются. Если в сумме две кости дают 9, то выигрывает один, а если 10 - другой. Первый объясняет, что шансы выиграть у обоих игроков одинаковые, т.к. в обоих случаях каждый имеет по две возможные комбинации (3, 6), (4, 5) и (4, 6), (5, 5). Второй соглашается с ним, но замечает, что проигрывает. Проверяет каждую кость отдельно:

$$\frac{N(1)}{N} \approx \dots \approx \frac{N(6)}{N} \approx \frac{1}{6}$$

т.е.  $1/n$ , где  $n$  - это число граней. И все равно проигрывает.

Парадокс не смогли разрешить ни Лейбниц, ни Даламбер. Разрешение между тем простое:

кость 2							
к	1						
о	2						
с	3						9
т	4					9	10
ь	5				9	10	
1	6			9	10		
		1	2	3	4	5	6

Здесь число возможных исходов  $n = 36$ :

$$n(9) = 4 \quad \frac{n(9)}{n} = \frac{4}{36}$$

$$n(10) = 3 \quad \frac{n(3)}{n} = \frac{3}{36}$$

## 39 Понятие вероятности.

	1	2	3	4	5	6
1	*	*	*	*	*	*
2	*	*	*	*	*	*
3	*	*	*	*	*	*
4	*	*	*	*	*	*
5	*	*	*	*	*	*
6	*	*	*	*	*	*

Каждая точка здесь - возможный исход эксперимента. Назовем **элементарными событиями**  $\omega_1, \omega_2, \dots, \omega_{36}$ .  $\Omega = \{\omega_i\}$  - **пространство элементарных событий**.

Условия  $\nu_1 + \nu_2 = 9$ ,  $\nu_1 + \nu_2 = 10$  определяют **составные события**. Как любое тело состоит из элементарных составляющих атомов, так любое события состоит из элементарных событий.

**Классическое определение вероятности** если  $n(\Omega)$  - число  $\omega_i$  в пространстве элементарных событий  $\Omega$ , а  $n(A)$  - число элементарных событий  $\omega_i$  в пространстве элементарных событий  $A$ , то вероятность  $P(A) = n(A)/n(\Omega)$  (все элементарные события считаются равноправными, т.е. ??? с вероятностью  $1/n$ ).

**Геометрическое определение вероятности:** если число точек - континуум, ??? как в отрезке, то  $P(A) = \text{mes}(A)/\text{mes}(\Omega)$ . Все точки здесь равноправны, поэтому вер???

**Статистич. определение вероятности:** в отличие от первых двух, это определение экспериментальное:  $P(A) \approx N(A)/N$ , где  $N$  - полное число опытов, а  $N(A)$  - число появляющихся событий  $A$ , причем  $N$  и  $N(A)$  должны быть  $\gg 1$  (строго говоря они должны стремиться к бесконечности ( $\rightarrow \infty$ )).

## 40 Алгебра событий.

Алгеброй чаще всего называют множество, на котором определены операции сложения и умножения. *Пример:* числовое множество.

Пусть  $A, B, C$  - события. Что могут значить  $A + B$  и  $BC$ ? И вообще, что такое сумма и произведение? Сложение и умножение - две операции обладающие свойствами коммутативности, ассоциативности и дистрибутивности???  $A(B + C) = AB + AC$ . Эти свойства будут иметь место, если определить:

$A + B$  - объединение всех элементов событий, входящих в  $A$  и в  $B$  (интерпретация: произошло  $A$  и  $B$  или оба).

$AB$  - множество всех общих для  $A$  и  $B$  элементарных событий (интерпретация: произошло и  $A$  и  $B$ )

$A - B$  - произошло  $A$ , но не произошло  $B$ .

$\bar{\Omega} - A \equiv \bar{A}$  - противоположное событие.

$\bar{\Omega} = \emptyset$  - невозможность события.

$AB = \emptyset$  - несовместные события.

$\Omega$  - достоверное событие,  $A_1 + A_2 + \dots + A_k = \Omega$  и  $A_i A_j = \emptyset$ ,  $i \neq j$  - полная группа событий.

## 41 Аксиоматическое определение вероятности.

Геометрическое определение вероятности допускает обобщение. В геометрическом подходе мера - это длина, площадь, объем:  $\text{mes} A = \int_A d^n x$ . Все точки равноправны и  $\text{mes} A$  определена для *измеримых областей*, множество которых обозначены  $\mathcal{A}$ :  $A \in \mathcal{A}$ . Вместо декартовой меры можно ввести произвольную меру на  $\mathcal{A}$ , т.е.:

1) неотрицательность функции  $\mu$  из множества  $\mathcal{A}$ ,

2) аддитивность  $\mu(A_1 + A_2) = \mu(A_1) + \mu(A_2)$ , для непересекающихся множеств  $A_1$  и  $A_2$ .

Если добавить условие  $P(\Omega) = 1$ , то получим вероятность (т.е. выполняются все ее свойства)  $P(A)$ .

**Аксиоматическое определение**(А.Н. Колмогоров): вероятность - это мера  $P(A)$ , заданная на  $\sigma$ -алгебре  $\mathcal{A}$  измеримого пространства  $\Omega$  и удовлетворяет условию  $P(\Omega) = 1$ .

**Тройка**  $(\Omega, \mathcal{A}, P)$  называется **вероятностным пространством**.

Если  $\Omega = \{\omega_1, \dots, \omega_n\}$ , то  $\mathcal{A} = \{\emptyset, \{\omega_i\}, \{\omega_i, \omega_j\}, \{\omega_i, \omega_j, \omega_k\}, \dots, \{\omega_1, \dots, \omega_n\} \equiv \Omega, \}$  - дискретное.

## 42 Множественно - вероятностный словарь.

	ТЕОРИЯ МНОЖЕСТВ	ТЕОРИЯ ВЕРОЯТНОСТЕЙ
$\Omega$	Пространство	Достоверное событие
$\omega$	точка $\Omega$	Элементарное событие
$\emptyset$	пустое множество	Невозможное событие
$A \subset \Omega$	подмножество $\Omega$	Событие
$A \subset B$	множество $A$ содержится в множестве $B$	Событие $A$ влечет за собой событие $B$
$A \cup B$	объединение множеств	Объединение событий (или $A$ или $B$ или оба)
$A \cap B$	??? множеств	совмещение событий (и $A$ и $B$ )
$\bar{A}$	дополнение к множеству $A$	событие противоположное $A$
$A \setminus B$	разность множеств	разность событий
$A \cap B = \emptyset$	??? множества	несовместные события
$A_i \cup A_j = \Omega$ $A_i \cap A_j = \emptyset \ i \neq j$	разбиение множества $\Omega$	полная группа событий

Chap-

ter 2. Комбинаторика

## 43 Принцип умножения

Задачи о числе различных порядков расположения множества объектов. Для того, чтобы понять суть данной задачи, можно попытаться ответить на вопросы: “Сколько существует способов расставить 10 книг на книжной полке?”, “Как можно разместить 8 шаров по 10 ящикам?”, “Как можно прийти из точки  $(0, 0)$  в точку  $(6, 4)$ , делая только единичные шаги и только и только в положительных направлениях оси  $x$  или  $y$ ?”. Общий способ: *построение дерева* возможных путей. Пусть необходимо выполнить одно за другим  $k$  действий, причем, первое -  $n_1$  способами, второе -  $n_2$  способами, ...,  $k - l - n_k$  - способами.

Принцип умножения: все  $k$  действий вместе могут быть выполнены  $n_{12...k} = n_1 n_2 ... n_k$  способами.

## 44 Перестановки и размещения.

Дано  $n$  предметов. Сколько существует способов их линейного размещения (упорядочивания):

$a, b, c, \dots$

$b, a, c, \dots$

$c, b, a, \dots$

$c, a, b, \dots$

и так далее.

Принцип умножения легко применить, если представить себе  $n$  ящиков и занумеровать предметы от 1 до  $n$ . Тогда на

**1-ое место** - можно поместить любой из  $n$  предметов

**2-ое место** -  $(n - 1)$  предмет

... ..

**последнее место** - 1 предмет

Число  $P_n = n(n-1)\dots 1 \equiv n!$  - называется числом перестановок из  $n$  элементов.

*ПРИМЕЧАНИЕ* мы должны уметь различать эти предметы, иначе нельзя говорить о порядке - любая перестановка будет давать тот же итог!

Пусть дано  $n$  предметов и их надо разместить по  $k < n$  ячейкам. Тогда принцип умножения дает:

**1-ое место** - помещаем любой из  $n$  предметов

**2-ое место** - помещаем любой из  $(n-1)$  предметов

... ..

**$k$ -ое место** - помещаем любой из  $n-k+1$  предметов.

Число размещений из  $n$  элементов по  $k$ :

$$A_n^k = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}$$

## 45 Сочетания

Сочетание есть набор элементов, рассматриваемых без учета их порядка. Например, мы хотим выбрать из 4-х книг 3 и заложить ими три места на нижней полке. Это можно сделать  $A_4^3 = 4*3*2 = 24$  способами (число размещений). Если же мы хотим эти три книги взять с собой в дорогу, то их порядок не имеет значения: из четырех книг  $A, B, C, D$  можно взять  $ABC, BCD, ABD$  или  $BCD$  - т.е. всего четыре способа.

Число сочетаний из  $n$  элементов по  $k$  обозначается как  $C_n^k$  или  $\binom{n}{k}$ .

**Вычисление  $C_n^k$ .** каждое сочетание путем перестановки  $k$  элементов можно превратить в  $k!$  размещений, так что  $C_n^k k! = A_n^k$ . Откуда

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!} \equiv \binom{n}{k}$$

Правило Паскаля:

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \quad 1 \leq k \leq n$$

Формула Вардермонда:

$$\sum_{k=0}^n \binom{m-l}{k} \binom{l}{n-k} = \binom{m}{n}$$

## 46 Перестановки объектов с повторениями

Мы рассматриваем размещение объектов, которые отличаются один от другого. Если они неразличимы, число размещений везде будет 1. Но что будет, если только некоторые объекты неразличимы?

Пусть дано множество из  $n$  элементов, в котором  $n_1$  элементов принадлежит 1-му типу,  $n_2$  - 2-му типу, ...,  $n_k$  -  $k$ -му типу, и элементы каждого типа неразличимы между собой. Тогда общее число перестановок общего множества  $n$  элементов равно

$$\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1! \dots n_k!}$$



Доказательство:

чтобы выполнить полное число перестановок в случае всех различных элементов, равное  $n!$  ( $n_1 + \dots + n_k = n$ ), надо

$$\binom{n}{n_1, \dots, n_k} n_1! n_2! \dots n_k! = n!$$

откуда и следует формула.

Частный случай:  $k = 2$

$$\binom{n}{n_1, n_2} = \frac{n!}{n_1! n_2!} = \frac{n!}{n_1! (n - n_1)!}$$

- число сочетаний из  $n$  различных предметов по  $k$ .

Биноминальная теоремы: если  $n$  - целое положительное число, то

$$(a + x)^n = \sum_{k=0}^n \binom{n}{k} a^n x^{n-k}$$

## 47 Формула Стирлинга.

Рассмотрим интеграл

$$\int_1^n \ln x dx$$

возьмем этот интеграл по частям:

$$x \ln x \Big|_1^n - \int_1^n dx = n \ln n - n + 1$$

По формуле ???:

$$\approx \frac{\ln 1 + \ln n}{2} + \ln 2 + \dots + \ln(n-1) = \ln n! - \frac{1}{2} \ln n$$

Таким образом, получаем равенство:

$$n \ln n - n + 1 \approx \ln n! - \frac{1}{2} \ln n$$

пренебрежем единицей:

$$\ln n! \approx \left(n + \frac{1}{2}\right) \ln n - n$$

$$n! \approx n^{n+1/2} e^{-n}$$

Более точная формула:

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

И еще более точная:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \left(1 + \frac{1}{12n}\right)$$

### Chapter 3. Классические задачи теории вероятности

## 48 Выборка без возвращений.

Пусть  $E$  состоит из  $n$  (выше 36) элементов и все исходы равновероятны:  $P(e) = \frac{1}{n}$ , для любого  $e \in E$  (выше 36). Тогда

$$P(A) = \sum_{e \in A} P(e) = \frac{1}{n} \sum_{e \in A} 1 = \frac{n(A)}{n}$$

классическое определение вероятности (выше  $P(9) = 4/36$ ).

Пусть задано множество  $\{a_1, a_2, \dots, a_n\}$ , которое мы будем называть генеральной совокупностью.

Выберем наугад один из этих элементов:  $\alpha_1$ . Выбирая  $\alpha_1$  мы не знаем какой именно попадетс:  $P\{\alpha_1 = a_1\} = \frac{1}{n}$ , как впрочем и  $P\{\alpha_1 = a_k\} = \frac{1}{n}$ . Из оставшихся выберем  $\alpha_2$  Какова вероятность  $P\{\alpha_1 = a_1, \alpha_2 = a_2\}$ ?

$$P\{\alpha_1 = a_1, \alpha_2 = a_2\} = \frac{1}{n^2 - n} = \frac{1}{n(n-1)}$$

$$P\{\alpha_1 = a_1, \alpha_2 = a_2, \alpha_3 = a_3\} = \dots$$

$$P\{\alpha_1 = a_1, \alpha_2 = a_2, \alpha_3 = a_3, \dots, \alpha_k = a_k\} = \frac{1}{(n)_k}$$

где  $(n)_k$  - число выборок объемом  $k$  из  $n$ , т.е. число размещений из  $n$  по  $k$ :

$$(n)_k = n(n-1)\dots(n-k+1)$$

если  $k = n$ , то  $(n)_k = n!$  - число перестановок.

## 49 Выборка с возвращением.

Здесь полное число различных выборок элементарных событий  $\underbrace{n * n * \dots * n}_k = n^k$

Вероятность любой заданной комбинации равна  $1/n^k$ .

В том числе и элементарные события:

$$P\{\nu_1 = 1, \nu_2 = 1, \dots, \nu_k = 1\} = \frac{1}{n^k}$$

$$P\{\nu_1 = 1, \nu_2 = 2, \dots, \nu_k = k\} = \frac{1}{n^k}$$

$$P\{\nu_1 = n_1, \nu_2 = n_2, \dots, \nu_k = n_k\} = \frac{1}{n^k}$$

Вероятность того, что все элементы выборки окажутся разными (раньше она была равна 0), теперь составляет событие

$$P\{\nu_1 \neq \nu_2 \neq \dots \neq \nu_k\} = \frac{\text{число выб. с несовп. элемен}}{n^k} = \frac{\text{число выб. без возвр.}}{n^k} = \frac{n!}{(n-k)!n^k}$$

## 50 Гипергеометрическое распределение.

Допустим в 1-ой урне  $n_1$  черных шаров и  $n_2$  белых ( $n_1 + n_2 = n$ ).

Если мы выберем один шар, то с вероятностью  $P\{1\} = \frac{n_1}{n}$  он черный и вероятностью  $P\{2\} = \frac{n_2}{n}$  он белый. Выберем (без возвращения) 2 шара. Вероятность того, что оба шара черные равна  $\frac{n_1(n_1-1)}{n(n-1)}$ .

Усложним задачу: выбираем (без возврата)  $k$  шаров. Какова вероятность, что  $k_1$  из них являются черными?

Всего выборов, различающихся составом  $\binom{n}{k}$ .

Число выборов черных шаров:  $\binom{n_1}{k_1}$  - столько способов выбрать  $k_1$  черных шаров из  $n_1$ .

Число выборов белых шаров:  $\binom{n_2}{k_2}$  - столько способов выбрать  $k_2$  белых шаров из  $n_2$ .

$$k_2 = k - k_1 \quad n_2 = n - n_1$$

$$P = \frac{\text{число способов выбрать } k_1 \text{ черных шаров}}{\text{полное число выборов}} = \frac{\binom{n_1}{k_1} \binom{n_2}{k_2}}{\binom{n}{k}} = \frac{\binom{n_1}{k_1} \binom{n-n_1}{k-k_1}}{\binom{n}{k}}$$

- гипергеометрическое распределение.

## 51 Спортлото

Участники лотереи из 49 наименований видов спорта (обозначенных просто цифрами) называют 6. Выигрыш определяется тем, сколько из них совпадет с шестью наименованиями, заранее выделенными комиссией. Назовем их черными шарами, остальные белыми. Вероятность угадать все шесть:  $k_1 = 6, k = 6, n_1 = 6, n = 49$

$$P = \frac{1}{\binom{49}{6}} = \frac{6!43!}{49!} \approx 7.2 * 10^{-8}$$

по формуле Стирлинга

$$n! \sim \sqrt{2\pi n} n^n e^{-n}$$

**Chapter. Независимость.**

## 52 Условные вероятности.

Вернемся к определению вероятности: дано  $N$  опытов, в  $N(A)$  из них наступило событие  $A$ .  $N(A)/N$  - частота наступления события  $A$ , при больших  $N$  она приблизительно равна  $P(A)$ . Пусть на ряду с событием  $A$  возможно и событие  $B$  и оно наступило (вместе с  $A$ ) в  $N(AB)$  опытах. Тогда

$$\frac{N(AB)}{N} \approx P(AB)$$

$$P(Ab) = P(A)P(B|A)$$

$$\frac{N(AB)}{N_A} \frac{N_A}{N} \approx P(B|A)P(A)$$

где  $P(B|A)$  - условная вероятность события  $B$  при условии события  $A$ . Если  $P(A) = 0$ , то

$$P(B|A) \approx \frac{P(AB)}{P(A)}$$

## 53 Основное правило исчисления вероятностей.

1.  $0 \leq P(A) \leq 1, \quad P(\emptyset) = 0, \quad P(\Omega) = 1$

2.  $P(A + B) = P(A) + P(B) - P(AB).$

Если несовместны, то  $P(A + B) = P(A) + P(B)$ ; в общем случае  $P(A + B) \leq P(A) + P(B)$ ,  $\sum P(A_i) = 1$ , если  $\{A_i\}$  - разбиение, то  $P(\bar{A}) = 1 - P(A)$

3.  $P(AB) = P(A|B)P(B), \quad P(A\Omega) = P(A)$

Если независимы, то  $P(AB) = P(A)P(B)$

$$P(A_n, \dots, A_1) = P(A_n|A_{n-1}, \dots, A_1)P(A_{n-1}|A_{n-2}, \dots, A_1) \dots P(A_2|A_1)P(A_1)$$

**Замечание:** независимые события совместны, несовместные события зависимы.

## 54 Формула полной вероятности.

Пусть  $B_1, \dots, B_n$  - разбиение  $\Omega$ :

$$\Omega = B_1 + \dots + B_n \quad B_i B_j = \emptyset \quad i \neq j$$

$$A = A\Omega = AB_1 + \dots + AB_n$$

$AB_i$  и  $AB_j$  - не пересекаются.

$$P(A) = P(AB_1 + \dots + AB_n) = \sum_{i=1}^n P(AB_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

## 55 Формула Байеса

Терминология:  $B_i$  - гипотезы,  $P(B_i)$  - априорные вероятности,  $P(B_i|A)$  - апостериорные вероятности.

Рассмотрим теперь не  $P\{A|B_i\}$ , а

$$P\{B_i|A\} = \frac{P\{B_i A\}}{P\{A\}} = \frac{P\{AB_i\}}{P\{A\}} = \frac{P\{A|B_i\}P\{B_i\}}{\sum_{j=1}^n P\{A|B_j\}P\{B_j\}}$$

по формуле полной вероятности.

**Пример:** Электрические лампы выпускаются двумя заводами:  $B_1$  и  $B_2$ . Пусть  $A$  - это количество бракованных лампочек, тогда

$P\{A|B_1\}$  - доля брака 1-го завода

$P\{A|B_2\}$  - доля брака 2-го завода

В корзине  $n_1$  ламп с первого завода и  $n_2$  - со второго завода.

1. Какова вероятность того, что наугад выбранная лампочка не горит?

$$P\{A\} = P\{A|B_1\}P\{B_1\} + P\{A|B_2\}P\{B_2\}$$

2. Пусть действительно вынутая лампочка не горит, т.е. выполнено первое условие.

Какова вероятность того, что бракованная лампочка изготовлена на 1-ом заводе:

$$P\{B_1|A\} = \frac{P\{A|B_1\}P\{B_1\}}{P\{A|B_1\}P\{B_1\} + P\{A|B_2\}P\{B_2\}}$$

**Chapter. Случайная величина.**

## 56 Дискретная случайная величина.

Пусть дан кубик. Все его грани принадлежат множеству  $\Omega$ . Занумеруем грани кубика:

$$\cdot \rightarrow 1 \quad \cdot \rightarrow 2 \quad \cdot \rightarrow 3 \quad \cdot \rightarrow 4 \quad \cdot \rightarrow 5 \quad \cdot \rightarrow 6$$

Таким образом, нам необходимо на множестве  $\Omega$  задать функцию  $\xi = f(\omega)$ . Здесь число различных значений функции конечно.

Обозначим:

$$P(\cdot) = P_1, \quad P(\cdot) = P_2, \quad P(\cdot \cdot) = P_3, \quad P(\cdot \cdot) = P_4, \quad P(\cdot \cdot \cdot) = P_5, \quad P(\cdot \cdot \cdot) = P_6$$

$P$	$P(\cdot)$	$P(\cdot)$	$P(\cdot \cdot)$	$P(\cdot \cdot)$	$P(\cdot \cdot \cdot)$	$P(\cdot \cdot \cdot)$
$\omega$	$\cdot$	$\cdot$	$\cdot \cdot$	$\cdot \cdot$	$\cdot \cdot \cdot$	$\cdot \cdot \cdot$
$\xi$	1	2	3	4	5	6
$P_\xi$	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$

- Случайной величиной называется вещественная функция, определенная на множестве  $\Omega$

Дискретная случайная величина задается следующим образом:

$$\xi = x_1, x_2, \dots, x_k$$

$$P_\xi = P_1, P_2, \dots, P_k$$

- Совокупность вероятностей  $P_1, P_2, \dots, P_k$  - называется распределением вероятностей дискретной случайной величины.

Свойства:

- $P_k \geq 0$
- $\sum_k P_k = 1$

Примеры:

- $P_k = \frac{1}{n}$ ,  $k = 1, 2, \dots, n$  - равномерное распределение.
- $P_k = \binom{n}{k} p^k (1-p)^{n-k}$ ,  $k = 1, 2, \dots, n$  - биномиальное распределение ( $0 < p < 1$ ).
- $P_k = e^{-a} \frac{a^k}{k!}$ ,  $k = 1, 2, 3, \dots$  - распределение Пуассона.

## 57 Схема Бернулли.

Пусть бросается непрерывно монета или производится другой опыт с двумя возможными исходами 1 — 0.

Последовательность независимых испытаний с двумя возможными исходами называется схемой Бернулли.

Дано множество элементарных событий для серии из  $n$  испытаний:

$$1\{000\dots 000$$

$$n \left\{ \begin{array}{l} 100\dots 000, \\ 010\dots 000, \\ 001\dots 000, \\ \dots\dots\dots, \\ 000\dots 010, \\ 000\dots 001, \end{array} \right.$$

$$\frac{n(n-1)}{2} \left\{ \begin{array}{l} 110\dots 00, \\ 101\dots 00, \\ \dots\dots\dots, \\ 000\dots 11, \end{array} \right.$$

$$1\{\underbrace{111\dots 11}_n\}$$

Как получить  $\frac{n(n-1)}{2}$  из  $n$ ? - добавить 1 вместо разных нулей и учесть удвоение (смотреть две первые строчки 110... 110...)

Что такое  $\frac{n(n-1)}{2}$ ? - это число способов разместить 2 шара по  $n$  ящикам или число способов взять 2 шара из  $n$  шаров.

Пусть  $C_n^k$  - число способов разместить  $k$  шаров по  $n$  ящикам. Тогда  $P(\nu = k) = C_n^k p^k (1-p)^{n-k}$ . Коэффициенты  $C_n^k$  найдем сравнивая условие нормировки

$$\sum_{k=0}^n C_n^k p^k (1-p)^{n-k} = 1$$

с биномом Ньютона:

$$(a+b)^n = \sum_{k=0}^n \frac{n!}{k!(n-k)!} a^k b^{n-k}$$

подставляя сюда  $a = p$  и  $b = 1-p$  находим Биномиальное распределение

$$P(\nu = k) = \frac{n! p^k (1-p)^{n-k}}{k!(n-k)!}$$

## 58 Теорема Пуассона

Зададим биномиальное распределение

$$P_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{1}{k!} \frac{n!(n-k+1)\dots n}{n!} p^k \frac{(1-p)^n}{(1-p)^k} =$$

Пусть  $a = np$  - среднее значение, тогда

$$= \frac{1}{k!} \left(1 - \frac{k}{n} + \frac{1}{n}\right) \dots 1 a^k \frac{(1 - \frac{a}{n})^n}{(1-p)^k}$$

Теперь пусть  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , но  $pn = a = \text{const}$

$$P_k \rightarrow \frac{a^k}{k!} \lim_{n \rightarrow \infty} \left(1 - \frac{a}{n}\right)^n = \frac{a^k}{k!} e^{-a}$$

Причем  $P_k = \frac{a^k}{k!} e^{-a}$  - распределение Пуассона.

ТЕОРЕМА ПУАССОНА

$$P_k \rightarrow \frac{a^k}{k!} \lim_{n \rightarrow \infty} \left(1 - \frac{a}{n}\right)^n = \frac{a^k}{k!} e^{-a}$$

## 59 Функция распределения и ее свойства.

Если  $\Omega$  - непрерывное множество, то  $\xi = f(\omega)$  называют непрерывной случайной величиной, которая характеризуется функцией распределения

$$F_\xi(x) \equiv P(f(\xi) < x)$$

Свойства функции распределения  $F_\xi(x)$

1.  $0 \leq F_{\xi}(x) \leq 1$  как вероятность.

$$2. F_{\xi}(x_2) - F_{\xi}(x_1) = P(\underbrace{\xi < x_2}_{A_2}) - P(\underbrace{\xi < x_1}_{A_1}) = P\left(\underbrace{(\xi < x_2)}_{A_2} - \underbrace{(\xi < x_1)}_{A_1}\right) =$$

$$= P(x_1 \leq \xi \leq x_2) \quad F_{\xi}(x_2) \geq F_{\xi}(x_1), \quad x_2 > x_1 - \text{монотонная.}$$

$$P(x_1 \leq \xi \leq x_2) = F_{\xi}(x_2) - F_{\xi}(x_1)$$

3. Если существует  $x_{min}$ , то  $F_{\xi}(x) = 0, x \leq x_{min}$

Если существует  $x_{max}$ , то  $F_{\xi}(x) = 1, x > x_{max}$

Если не существует максимального и минимального значений, то должны существовать пределы:

$$F_{\xi}(x) \rightarrow 0, \quad \text{при} \quad x \rightarrow -\infty$$

$$F_{\xi}(x) \rightarrow 1, \quad \text{при} \quad x \rightarrow +\infty$$

4.  $F_{\xi}(x)$  - непрерывная слева:  $\lim_{x \rightarrow x_0} F_{\xi}(x) = F_{\xi}(x_0)$

5.  $F_{a+b\xi}(x) = F_{\xi}\left(\frac{x-a}{b}\right), b > 0$

6. Функция распределения дискретной случайной величины:

$$F_{\xi}(x) = \sum_{j=1}^n P_j u(x - x_j)$$

$$u(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

## 60 Плотность распределения.

Пусть  $F(x)$  не только непрерывна, но и дифференцируема:

$$\Delta F(x) \sim F'(x) \Delta x.$$

Производная  $F'_{\xi}(x) = P_{\xi}(x)$  называется плотностью распределения вероятности.

Такое название вытекает из определения:

$$P_{\xi}(x) = \lim_{\Delta x \rightarrow 0} \frac{F_{\xi}(x + \Delta x) - F_{\xi}(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(x < \xi < x + \Delta x)}{\Delta x}$$

### СВОЙСТВА

1. Если  $F_{\xi}(x) = const$ ,  $a < x < b$ , то  $P_{\xi}(x) = 0$  в этом интервале.

2.  $\int_a^b P_{\xi}(x) dx = P(a \leq \xi \leq b)$  (из 2)

3.  $\int_{-\infty}^x P_{\xi}(x) dx = F_{\xi}(x) - F_{\xi}(-\infty) = F_{\xi}(x)$  (смотри 3) - обратное соотношение.

4. Если существует  $x_{min}$ , то  $P_\xi(x) = 0$ ,  $x < x_{min}$ , если существует  $x_{max}$ , то  $P_\xi(x) = 0$ ,  $x > x_{max}$ , и выполняется  $\int_{x_{min}}^{x_{max}} P_\xi(x) dx = 1$ .

Если не существует минимального и максимального значений, то  $P_\xi(x) \rightarrow 0$ ,  $x \rightarrow \infty$  и  $x \rightarrow -\infty$  и выполняется условие нормировки  $\int_{-\infty}^{\infty} P_\xi(x) dx = 1$  (из 3)

5.  $P_{a+b\xi}(x) = \frac{1}{b} P_\xi\left(\frac{x-a}{b}\right)$  - (см. 5)

6. В дискретном случае  $P_\xi(x) = \sum_{j=1}^n p_j u'(x-x_j) = \sum_{j=1}^n p_j \delta(x-x_j)$ , где  $\delta(x-x_j)$  -  $\delta$ -функция. Общее название и для  $P_\xi(x)$  и для  $p_j$  - распределения.

## 61 Дельта-функция Дирака.

Что такое  $u'(x-x_i)$ ? Эта функция везде равна нулю кроме одной точки  $x = x_i$ , где она не существует. Но это в классическом смысле. Сегодня математики и физики работают с такими функциями, называемыми обобщенными функциями или распределениями.

Пусть  $f(x)$  - непрерывная, дифференцируемая функция, обращающаяся в нуль при  $x \leq a$  и  $x \geq b$  и  $x_i \in (a, b)$

$$\int_a^b f'(x) u(x-x_i) dx = u(x-x_i) f(x) \Big|_a^b - \int_a^b f(x) u'(x-x_i) dx = - \int_a^b f(x) \delta(x-x_i) dx$$

С другой стороны, по определению  $u(x-x_i)$ :

$$\int_a^b f'(x) u(x-x_i) dx = \int_{x_i}^b f'(x) dx = f(b) - f(x_i) = -f(x_i)$$

Таким образом:

$$\int_a^b f(x) \delta(x-x_i) dx = \begin{cases} f(x_i), & x_i \in (a, b) \\ 0, & x_i \notin [a, b] \end{cases}$$

- это одно из определений  $\delta$ -функции.

### Chapter. Математическое ожидание (среднее значение) случайной величины.

Пусть  $\xi$  - случайная величина со значениями  $x_1, \dots, x_n$  и вероятностями  $p_1, \dots, p_n$ .

$$S_N = \sum_{j=1}^N \xi_j = N_1 x_1 + \dots + N_n x_n$$

$$\frac{S_N}{N} = \frac{N_1}{N} x_1 + \dots + \frac{N_n}{N} x_n$$

При больших  $N$   $N_k/N \sim p_k$ .  $S_N/N \sim \sum_{k=1}^n x_k p_k = \bar{\xi}$  - математическое ожидание (среднее значение) =  $M\xi$ .

В случае непрерывной случайной величины:

$$\sum_{k=1}^n x_k p(x_k) \Delta x_n = \int_{-\infty}^{\infty} x p(x) dx = M\xi$$



## СВОЙСТВА МАТ. ОЖИДАНИЯ.

1. Если  $\xi = c$  - постоянная, то  $M\xi = c$
2.  $Mc\xi = cM\xi$
3.  $M(\xi + c) = M\xi + c$ . Вообще  $M(\xi + \nu) = M\xi + M\nu$ , но об этом позже.
4. Если  $a \leq \xi b$ , то  $a \leq M\xi \leq b$ . Всегда  $M\xi \leq M|\xi|$ .
5. Если  $\xi \geq 0$  и  $M\xi = 0$ , то  $\xi = 0$  с вероятностью 1.
6.  $Mu(x - \xi) = F_\xi(x)$

## 62 Моменты случайной величины.

$$m_n \equiv \bar{\xi}^n = \int_{-\infty}^{\infty} x^n dF_\xi(x) = \int_{-\infty}^{\infty} x^n P_\xi(x) dx$$

- момент  $n$ -порядка.  $n = 0, 1, 2, \dots$ ,  $\bar{\xi}^0 = 1$ ,  $\bar{\xi}^1 = \bar{\xi}$ .

Если интеграл сходится, то мы говорим, что существует  $n$ -ый момент. Нулевой момент существует всегда.

$|\bar{\xi}|^\nu \equiv M|\xi|^\nu$  - **абсолютный момент**.

$\mu_\nu = M[(\xi - \bar{\xi})^\nu] = \int_{-\infty}^{\infty} (x - \bar{\xi})^\nu dF_\xi(x)$  - **центральные моменты**.

Раскрывая  $(\xi - m)^\nu$ , при  $m = \bar{\xi}$ , находим

$$\mu_0 = 1$$

$$\mu_1 = 0$$

$$\mu_2 = \alpha_2 - m^2$$

$$\mu_3 = \alpha_3 - 3m\alpha_2 + 2m^3$$

$$\mu_4 = \alpha_4 - 4m\alpha_3 + 6m^2\alpha_2 - 3m^4$$

## 63 Медиана и мода.

1. Если точка  $x_0$  разделяет всю вероятность на две равные части, она называется медианой распределения:  $F_\xi(x_0) = 1/2$ . ??? распределение имеет по крайней мере одну медиану. ??? не одной медианы.
2. Для непрерывного распределения любая точка  $x_0$  максимума плотности вероятности называется модой. Унимодальные, мультимодальные распределения.

## 64 Дисперсия.

Дисперсией называют:

$$D\xi = M(\xi - \bar{\xi})^2 = \sigma^2$$

где  $\sigma$  - стандартное отклонение (мера рассеяния, ширина распределения, статистическая погрешность).

## СВОЙСТВА ДИСПЕРСИИ

1.  $D\xi \geq 0$ ;  $D\xi = 0$ , тогда и только тогда, когда  $D(\xi = c) = 1$ , где  $c = \text{const}$ .
2.  $D(c\xi) = c^2 D\xi$
3.  $D(\xi + c) = D\xi$
4.  $D\xi = M\xi^2 - (M\xi)^2$

#### ПРИМЕРЫ:

1. ширина  $2a$ , полуширина  $a$

$$D\xi = \frac{1}{2a} 2 \int_0^a x^2 dx = \frac{1}{3} a^2 \quad \sigma = \frac{a}{\sqrt{3}}$$

2.  $\frac{1}{a} e^{-|x|/a}$

$$D\xi = \frac{2}{a} \int_0^\infty x^2 e^{-x/a} dx = 2a^2 \cdot 2 = 4a^2 \quad \sigma = \sqrt{D\xi} = 2a$$

## 65 Симметричные случайные величины.

$P(\xi < c - x) = P(\xi > c + x)$  - симметричная относительно  $c$ .

$$F_\xi(c - x) = 1 - F_\xi(c + x)$$

Если имеется плотность  $P_\xi(c - x) = P_\xi(c + x)$  (четно относительно  $x$ )

Доказательство  $M\xi = c$ :

$$\int_{-\infty}^{\infty} x P_\xi(x) dx = \int_{-\infty}^{\infty} (x - c + c) P_\xi(x) dx = \int_{-\infty}^{\infty} x' P_\xi(x' + c) dx' + c = c$$

Здесь была сделана замена  $x - c = x'$  и  $x = x' + c$ .

$$\mu_{2k+1} = 0$$

$$D\xi = \alpha_2 - \alpha_1^2 = \alpha_2. \text{ если } c = 0 \text{ (симметричное относительно нуля).}$$

## 66 Характеристики формы распределения.

Случайная величина  $\xi$  называется симметричной относительно нуля, если  $P(\xi > x) = P(\xi < -x)$ , для любого  $x > 0$

$$1 - F_\xi(x) = F_\xi(-x) \text{ и } P_\xi(x) = P_\xi(-x)$$

Если случайная величина симметричная, то все нечетные моменты равны нулю.

Простейшие - 1-ый (математическое ожидание) и 3-ий -  $\mu_3$ . В результате получаем  $\gamma_1 = \mu_3/\sigma^3$  - коэффициент асимметрии (может принимать как положительные значения, так и отрицательные).  $\gamma_2 = \mu_4/\sigma^4 - 3$  - коэффициент эксцесса.

**Chapter. Многомерные случайные векторы.**

## 67 Совместное распределение случайной величины.

Рассмотрим две случайные величины  $\xi$  и  $\eta$ :

$$F_{\xi}(x) = P\{\xi < x\} \quad P_{\xi}(x) = F'_{\xi}(x)$$

$$F_{\eta}(y) = P\{\eta < y\} \quad P_{\eta}(y) = F'_{\eta}(y)$$

Совместная функция распределения:

$$F_{\xi\eta}(x, y) = P\{\xi < x, \eta < y\}$$

$$F_{\xi\eta}(x, \infty) = F_{\xi}(x) \quad F_{\xi\eta}(\infty, y) = F_{\eta}(y)$$

Совместная плотность:

$$\frac{\partial^2 F_{\xi\eta}}{\partial x \partial y} = P_{\xi\eta}(x, y)$$

$$P_{\xi\eta}(x, y) dx dy = P\{\xi \in dx, \eta \in dy\}$$

Случайный вектор  $\vec{\rho} = (\xi, \eta)$  или  $\xi \vec{i} + \eta \vec{j}$ ,

$$dx dy = ds$$

$$P_{\xi} = \int P_{\xi\eta} dy$$

$$\vec{r} = x \vec{i} + y \vec{j}$$

Это можно распространить на трехмерные случайные вектора и многомерные случайные вектора:

$$\vec{\rho} = \xi \vec{i} + \eta \vec{j} + \zeta \vec{k}, \quad P_{\vec{\rho}}(\vec{r}) ds = P\{\vec{\rho} \in ds\}, \quad \int \rho ds = 1$$

$$\vec{r} = x \vec{i} + y \vec{j} + z \vec{k}, \quad P_{\vec{\rho}}(\vec{r}) dV = P\{\vec{\rho} \in dV\}, \quad \int \rho dV = 1$$

## 68 Условная плотность распределения.

Напомним выражение для условной вероятности:

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}$$

Пусть  $A$  такое, что  $\xi \in dx$ ,  $P\{A \cap B\} = P\{\xi \in dx, \eta \in dy\} = p_{\xi\eta}(x, y) dx dy$ , и пусть  $B$  такое, что  $\eta \in dy$ ,  $P\{B\} = P\{\eta \in dy\} = p_{\eta}(y) dy$

$$P\{\xi \in dx | \eta \in dy\} = \frac{P_{\xi\eta}(x, y)}{P_{\eta}(y)} dx$$

$$P_{\xi}(x | \eta = y) dx = \frac{P_{\xi\eta}(x, y)}{P_{\eta}(y)} \quad \text{если} \quad P_{\eta}(y) \neq 0$$

$$P_{\xi}(x, y) = P_{\xi}(x | \eta = y) P_{\eta}(y)$$

## 69 Формула полного математического ожидания.

$$\begin{aligned} M\xi &= \int_{-\infty}^{\infty} xP_{\xi}(x)dx = \int_{-\infty}^{\infty} dx x \int_{-\infty}^{\infty} dy P_{\xi}(x, y) = \int_{-\infty}^{\infty} dx x \int_{-\infty}^{\infty} dy P_{\xi}(x|\eta = y)P_{\eta}(y) = \\ &= \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx x P_{\xi}(x|\eta = y)P_{\eta}(y) = \int_{-\infty}^{\infty} dy M(\xi|y)P_{\eta}(y) = \end{aligned}$$

Учитывая, что  $\int f(y)P_{\eta}(y)dy = Mf(\eta)$ , получим

$$= MM(\xi|\eta)$$

Сначала  $\bar{\xi}(\eta = y)$ , затем  $\bar{\xi} = \int_{-\infty}^{\infty} \bar{\xi}(y)P_{\eta}(y)dy$ .

Таким образом, получили соотношение:

$$M\xi = MM(\xi|\eta)$$

## 70 Независимые случайные величины.

Если события  $\xi < x$  и  $\eta < y$  независимы, то и сами случайные величины называются независимыми.

$$F_{\xi\eta}(x, y) = P\{\xi < x, \eta < y\} = P\{\xi < x\}P\{\eta < y\} = F_{\xi}(x)F_{\eta}(y)$$

$$P_{\xi\eta}(x, y) = P_{\xi}(x)P_{\eta}(y)$$

$$M_{\xi\eta} = \int \int xy P_{\xi}P_{\eta} dx dy = M_{\xi}M_{\eta}$$

Если  $\xi$  и  $\eta$  - независимые величины, то  $f(\xi)$  и  $g(\eta)$  тоже независимы.

$$\Psi_{\xi_1\xi_2}(k_1, k_2) = \Psi_{\xi_1}(k_1)\Psi_{\xi_2}(k_2)$$

## 71 Изотропный вектор с независимыми координатами.

Изотропным называется случайный вектор, совместная плотность которого не меняется при поворотах и отражениях систем координат:

$$P_{\xi\eta}(B) = P'_{\xi\eta}(B).$$

Выберем точку  $B$  так, чтобы  $x = y$ :

$$P_{\xi\eta}(x, y) = P_{\xi\eta}(y, x)$$

$$P_{\xi}(x)P_{\eta}(y) = P_{\xi}(y)P_{\eta}(x) \quad P_{\xi} = P_{\eta}$$

Тогда, с учетом независимости координат  $P_{\xi\eta}(x, x) = P_{\xi}(x)P_{\eta}(x) \equiv P^2(x)$  (распределения  $\xi$  и  $\eta$  должны быть одинаковыми).

Теперь повернем систему так, чтобы ось  $x'$  проходила через точку  $B$ . Тогда  $p(\sqrt{2}x)p(0)$

$$p^2(x) = p(0)p(\sqrt{2}x)$$

Обозначим  $f(x) = \ln p(x)$ ,  $p = e^{f(x)}$

$$2f(x) = f(0) + f(\sqrt{2}x)$$

будем искать решение в виде ряда:

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots$$

$$2a_0 + 2a_1x + 2a_2x^2 + 2a_3x^3 + \dots = a_0 + a_0 + a_1\sqrt{2}x + a_22x^2 + a_32\sqrt{2}x^3 + \dots$$

$a_1 = 0$ ,  $a_3 = 0$  и все  $a_n$  с  $n > 3$  будут равны нулю.

Таким образом:

$$P(x) = e^{a_0 + a_2x^2} = Ae^{a_2x^2}$$

Из условия нормы  $\int_{-\infty}^{\infty} P(x)dx = 1$  видно, что  $a_2 < \infty$ .

$P(x) = Ae^{-\frac{x^2}{2\sigma^2}}$  - нормальный закон Гаусса. Следовательно  $a_2 = -\frac{1}{2\sigma^2}$

**Chapter. Числовые характеристики случайных величин.**

## 72 Математическое ожидание (среднее значение).

$$M\xi = \bar{\xi} = \begin{cases} \sum x_k p_k, & \text{дискретные величины;} \\ \int_{-\infty}^{\infty} x P_{\xi}(x) dx, & \text{непрерывные величины.} \end{cases}$$

Статистическая интерпретация:

Пусть имеется  $\xi_1, \xi_2, \dots, \xi_n$  - различные наблюдения.

$$\frac{1}{n} \sum_{j=1}^n \xi_j = \frac{1}{n} \sum_k x_k \nu_k = \sum_k x_k \frac{\nu_k}{n} \approx \sum x_k p_k$$

при больших  $n$  ( $\frac{\nu_k}{n} \sim p_k$ )

Механическая интерпретация:

$$\sum m_k = m \quad p_k = \frac{m_k}{m}$$

$$\frac{\sum m_k x_k}{\sum m_k} \quad \text{положение центра масс стержня неизменной погонной плотности}$$

$$\frac{\int_a^b x \rho(x) dx}{\int_a^b \rho(x) dV}$$

## 73 Математическое ожидание и моменты суммы случайной величины.

$$Mf(\xi, \eta) \int \int (x+y)p(x,y) dx dy$$

Пусть  $\zeta = \xi + \eta$  - новая случайная величина. Тогда

$$M(\xi + \eta) = \int \int (x+y)p_{\xi\eta}(x,y) dx dy = \int x \left[ \int p_{\xi\eta}(x,y) dy \right] dx + \int y \left[ \int p_{\xi\eta}(x,y) dx \right] dy =$$

$$= \int x p_{\xi}(x) dx + \int y p_{\eta}(y) dy = M\xi + M\eta$$

мы просто подтвердили свойство мат. ожидания, т.е., что  $p_{\xi\eta}(x, y)$  согласована с этими свойствами. Можно совместить плотностью и не пользоваться:

$$M(\xi + \eta)^n = M \sum_{k=0}^n \binom{n}{k} \xi^k \eta^{n-k} = \sum_{k=0}^n \binom{n}{k} M \xi^k \eta^{n-k} = \sum_{k=0}^n \binom{n}{k} \underbrace{\overline{\xi^k \eta^{n-k}}}_{\text{смешанные моменты } \xi \text{ и } \eta}$$

Если  $\xi$  и  $\eta$  - независимы,  $\overline{\xi^k \eta^{n-k}} = \overline{\xi^k} \overline{\eta^{n-k}}$

$$\int \int x^k y^{n-k} p_{\xi\eta}(x, y) dx dy$$

## 74 Дисперсия суммы, ковариация.

$$D(\underbrace{\xi + \eta}_{\zeta}) = M(\zeta - \bar{\zeta})^2 = M(\xi + \eta - \bar{\xi} - \bar{\eta})^2 = M\{(\xi - \bar{\xi})^2 + (\eta - \bar{\eta})^2 + 2(\xi - \bar{\xi})(\eta - \bar{\eta})\} =$$

$$= M(\xi - \bar{\xi})^2 + M(\eta - \bar{\eta})^2 + 2M\{(\xi - \bar{\xi})(\eta - \bar{\eta})\} = D\xi + D\eta + 2\text{cov}(\xi, \eta)$$

причем  $\bar{\zeta} = M(\xi + \eta) = M\xi + M\eta = \bar{\xi} + \bar{\eta}$

Смешанный центральный момент  $M(\xi - \bar{\xi})(\eta - \bar{\eta}) = \text{cov}(\xi, \eta)$  - называется *ковариацией*.

Если  $\xi = \eta$ , то  $D(2\xi) = 2D\xi + 2D\xi = 2D\xi$ , ковариация должна быть преобразована подобно дисперсии:

$$\text{cov}(\xi, \eta) = \overline{\xi\eta} - \bar{\xi}\bar{\eta}$$

Если  $\xi$  и  $\eta$  - независимы, то  $\overline{\xi\eta} = \bar{\xi}\bar{\eta}$  и  $\text{cov}(\xi, \eta) = 0$

отметим свойство ковариации:  $\text{cov}(a\xi, b\eta) = ab\text{cov}(\xi, \eta)$

### СВОЙСТВА ДИСПЕРСИИ:

1.  $D\xi \geq 0, Dc = 0$
2.  $D(c\xi) = c^2 D\xi$
3.  $D(c + \xi) = D\xi$
4.  $D(\xi + \eta) = D\xi + D\eta + 2\text{cov}(\xi, \eta)$

## 75 Коэффициент корреляции.

$$D(a\xi + b\eta) = a^2 D\xi + b^2 D\eta + 2ab\text{cov}(\xi, \eta) \geq 0$$

учитывая условие, что  $Ax^2 + Bx + C \geq 0$ , тогда и только тогда, когда  $B^2 - 4AC \leq 0$ , находим:

$$B^2 - 4AC = 4b^2 \text{cov}^2(\xi, \eta) - 4D\xi b^2 D\eta \leq 0$$

$$|\text{cov}(\xi, \eta)| \leq \sqrt{D\xi D\eta}$$

$$\rho(\xi, \eta) = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi D\eta}} \quad \text{коэффициент корреляции.}$$

СВОЙСТВА:

1.  $-1 \leq \rho(\xi, \eta) \leq 1$
2. Если  $\xi$  и  $\eta$  независимы,  $\rho(\xi, \eta) = 0$ ,  $\text{cov}(\xi, \eta) = \overline{\xi\eta} - \bar{\xi}\bar{\eta}$
3. Если  $\eta = a + b\xi$ ,  $|p(\xi, \eta)| = 1$

Доказательство:

$$\begin{aligned}
 D\eta &= B^2 D\xi, & M\eta &= a + b\bar{\xi} \\
 M\xi\eta &= M(a\xi + b\xi^2) = a\bar{\xi} + b\bar{\xi}^2 = a\bar{\xi} + bD\xi + b\bar{\xi}^2 \\
 \text{cov}(\xi, \eta) &= a\bar{\xi} + bD\xi + b\bar{\xi}^2 - \bar{\xi}a - b\bar{\xi}^2 \\
 \rho &= \frac{bD\xi}{|b|D\xi} = \pm 1
 \end{aligned}$$

Но если  $\rho = 0$ , это значит, что случайные величины независимы. Пусть  $\xi$  - симметричная случайная величина и  $\eta = \xi^2$ , тогда получаем, что  $\text{cov} = \bar{\xi}^3 - \bar{\xi}^2\bar{\xi} = 0$ .

## 76 Распределение суммы случайной величины.

Пусть задана случайная величина, как сумма двух других случайных величин  $\zeta = \xi + \eta$ , тогда для нее справедливо:

$$P_\zeta(z) = F'_\zeta(z)$$

$$F_\zeta(z) = P\{\zeta < z\} = \int_{-\infty}^z P_{\zeta'} dz'$$

С другой стороны:

$$\begin{aligned}
 P\{\zeta < z\} &= P\{\xi, \eta < z\} = \int_{x+y < z} P_{\xi\eta}(x, y) dx dy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} dy P_{\xi\eta}(x, y) \\
 \int_{-\infty}^z P_\zeta(z') dz' &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} dy P_{\xi\eta}(x, y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^z dy' P_{\xi\eta}(x, y' - x)
 \end{aligned}$$

здесь была произведена замена  $y' = y + x$

Дифференцируем

$$P_\zeta(z) = \int_{-\infty}^{\infty} dx \frac{d}{dz} \int_{-\infty}^z dy' P_{\xi\eta}(x, y' - x) = \int_{-\infty}^{\infty} dx P_{\xi\eta}(x, z - x)$$

Для независимых случайных величин:

$$P_\zeta(z) = \int_{-\infty}^{\infty} dx P_\xi(x) P_\eta(z - x)$$

Для неотрицательных независимых случайных величин:

$$P_\zeta(z) = \int_0^z dx P_\xi(x) P_\eta(z - x)$$

## 77 Функция от случайной величины.

1. Пусть на  $[a, b]$  задана случайная величина  $\xi$  с плотностью распределения  $P_\xi(x)$  и монотонно возрастающая функция  $f(x)$  с областью значений  $[c, d]$ . Случайная величина  $\eta = f(\xi)$  называется функцией случайной величины  $\xi$ .

найдем плотность функции случайной величины:

$$F_\eta(y) = P\{\eta < y\} = P\{f(\xi) < y\} = P\{\xi < f^{-1}(y)\} = F_\xi(f^{-1}(y));$$

$$P_\eta(y) = P_\xi(f^{-1}(y)) \frac{df^{-1}(y)}{dy} = \frac{P_\xi(f^{-1}(y))}{df(y)/dy}$$

2. Если убывает, то

$$P\{f(\xi) < y\} = P\{\xi > f^{-1}(y)\} = 1 - F_\xi(f^{-1}(y))$$

$$P_\eta(y) = -P_\xi(f^{-1}(y)) \frac{df^{-1}}{dy} = -\frac{P_\xi(f^{-1}(y))}{df/dy} = \frac{P_\xi(f^{-1}(y))}{|df/dy|}$$

## 78 Математическое ожидание.

$$M\eta = \int y P_\eta(y) dy = \int y P_\xi(f^{-1}(y)) \frac{dy}{df/dy} = \int f(x) P_\xi(x) dx$$

с учетом  $x = f^{-1}(y)$  и  $y = f(x)$ , получаем

$$= \int f(x) P_\xi(x) dx$$

Обобщение на многомерные случайные величины:

$$M\eta = \int \dots \int f(x_1, \dots, x_n) P(x_1, \dots, x_n) dx$$

**Chapter. Характеристические функции.**

## 79 Интегральное представление $\delta$ -функции.

Рассмотрим функцию

$$\frac{1}{2\pi} \int_{-A}^A e^{ikx} dk \equiv \Psi_A(x) = \frac{1}{\pi} \frac{\sin Ax}{x}$$

Эта функция обладает следующими свойствами:

1. симметрична
2. в точке  $k = 0$  функция принимает значение  $\Psi_A(0) = \frac{A}{\pi}$
3. в других точках функция осциллирует с убывающей асимптотикой.
4.  $\int_{-\infty}^{\infty} \Psi_A(x) dx = 1$



Рассмотрим

$$\int_{-\infty}^{\infty} f(x) \Psi_A(x) dx \longrightarrow f(0) \int_{-\infty}^{\infty} \Psi_A(x) dx = f(0)$$

при  $A \rightarrow \infty$  осцилляции зарежут  $f(k)$  везде где  $k < 0$ .

Итак, при  $A \rightarrow \infty$   $\Psi_A(x) \rightarrow \Psi_{\infty}(x)$  со свойствами:

1. симметричность
2. равенство  $\infty$  при  $k = 0$ .
3.  $\int_{-\infty}^{\infty} \Psi_{\infty}(x) dx = 1$
4.  $\int_{-\infty}^{\infty} f(x) \Psi_{\infty}(x) dx = f(0)$
5.  $\delta(\alpha x) = \frac{\delta(x)}{|\alpha|}$

Дельта функция:

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\pm ik(x-x')} dk$$

## 80 Характеристическая функция.

Характеристическая функция:

$$\varphi_{\xi}(k) = M e^{ik\xi} = \int_{-\infty}^{\infty} e^{ikx} p_{\xi}(x) dx$$

$$P_{xi}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ikx} \varphi_{\xi}(k) dk \quad \text{преобразование Фурье}$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} dk e^{-ikx} \int_{-\infty}^{\infty} e^{ikx'} p_{\xi}(x') dx' = \int_{-\infty}^{\infty} dx' p_{\xi}(x') \underbrace{\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ik(x-x')} dk}_{\delta(x-x')} = p_{\xi}(x)$$

## 81 Свойства характеристической функции.

1. Наличие взаимно однозначного соответствия между характеристической функцией и функцией распределения.
2.  $\varphi_{\xi}(k)$  определена при любом  $k \in (-\infty, \infty)$
3.  $\varphi_{\xi}(0) = 1, |\varphi_{\xi}(k)| \leq 1$
4.  $\varphi_{a\xi+b}(k) = e^{ikb} \varphi_{\xi}(ak)$
5.  $\varphi_c(k) = e^{ikc}$
6.  $\varphi_{\xi+\eta}(k) = \varphi_{\xi}(k) \varphi_{\eta}(k)$ , если  $\xi$  и  $\eta$  независимы.

## 82 Характеристические функции и моменты.

$$\begin{aligned}\varphi'_\xi(k) &= M(i\xi)e^{ik\xi} & \varphi'_\xi(0) &= i\bar{\xi} \\ \varphi''_\xi(k) &= (i\xi)^2 e^{ik\xi} & \varphi''_\xi(0) &= -\bar{\xi}^2 \\ &\dots\dots\dots \\ \varphi_\xi^{(n)}(k) &= M(i\xi)^n e^{ik\xi} & \varphi_\xi^{(n)}(0) &= i^n \bar{\xi}^n\end{aligned}$$

Разложение характеристической функции в ряд:

$$\varphi_\xi(k) = \varphi_\xi(0) + \varphi'_\xi(0)k + \frac{\varphi''_\xi(0)}{2}k^2 + \dots = 1 + i\bar{\xi}k - \frac{\bar{\xi}^2}{2}k^2 + \dots$$

## 83 Характеристическая функция нормального распределения.

$$\varphi_\xi(k) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{ikx - \frac{x^2}{2\sigma}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \cos kx dx$$

Дифференцируем по  $k$ :

$$\begin{aligned}\varphi'_\xi(k) &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \sin kx dx = \text{по частям} = -k\varphi_\xi(k) \\ \frac{d\varphi}{dk} &= -k\varphi, & \varphi(0) &= 1 \\ \frac{d\varphi}{\varphi} &= -kdk\end{aligned}$$

Следовательно

$$\varphi_\xi(k) = e^{-\frac{\sigma^2 k^2}{2}}$$

легко запомнить, так как это распределение имеет такой же вид как и плотность (только умножается на  $\frac{1}{\sqrt{2\pi}}$ ). **Chapter. Последовательности независимых случайных величин.**

## 84 Закон больших чисел.

**Теорема** Пусть заданные случайные величины  $\xi_1, \dots, \xi_n$  - независимы, одинаково распределены и  $M\xi = a$ . Тогда, при любом  $\varepsilon > 0$  существует

$$\lim_{n \rightarrow \infty} P\left\{\left|\underbrace{\xi_1, \dots, \xi_n}_n - a\right| < \varepsilon\right\} = 1$$

Другими словами:

$$\frac{S_n}{n} \rightarrow a$$

доказательство:

$$\varphi_{\frac{S_n}{n}}(k) = \varphi_{S_n}\left(\frac{k}{n}\right) = \varphi_\xi^n\left(\frac{k}{n}\right) = \left\{\left(1 + \frac{1}{N}\right)^N\right\}^{i\bar{\xi}k} \rightarrow e^{i\bar{\xi}k} = e^{iak}$$

## 85 Центральная предельная теорема.

Обозначим функцией распределения стандартную нормальную случайную величину  $\eta_0(0, 1)$  через  $\Phi(x)$ , такую что:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx$$

по свойствам функции распределения  $F_{a+\sigma\eta_0}(x) = \Phi\left(\frac{x-a}{\sigma}\right)$ . Обратно  $\Phi\left(\frac{x-a}{\sigma}\right)$  - функция распределения нормальной случайной величины со средним  $a$  и дисперсией  $\sigma^2$ .

**ЦПТ** Если случайные величины  $\xi_1, \dots, \xi_n$  независимы, одинаково распределены и имеют конечные дисперсии  $D\xi = \sigma^2$ , то распределение их суммы  $S_n = \sum_{j=1}^n \xi_j$  при большом числе слагаемых является нормальным со средним  $MS_n = na$  и дисперсией  $n\sigma^2$  независимо от того, как распределены определенные слагаемые:

$$F_{S_n}(x) \sim \Phi\left(\frac{x-na}{\sigma\sqrt{n}}\right), \quad n \rightarrow \infty$$

Доказательство:

$$S_n - na = \dot{S}_n = \sum_{j=1}^n \dot{\xi}_j$$

$$\varphi_{\dot{\xi}}(k) \sim 1 - \sigma^2 k^2 / 2 \quad \text{при} \quad k \rightarrow 0$$

$$\frac{\varphi_{\dot{S}_n}(k)}{\sigma\sqrt{n}} = \varphi_{\dot{S}_n}\left(\frac{k}{\sigma\sqrt{n}}\right) = \left[\varphi_{\dot{\xi}}\left(\frac{k}{\sigma\sqrt{n}}\right)\right]^n \sim \left(1 - \frac{k^2}{2n}\right)^n \sim e^{-\frac{k^2}{2}}$$

- характеристическая функция нормального распределения  $\Phi(x)$ .

## 86 Теорема Муавра-Лапласа

Первоначально ЦПТ была доказана для частного случая - схемы Бернулли,  $\xi = 1$  с вероятностью  $p$  и равной нулю с вероятностью  $q$ ,  $a = M\xi = p$ ,  $\sigma^2 = D\xi = pq$ ,  $S_n = \nu$  (число "успехов")

$$\text{ТМЛ } F_{\nu}(x) \sim \Phi\left(\frac{x-np}{\sqrt{npq}}\right), \quad n \rightarrow \infty$$

$$\text{Следствие 1 } \sum_{k=0}^{[x]} \binom{n}{k} p^k q^{n-k} \sim \Phi\left(\frac{x-np}{\sqrt{npq}}\right), \quad n \rightarrow \infty$$

$$\text{Следствие 2 } \sum_{k=0}^{[x]} \frac{a^k}{k!} e^{-a} \sim \Phi\left(\frac{x-a}{\sqrt{a}}\right), \quad n \rightarrow \infty$$

## 87 Устойчивость нормального закона.

Теперь пусть  $\eta_1$  и  $\eta_2$  - стандартные нормальные случайные величины ( $M\eta = 0$ ,  $D\eta = 1$ ).

Рассмотрим новую случайную величину  $\zeta = \frac{\eta_1 + \eta_2}{c}$  такую, что:

$$\varphi_{\zeta}(k) = \varphi_{\eta_1 + \eta_2}(k/c) = [\varphi_{\eta}(k/c)]^2 = e^{-\frac{1}{2} \frac{k^2}{c^2}} = e^{-\frac{k^2}{2}}$$

если  $c^2 = 2$ , т.е.  $c = \sqrt{2}$ .

Таким образом  $\frac{\eta_1 + \eta_2}{\sqrt{2}} =^d \eta$

$\frac{\eta_1 + \eta_2 + \dots + \eta_n}{n^{1/2}} =^d \eta$  - свойство устойчивости: сумма независимых нормально распределенных случайных величин является нормальной случайной величиной.

Вообще, если  $\frac{\xi_1 + \dots + \xi_n}{n^{1/2}} =^d \xi$ , т.е. случайная величина  $\xi$  называется устойчивой.

## 88 Характеристическая функция симметричного устойчивого закона.

1. Характеристическая функция симметричного закона зависит лишь от модуля  $|k|$ :

$$\varphi(k) = \int_{-\infty}^{\infty} e^{ikx} p(x) dx = - \int_{-\infty}^{\infty} e^{-ikx} p(x) dx = \int_{-\infty}^{\infty} e^{-ikx} p(x) dx = \varphi(-k)$$

2.  $\varphi^2\left(\frac{k}{c}\right) = \varphi(k)$

$$\ln \varphi(k) = -\psi(k)$$

$$2\psi\left(\frac{k}{c}\right) = \psi(k) \text{ этому уравнению удовлетворяет}$$

$$\begin{cases} \psi(k) = |k|^\alpha \\ c = 2^{1/\alpha}, \end{cases} \quad \alpha \in [0, 2] \quad \text{характеристические показатели устойчивого закона}$$

$$\varphi(k) = e^{-|k|^\alpha} \quad \text{устойчивый закон}$$

При  $\alpha = 1$ , получаем

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ikx - |k|} dk = \frac{1}{\pi(1+x^2)} \quad \text{закон Коши}$$

$c = 2$

$\frac{\eta_1 + \eta_2}{2} =^d \eta$  - ЦПТ не работает.

## 89 Многократные свертки распределений.

$$P^{*n}(x) = \underbrace{p(x) * \dots * p(x)}_n = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_{n-1} p(x - x_1 - \dots - x_{n-1})$$

$$\varphi(k) \rightarrow [\varphi(k)]^n$$

$$P^{*n_1}(x) * P^{*n_2}(x) = P^{*n_1+n_2}$$

$$P^{*n-1}(x) * P(x) = \int_{-\infty}^{\infty} P^{*n-1}(x - x') P(x') dx'$$

Если  $P(x) = P_\xi(x)$ , то  $P^{*n}(x) = P_{S_n}(x)$ ,  $S_n = \xi_1 + \dots + \xi_n$

Если  $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ ,  $P^{*n}(x) =$

Если  $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$ ,  $P^{*n}(x) =$

Если  $p(x) = g(x; \alpha)$ ,  $P^{*n}(x) = n^{-1/\alpha} p(xn^{-1/\alpha})$

Воспроизводимые распределения

$\varphi(k; \theta) \rightarrow \varphi(k; n\theta)$

$P^{*n}(x; \theta) \rightarrow p(x, n\theta)$

В том числе воспроизводящих:

Биномиальные -  $\varphi(k) = (q + e^{ik}p)^n$

Распределения Пуассона -  $\varphi(k) = e^{-\mu}(1 - e^{ik})^\mu$

Гамма распределение -  $\varphi(k) = \left(\frac{\mu}{\mu - ik}\right)^\nu$

ЦПТ:

$$P^{*n}(x) \sim \frac{1}{\sqrt{2\pi n\sigma}} e^{-\frac{(x-na)^2}{2n\sigma^2}}, \quad n \rightarrow \infty$$

$$a = \int_{-\infty}^{\infty} xp(x)dx$$

$\sigma^2$  - дисперсия  $p(x)$ .

**chapter. Специальные распределения.**

## 90 Многомерное нормальное распределение.

$$\vec{\xi} = \xi_1, \dots, \xi_n$$

$$\vec{x} = x_1, \dots, x_n$$

$$d\vec{x} = dx_1, \dots, dx_n$$

$$P_{\vec{\xi}}(\vec{x}) = P_{\xi_1, \dots, \xi_n}(x_1, \dots, x_n)$$

**a)** независимое нормальное распределение

$$P_{\xi_1}(x_1) \dots P_{\xi_n}(x_n) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x_1-a_1)^2}{2\sigma_1^2}} \dots \frac{1}{\sigma_n \sqrt{2\pi}} e^{-\frac{(x_n-a_n)^2}{2\sigma_n^2}}$$

**b)** Одинаково распределенные (сферическое нормальное распределение)

$$= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n x_k^2}$$

это вырожденные распределения:

поверхности уровня в случае b) - сферические, в случае a) - эллипсоиды, главные оси которых направлены вдоль осей координат.

Если повернуть оси, то получим:

$$P_{\vec{\xi}}(x_1, \dots, x_n) = \frac{\sqrt{\det ||a_{ij}||}}{(2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i,j=1}^n a_{ij} x_i x_j}$$

## 91 Гамма-распределения.

$$n! = \int_0^{\infty} e^{-t} t^n dt \rightarrow \int_0^{\infty} e^{-t} t^{\nu} dt = \nu! \equiv \Gamma(\nu + 1) \quad \text{все } n - \text{целые}$$

Так, что

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt \quad - \text{ гамма функция}$$

Пусть  $\xi_1, \xi_2, \dots$  - независимые случайные величины.

$$f_{\xi_1}(x) = \mu e^{-\mu x}$$

$$f_{\xi_1+\xi_2}(x) = \mu \mu x e^{-\mu x}$$

$$\dots\dots\dots f_{\xi_1+\dots+\xi_n} = \mu \frac{(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} \equiv g_{\mu n}(x) - \text{ формула Эрланга.}$$

$$\underbrace{\xi_1 + \dots + \xi_{n_1}}_{g_{\mu, n_1}} + \underbrace{\xi'_1 + \dots + \xi'_{n_2}}_{g_{\mu, n_2}} \text{ имеет плотность } g_{\mu, n_1} * g_{\mu, n_2} = g_{\mu, n_1+n_2}(x) - \text{ теорема}$$

**сложения.**

Последняя формула похожа на устойчивость, но - нет! Чтобы понять, что это действительно так, надо получить распределение с другими параметрами.

До этого мы рассматривали только целое значение индекса  $n$ , теперь сделаем обобщение на нецелое значение:

$$g_{\mu, \nu}(x) = \mu \frac{(\mu x)^{\nu-1}}{\Gamma(\nu)} e^{-\mu x}$$

## 92 Бета-распределение.

Пусть  $\xi_1$  и  $\xi_2$  независимы и имеют гамма-распределение с параметрами  $(1, \nu_1)$  и  $(1, \nu_2)$ .

$$\xi = \frac{\xi_1}{\xi_1 + \xi_2} \in (0, 1)$$

$$F_{\xi}(x) = \int_0^{\infty} \int_0^{\infty} u \left( x - \frac{x_1}{x_1 + x_2} \right) \frac{x_1^{\nu_1-1} x_2^{\nu_2-1}}{\Gamma(\nu_1) \Gamma(\nu_2)} e^{-x_1-x_2} dx_1 dx_2$$

$$F'_{\xi}(x) = \int_0^{\infty} \int_0^{\infty} \delta \left( x - \frac{x_1}{x_1 + x_2} \right) \dots$$

$$\delta(f(x_2)) = \frac{\delta(x_2 - x_2^{(0)})}{|f'(x_2^{(0)})|}$$

$$f(x_2^{(0)}) = 0$$

$$x - \frac{x_1}{x_1 + x_2^{(0)}} = 0$$

$$x_2^{(0)} = x_1 \left( \frac{1}{x} - 1 \right)$$

$$f'(x_2^{(0)}) = x^2/x_1$$

$$\begin{aligned} P_\xi(x) &= \int_0^\infty \int_0^\infty \frac{\delta(x_2 - x_1(1/x - 1))}{x^2/x_1} \frac{x_1^{\nu_1-1} x_2^{\nu_2-1}}{\Gamma(\nu_1)\Gamma(\nu_2)} e^{-x_1-x_2} dx_1 dx_2 = \\ &= \frac{(1/x - 1)^{\nu_2-1}}{\Gamma(\nu_1)\Gamma(\nu_2)x^2} \int_0^\infty x_1^{\nu_1+\nu_2-1} \underbrace{e^{-x_1-x_1(1/x-1)}}_{e^{-x_1/x}=e^{-t}, t=x_1/x} dx_1 = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma_1(\nu_1)\Gamma(\nu_2)} x^{-2+\nu_1+\nu_2} (1/x - 1)^{\nu_2-1} = \\ &= \frac{1}{B(\nu_1, \nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1} \quad 0 < x < 1 \\ P_\xi(x) &= \frac{1}{B(\nu_1, \nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1} \quad \textbf{бета-распределение} \end{aligned}$$

$k$ -мерным аналогом бета-распределения является распределение Дирихле:

$$\begin{aligned} \xi &= \frac{\xi_1}{\xi_1 + \xi_2 + \dots + \xi_{k+1}}; \\ P_\xi(x_1, \dots, x_k) &= \frac{\Gamma(\nu_1 + \dots + \nu_{k+1})}{\Gamma(\nu_1) \dots \Gamma(\nu_{k+1})} x_1^{\nu_1-1}, \dots, x_k^{\nu_k-1} (1 - x_1 - \dots - x_k)^{\nu_{k+1}-1} \end{aligned}$$

в любой точке симплекса:  $\sum_{i=1}^k x_i \leq 1$  и равняется нулю за его пределами.

## 93 $\chi^2$ -распределение.

Пусть  $\eta_i$  - независимые стандартные нормальные случайные величины.

$$P_\eta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Обозначим  $\chi_n^2 = \sum_{i=1}^n \eta_i^2$  (квадрат длины изотропного нормального вектора в  $\mathbf{R}^n$  или сумма независимых квадратов.)

Сначала найдем  $P_{\chi_1^2}(x) = P_{\eta^2}(x)$ :

$$F_{\eta^2}(x) = P(\eta^2 < x) = P(-\sqrt{x} < \eta < \sqrt{x}) = 2P(0 < \eta < \sqrt{x}) = 2 \int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

$$P_{\eta^2}(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x}{2}} \frac{1}{2} x^{-1/2} = \frac{1}{2} \frac{(x/2)^{\frac{1}{2}-1}}{\Gamma(\frac{1}{2})} e^{-\frac{x}{2}} = f_{\frac{1}{2}, \frac{1}{2}}(x)$$

причем  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ .

$$P_{\chi_2^2}(x) = P_{\eta_1^2 + \eta_2^2}(x) = g_{\frac{1}{2}, \frac{1}{2}}(x) = g_{\frac{1}{2}, \frac{1}{2} + \frac{1}{2}}(x)$$

$$P_{\chi_n^2}(x) = g_{\underbrace{\frac{1}{2}, \frac{1}{2} + \dots + \frac{1}{2}}_n}(x) = g_{\frac{1}{2}, \frac{n}{2}}(x) = \frac{1}{2} \frac{(x/2)^{n/2-1}}{\Gamma(n/2)} e^{-\frac{x}{2}} \equiv k_n(x)$$

Таким образом, мы получили частный случай гамма-распределения. Здесь  $\chi_n^2$  - плотность распределения,  $n$  - число степеней свободы.

Теорема сложения:  $K_{n_1+n_2}(x) = K_{n_1}(x) * K_{n_2}(x)$

## 94 Распределение Фишера ( $F$ -распределение Фишера).

Пусть случайные величины  $\eta_1, \dots, \eta_m$  и  $\zeta_1, \dots, \zeta_n$  независимы и нормальны  $(0, \sigma)$ . Обозначим

$$\eta = \sum_{i=1}^m \frac{\eta_i^2}{m} \quad \zeta = \sum_{i=1}^n \frac{\zeta_i^2}{n} \quad \xi = \frac{\eta}{\zeta}$$

$$F_\xi(x) = P(\eta|\zeta < x) = \int_0^\infty \int_0^\infty P_{\eta\zeta}(y, z) u(x - y|z) dy dz$$

$$\begin{aligned} P_\xi(x) = F'_\xi(x) &= \int_0^\infty \int_0^\infty P_\eta(y) P_\zeta(z) \delta(x - y|z) dy dz = \int_0^\infty \int_0^\infty \frac{m}{\sigma^2} K_m\left(\frac{my}{\sigma^2}\right) \frac{n}{\sigma^2} K_n\left(\frac{nz}{\sigma^2}\right) z \delta(y - x|z) dy dz = \\ &= \frac{mn}{\sigma^4} \int_0^\infty K_m\left(\frac{mxz}{\sigma^2}\right) K_n\left(\frac{nz}{\sigma^2}\right) z dz = \frac{mn}{\sigma^4} \int_0^\infty \frac{(mxz)^{m/2-1}}{\sigma^{m-1} 2^{m/2} \Gamma(\frac{m}{2})} \frac{(nz)^{n/2-1}}{2^{n/2} \sigma^{n-2} \Gamma(\frac{n}{2})} e^{-\frac{mxz+nz}{2\sigma^2}} z dz = \\ &= \frac{m^{m/2} n^{n/2}}{\sigma^m} \frac{1}{2^{(m+n)/2}} \frac{1}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \int_0^\infty x^{m/2-1} z^{(m+1)/2-1} e^{-\frac{(mx+n)z}{2\sigma^2}} dz = \end{aligned}$$

Переобозначая  $\frac{mx+n}{2\sigma^2} z = t$ ,  $dz = \frac{2\sigma^2 dt}{mx+n}$ ,  $z = \frac{2t\sigma^2}{mx+n}$  получаем

$$= \frac{m^{m/2} n^{n/2} x^{m/2-1}}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \frac{\Gamma(\frac{m+n}{2})}{(mx+n)^{(m+n)/2}}$$

В качестве частного случая можно рассмотреть  $m = 1$ .

В результате получим плотность распределения Фишера:

$$f_{mn}(x) = \frac{m^{m/2} n^{n/2}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}$$

Плотность распределения Фишера не зависит от  $\sigma$ .

## 95 Распределение Стьюдента.

В. Госсет писал под псевдонимом Student.

Положим, что имеется  $n + 1$  случайных величин  $\eta, \eta_1, \dots, \eta_n$ , независимых и  $(0, \sigma)$  нормальные. Положим

$$\zeta = \sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i^2}$$

и рассмотрим

$$\tau = \frac{\eta}{\zeta} = \frac{\eta}{\sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i^2}} =^d \sqrt{n} \frac{\eta}{\chi_n}$$

Найдем  $P_\tau(x)$

$$F_\tau(x) = P(\tau < x) = \int \int_{y/z < x} P_{\eta\zeta}(y, z) dy dz = \int_0^\infty \int_0^\infty P_{\eta\zeta}(y, z) u(x - y|z) dz dy$$



Совместная плотность равна произведению:

$$P_{\eta\zeta}(y, z) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{y^2}{2\sigma^2}} P_{\zeta}(z)$$

$$P_{\eta\zeta} = C_n e^{-\frac{y^2 + nz^2}{2\sigma^2}} z^{n-1}$$

$$n = \sqrt{\frac{2}{\pi}} \frac{(n/2)^{n/2}}{\Gamma(\frac{n}{2}) \sigma^{n+1}}$$

$$\frac{2nz}{\sigma^2} K_n \left( \frac{nz^2}{\sigma^2} \right) = \frac{2 \left( \frac{n}{2} \right)^{n/2} z^{n-1}}{\sigma^n \Gamma(\frac{n}{2})} e^{-\frac{nz^2}{2\sigma^2}}$$

Запишем свойства дельта-функции:

$$\delta(\alpha x) = \delta(x)$$

$$\delta(x) = \delta(-x)$$

$$P_{\tau}(x) = F'_{\tau}(x) = \int_0^{\infty} \int_0^{\infty} P_{\eta\zeta}(y, z) \delta(x - y|z) dy dz =$$

Учитывая, что  $\delta((xz - y)|z) = z\delta(y - xz)$ , получаем

$$= \int_0^{\infty} P_{\eta\zeta}(xz, z) z dz = C_n \int_0^{\infty} e^{-\frac{(xz)^2 + nz^2}{2\sigma^2}} z^n dz = C_n \frac{\sigma^{n+1} 2^{\frac{n+1}{2}}}{(x^2 + n)^{\frac{n+1}{2}}} \int_0^{\infty} e^{-t^2} t^n dt$$

делая переобозначения  $t^2 = s$ ,  $t = s^{1/2}$ ,  $dt = \frac{1}{2} s^{-1/2} ds$ , получаем:

$$P_{\tau}(x) = \frac{1}{2} \sqrt{\frac{2}{\pi}} \frac{(n/2)^{n/2} \Gamma(\frac{n+1}{2})}{\Gamma(n/2) \sigma^{n+1}} \sigma^{n+1} \left[ 1 + \frac{x^2}{n} \right]^{-\frac{n+1}{2}} n^{-\frac{n}{2}} n^{-\frac{1}{2}} 2^{\frac{n+1}{2}} = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left( 1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}} \equiv S_n(x)$$

Отсюда видно, что  $S_n(x)$  не зависит от  $\sigma$ .

## 96 Обобщенная ЦПТ.

Пусть  $\xi_1, \dots, \xi_n$  одинаково распределены, так что

$$P\{\xi > x\} \sigma x^{-\alpha}, \quad x \rightarrow \infty$$

$$P\{\xi < x\} \sim b|x|^{-\alpha}, \quad x \rightarrow -\infty$$

Причем  $a, b > 0$  и  $\alpha < 2$ .

Тогда

$$P\left\{ \frac{S_n - A_n}{B_n} < x \right\} \Rightarrow F_{\alpha, \beta}(x), \quad n \rightarrow \infty$$

где  $S_n = \sum_{i=1}^n \xi_i$ ,  $B_n = B_1 n^{1/\alpha}$ ,  $B_1 = (a + b) \Gamma(1 - \alpha) \cos \frac{\alpha}{2}$

$A_n = na$ , если существует  $M\xi = a$  (т.е. при  $\alpha > 1$ )

$A_n = 0$ , если  $\alpha < 1$

$A_n = (a^2 - b^2)n \ln n$ , если  $\alpha = 1$

$\beta = \frac{a-b}{a+b} \in [-1, 1]$

$F_{\alpha, \beta}(x)$  - устойчивая функция распределения с характеристической функцией с характеристическим показателем  $\alpha$  и параметром асимметрии  $\beta$

$$\varphi(k) = e^{-|k|^{\alpha} \left[ 1 - i\beta \operatorname{tg}\left(\frac{\pi\alpha}{2}\right) \frac{k}{|k|} \right]}, \quad \alpha \neq 1$$

$$e^{-\frac{\pi}{2}|k| - ik\beta \ln |k|}$$

## 97 Система Пирсона.

Система Пирсона - это семейство плотностей, удовлетворяющих уравнению:

$$\frac{dP}{dx} = -\frac{a+x}{b_0+b_1x+b_2x^2}P$$

где  $a$  и  $b_n$  - действительные числа.

называются кривыми Пирсона. Классифицируются в зависимости от уравнения  $b_0 + b_1x + b_2x^2 = 0$ . Семейство состоит из 12 типов и нормального распределения.

Наиболее важны:

**тип I**  $P(x) = k(1+x/a_1)^{m_1}(1-x/a_2)^{m_2}$ ,  $-a_1 \leq x \leq a_2$ ,  $m_1, m_2 \geq -1$  - частный случай бета-распределения.

**тип III**  $P(x) = k(1+x/a)^{\mu}e^{-\mu x}$ ,  $-a \leq x < \infty$ ,  $\mu > 0$ ,  $a > 0$  - частный случай гамма-распределения,  $\chi^2$ -распределение.

**тип VI**  $P(x) = kx^{-q_2}(x-a)^{q_1}$ ,  $a_1 \leq x < \infty$ ,  $q_1 > q_2 - 1$  - частный случай бета-распределения.

**тип VII**  $P(x) = k(1+\frac{x^2}{a^2})^{-m}$ ,  $-\infty < x < \infty$ ,  $m > \frac{1}{2}$  - распределение Стьюдента.

**тип X**  $P(x) = ke^{-(x-m)/\sigma}$ ,  $m \leq x < \infty$ ,  $\sigma > 0$

**тип XI**  $P(x) = kx^{-m}$ ,  $b \leq x < \infty$ ,  $m > 0$

Всякое П. распределение определено своими ???? моментами. Это свойство используется для приближения отношения ??? распределений: находят моменты, определяют тип подходящего П.Р. и находят значение неизвестных параметров.

## 98 Таблица плотностей.

$\eta_i$  - независимые,  $(0, \sigma)$  - нормальные случайные величины.

$$K_n(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(\frac{n}{2})}$$

$\xi$	$P_\xi(x)$
$\chi_n^2 = \sum_{i=1}^n \eta_i^2$	$\frac{1}{\sigma^2} K_n\left(\frac{k}{\sigma^2}\right) = \frac{1}{2^{n/2}\sigma^n\Gamma(\frac{n}{2})} x^{n/2-1} e^{-\frac{x}{2\sigma^2}}$
$\frac{1}{n} \sum_{i=1}^n \eta_i^2$	$\frac{n}{\sigma^2} K_n\left(\frac{nx}{\sigma^2}\right) = \frac{(n/2)^{n/2}}{\sigma^n\Gamma(n/2)} x^{n/2-1} e^{-\frac{nx}{2\sigma^2}}$
$\sqrt{\sum_{i=1}^n \eta_i^2}$	$\frac{2x}{\sigma^2} K_n\left(\frac{x^2}{\sigma^2}\right) = \frac{2}{2^{n/2}\sigma^n\Gamma(\frac{n}{2})} x^{n-1} e^{-\frac{x^2}{2\sigma^2}}$
$\sqrt{\frac{1}{n} \sum_{i=1}^n \eta_i^2}$	$\frac{2nx}{\sigma^2} K_n\left(\frac{nx^2}{\sigma^2}\right) = \frac{2}{2^{n/2}\sigma^n\Gamma(\frac{n}{2})} x^{n-1} e^{-\frac{nx^2}{2\sigma^2}}$

## Часть I

# Математическая статистика.

chapter. Математическое ожидание и дисперсии основных статистик.

## 99 Понятие выборки.

Пусть требуется измерить некоторую величину  $a$ . Число яблок в корзине - 15, следовательно  $a = 15$ . ????. Результат измерения один и тот же. Все ???? тоже, но если будем измерять весами с большей точностью (??? до мили грамм), то будем получать немного разные значения:  $\xi_1, \dots, \xi_N$ . В физических экспериментах число распадов за время  $t$  в радиоактивном образце, число частиц космического излучения, попадающих на детектор за одно и тоже время, число фотонов, приходящих на приемник, число молекул газа в объеме  $\Delta V$  и так далее - флуктуируют. Вместо  $a$  получали  $\xi_1, \dots, \xi_N$ . **Математическая статистика** рассматривает эти значения как реализации случайной величины  $\xi$  со средним значением  $M\xi = a$ .

**Математической моделью, независимых измерений**, проводимых в одинаковых условиях является **случайный вектор**  $(\xi_1, \dots, \xi_N)$ , где  $\xi_i$  независимы и одинаково распределены. Случайный вектор называется также случайной выборкой объема  $N$ .

Функция  $\zeta = f(\xi_1, \dots, \xi_N)$  - случайная величина - называется статистикой.

Примеры: выборочная сумма, выборочное среднее, наибольший элемент выборки, наименьший элемент выборки и т.д.

## 100 Математическое ожидание и дисперсия частоты события.

Выборка  $\xi_1, \xi_2, \xi_3, \dots, \xi_n$

События  $A, B, \dots, C$

Частота

$$\hat{P}(A) = \frac{\nu(A)}{n} = \frac{1}{n} \sum_{j=1}^n 1(A_j, A)$$

$$M\hat{P}(A) = M1(A_j, A)1 * P(A) + 0 * P(\bar{A}) = P(A) \quad \text{несмещенность}$$

$$\begin{aligned} D\hat{P}(A) &= \frac{1}{n^2} n D1(A_j, A) = \frac{1}{n} \left( M1^2(A_j, A) - [M1(A_j, A)]^2 \right) = \frac{1}{n} \left( M1(A_j, A) - [M1(A_j, A)]^2 \right) = \\ &= \frac{1}{n} \left( P(A) - P^2(A) \right) = \frac{1}{n} P(A) [1 - P(A)] \rightarrow 0 \quad \text{состоятельность} \end{aligned}$$

## 101 Математическое ожидание и дисперсия выборочного среднего.

Если  $\xi$  - случайная величина, то ее мат. ожидание (теоретическое среднее)  $M\xi = a$  - детерминированное число равно

$$\sum_{n=1}^{\infty} x_n p_n \quad \text{или} \quad \int_{-\infty}^{\infty} x p_{\xi}(x) dx$$

Но если мы имеем выборку  $\xi_1, \dots, \xi_N$  и больше ничего не знаем, то  $M\xi$  мы вычислить не сможем. Но мы можем вычислить

$$\frac{1}{N} \sum_{i=1}^N \xi_i \equiv \hat{a} \quad \underline{\text{выборочное среднее.}}$$

Вместо  $M\ldots \rightarrow \frac{1}{N} \sum_{i=1}^N$ .

Чем оно отличается от теоретического?

1. это случайная величина
2. ее распределение зависит от  $N$

Найдем ее мат. ожидание и дисперсию.

Мат. ожидание:  $M\hat{a} = M\xi = a$  при любом  $N$ , потому что  $\hat{a}$  называется несмещенной оценкой  $a$ .

Дисперсия:  $D\hat{a} = \frac{1}{N} D\xi = \frac{\sigma^2}{N}$  и  $N \rightarrow \infty$ , поэтому  $\hat{a}$  называют состоятельной оценкой  $a$ .

## 102 Математическое ожидание выборочной дисперсии.

Мат. ожидание  $M\xi = a \rightarrow \hat{a} = \frac{1}{N} \sum_{i=1}^N \xi_i$ . Можно ожидать, что дисперсия будет:

$$D\xi = M(\xi - a)^2 = \begin{cases} \frac{1}{N} \sum_{i=1}^N (\xi_i - a)^2, & ?; \\ \frac{1}{N} \sum_{i=1}^N (\xi_i - \hat{a})^2, & ?. \end{cases}$$

Две разные задачи:

1.  $a$  известно, надо найти  $\hat{D}\xi$ .
2.  $a$  неизвестно, найти  $\hat{a}$  и  $\hat{D}\xi$  по известной выборке  $\xi_1, \dots, \xi_N$

Проверим несмещенность:

1.  $MD\xi = \frac{1}{N} \sum_{i=1}^N (\xi_i - a)^2 = \sigma^2$  - выполняется  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (\xi_i - a)^2$  - несмещенность.
- 2.

$$\begin{aligned} M\hat{\sigma}^2 &= M \frac{1}{N} \sum_{i=1}^N (\xi_i - \hat{a})^2 = \frac{1}{N} \sum_{i=1}^N ((\xi_i - a) - (\hat{a} - a))^2 = M \frac{1}{N} \sum_{i=1}^N \left[ (\xi_i - a) - \frac{1}{N} \sum_{i=1}^N (\xi_i - a) \right]^2 = \\ &= M \frac{1}{N} \sum_{i=1}^N \left[ \eta_i - \frac{1}{N} \sum_{j=1}^N \eta_j \right]^2 = \frac{1}{N} \sum_{i=1}^N \left[ \frac{N-1}{N} \eta_i - \frac{1}{N} \sum_{i \neq j} \eta_j \right]^2 = \\ &= \frac{1}{N} \left( \frac{N-1}{N} \right)^2 N \overline{\eta^2} - 0 + \frac{1}{N} \frac{N-1}{N} N M \eta^2 = \frac{N^2 - 2N + 1 + N - 1}{N^2} \sigma^2 = \frac{N^2 - N}{N^2} \sigma^2 = \\ &= \frac{N-1}{N} \sigma^2 \\ S^2 &= \frac{N-1}{N} \sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (\xi_i - \hat{a})^2 \end{aligned}$$

$\hat{\sigma}^2$  - асимптотическое несмещение.  $\hat{S}^2$  - не смещается при любом  $N$ .

## 103 Дисперсия выборочной дисперсии.

Рассмотрим два случая:

1.  $a$  известно:

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \eta_i^2$$

$\eta_i$  - центральная случайная величина,  $\overline{\eta_i^2} = \sigma^2$

$$D\hat{\sigma}_N^2 = \frac{1}{N^2} ND\eta^2 = \frac{1}{N} (\overline{\eta^4} - \overline{\eta^2}^2) = \frac{1}{N} (\mu_4 - \sigma^4)$$

2.  $a$  неизвестно:

$$S_N^2 = \frac{1}{N-1} \sum_{j=1}^N (\xi_j - \hat{a})^2$$

$$DS_N^2 = \frac{1}{N} \left( \mu_4 - \frac{N-3}{N-1} \sigma^4 \right) \approx \frac{1}{N} (\mu_4 - \sigma^4), \quad N \gg 1$$

Если  $\xi$  - нормальная случайная величина, то

$$\mu_{2n} = \frac{(2\sigma^2)^n}{\sqrt{\pi i}} \Gamma(n + \frac{1}{2})$$

$$\mu_n = \frac{4\sigma^4}{\sqrt{\pi i}} \frac{3}{2} \frac{1}{2} \sqrt{\pi} = 3\sigma^4$$

откуда:

$$D\hat{\sigma}_N^2 \approx DS_N^2 \approx \frac{2\sigma^4}{N}$$

## 104 Распределение выборочной дисперсии.

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \xi_i^2 - \left( \frac{1}{N} \sum_{i=1}^N \xi_i \right)^2$$

Рассмотрим  $N = 2$ :

$$\hat{\sigma}_2^2 = \frac{\xi_1^2 + \xi_2^2}{2} - \left( \frac{\xi_1 + \xi_2}{2} \right)^2 = \frac{\xi_1^2}{4} + \frac{\xi_2^2}{4} - 2 \frac{\xi_1 \xi_2}{4} = \frac{(\xi_1 - \xi_2)^2}{4}$$

Пусть  $\eta$  стандартная нормальная величина  $(0, 1)$ . Тогда

$$\xi = {}^d \sigma \eta$$

$$\xi_1 - \xi_2 = {}^d \xi_1 + \xi_2 = {}^d \sqrt{2} \sigma \eta$$

$$\hat{\sigma}_2^2 = {}^d \frac{2\sigma^2 \eta^2}{4} = \frac{\sigma^2}{2} \eta^2$$

Проверяя, получим

$$\hat{\sigma}_N^2 = {}^d \frac{\sigma^2}{N} \left( \underbrace{\eta_1^2 + \dots + \eta_{N-1}^2}_{\chi_{N-1}^2} \right)$$

Величина  $\frac{N\hat{\sigma}_N^2}{\sigma^2}$  имеет  $\chi^2$ -распределение с  $N - 1$  степенями свободы.

## 105 Распределение относительной погрешности.

$$a = 0 \quad \sigma = 1$$

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^N (\xi_i^2 - \hat{a})^2 = \frac{1}{N-1} \sum_{i=1}^N (\xi_i^2 - 2\xi_i\hat{a} + \hat{a}^2) = \frac{1}{N-1} \left[ \sum_{i=1}^N \xi_i^2 - 2\hat{a} \sum_{i=1}^N \xi_i + N\hat{a}^2 \right] = \\ &= \frac{1}{N-1} \left[ \sum_{i=1}^N \xi_i^2 - N\hat{a}^2 \right] = \frac{1}{N-1} \left[ \sum_{i=1}^N \xi_i^2 - \frac{1}{N} \left( \sum_{i=1}^N \xi_i \right)^2 \right] \end{aligned}$$

Линейное ортогональное преобразование  $\eta_j = \sum_{i=1}^N C_{ji} \xi_i$  (поворот!), причем

$$\eta_1 = \sum_{i=1}^N C_{1i} \xi_i = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i$$

т.е.  $C_{1i} = \frac{1}{\sqrt{N}}$ . Все  $\eta_i$  - нормальные и независимые (зависимость возникает лишь когда  $\xi_i$  имеют разные дисперсии)

Заметим, что

$$\sum_{i=1}^N \xi_i^2 = \sum_{i=1}^N \eta_i^2$$

и

$$\sum_{i=1}^N \xi_i^2 - \frac{1}{N} \left( \sum_{i=1}^N \xi_i \right)^2 = \sum_{i=1}^N \eta_i^2 - \eta_1^2 = \sum_{i=2}^N \eta_i^2 =^d \chi_{N-1}^2$$

так что

$$\tau =^d = \frac{\frac{1}{N} \sum_{i=1}^N \xi_i}{\frac{1}{\sqrt{N-1}} \chi_{N-1}} \sqrt{N} = \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i}{\chi_{N-1}} \sqrt{N-1} = \frac{\eta_1}{\chi_{N-1}} \sqrt{N-1}$$

При  $N \rightarrow \infty$  сходится к норм  $\frac{\eta_1}{\chi_{N-1}} \sqrt{N-1}$  - распределение с  $N-1$  степенями свободы  $\tau_{N-1}$ .

**Chapter. Распределение основных оценок.**

## 106 Распределение частоты событий.

Частота событий  $A$   $\nu(A)/n$  есть ст. оценка вероятности  $P_N(A)$

Выборка  $\xi_1, \dots, \xi_N$ ,  
События  $A, \bar{A}, \dots, \bar{A}$  } . схема Бернулли - биномиальное распределение.

$$P(\nu(A) = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{(np)^k}{k!} e^{-np}$$

$$P(k_1 \leq \nu(A) \leq k_2) = \sum_{k=k_1}^{k_2} \binom{n}{k} p^k (1-p)^{n-k}$$

$$P\left(x_1 \leq \frac{\nu(A)}{n} \leq x_2\right) = \sum_{k=nx_1}^{nx_2} \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi\sigma}} \int_{nx_1}^{nx_2} e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

причем  $x_1 = \frac{k_1}{n}$  и  $x_2 = \frac{k_2}{n}$ ,  $np \gg 1$

$a = np$ ,  $\sigma = np(1-p) \approx np$ , если  $p \ll 1$ .

## 107 Распределение выборочного среднего.

Если мы знаем, что  $M\xi = a$  и  $D\xi = \sigma^2$ , то согласно ЦПТ при больших  $N$  выполняется

$$\frac{S_N - N_a}{\sigma\sqrt{N}} \stackrel{d}{=} \eta$$

-стандартная нормальная случайная величина, т.е.

$$\frac{1}{N} \sum_{i=1}^N \xi_i = a + \frac{\sigma}{\sqrt{N}} \eta$$

$\hat{a} = a + \Delta$ , где  $\Delta$  - случайная ошибка.

$$F_{\hat{a}}(x) = p(\hat{a} < x) = P(a + \Delta < x) = D(\Delta < x - a) = P(\eta < \frac{x-a}{\sigma} \sqrt{N}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-a}{\sigma} \sqrt{N}} e^{-\frac{z^2}{2}} dz$$

$$\begin{aligned} D\hat{a} &= D\Delta = \overline{\Delta^2} \\ P_{\hat{a}}(x) &= \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2 N}{2\sigma^2}} \\ N\hat{a} &= a \quad D\hat{a} = \frac{\sigma^2}{N} \\ \sqrt{\overline{\Delta^2}} &= \frac{\sigma}{\sqrt{N}} \end{aligned}$$

## 108 Распределение выборочной дисперсии.

Мы уже показали в пункте (13.3), что выборочное среднее

$$\hat{a} = \frac{1}{N} \sum_{i=1}^N \xi_i$$

асимптотически ( $N \rightarrow \infty$ ) распределено по нормальному закону с  $M\hat{a} = a$  и  $D\hat{a} = \frac{\sigma^2}{N}$ , где  $a = M\xi$ ,  $\sigma^2 = D\xi$  независимо от вида распределения  $\xi_i$ , только бы  $D\xi_i < \infty$ .

А как распределена выборочная дисперсия

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N (\xi_i - \hat{a})^2?$$

В общем случае - это сложная задача. Самым простым случаем этой задачи является предположение, что  $\xi_i$  распределено по нормальному закону с  $M\xi_i = 0$  и  $D\xi_i = \sigma^2$ , тогда  $\hat{a} \rightarrow 0$ , при  $N \rightarrow \infty$  и

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \xi_i^2$$

Каково распределение  $\xi^2$ ?

$$F_{\xi^2}(x) = P\{\xi^2 < x\} = P\{-\sqrt{x} < \xi < \sqrt{x}\} = 2P\{0 < \xi < \sqrt{x}\} = \frac{2}{\sqrt{2\pi}\sigma} \int_0^{\sqrt{x}} e^{-\frac{x'^2}{2\sigma^2}} dx'$$

$$P_{\xi^2}(x) = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-\frac{x}{2\sigma^2}} \frac{1}{2} x^{-1/2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x}{2\sigma^2}} x^{-1/2} = f_{\frac{1}{2\sigma^2}, \frac{1}{2}}(x)$$

$$P_{\sum_{i=1}^N \xi_i^2}(x) = f_{\frac{1}{2\sigma^2}, \frac{N}{2}}(x) \quad \nu - 1 = -\frac{1}{2}$$

chapter. Интервальные оценки.

## 109 Квантили.

Пусть  $F(x)$  - несобственная функция распределения,  $F(x) = P(\xi < x)$ . Корень уравнения  $F(x) = 1 - \alpha$ ,  $x_\alpha$  называется квантилью распределения  $F(x)$ . Квантиль - это такое значение  $x$ , т которого вероятность  $\alpha$ , естественно,  $1 - \alpha$ , которое может быть превышено случайной величиной  $\xi$  с вероятностью  $\alpha$ :  $P(\xi \geq x) = \alpha$ . Если известно два квантиля  $x_\alpha$  и  $x_{1-\beta}$ ,  $\beta < 1 - \alpha$ , то  $P(x_{1-\beta} < \xi < x_\alpha) = 1 - \alpha - \beta$ . В частности,

$$P(x_{1-\alpha} < \xi < x_\alpha) = 1 - 2\alpha$$

доверительный интервал или доверительная вероятность.

Квантили нормального распределения

$$\Phi(u) = 1 - \alpha$$

$$P(-u_\alpha < \eta < u_\alpha) = 1 - 2\alpha$$

Квантили  $\chi^2$ -распределения:

$$K_n(q_\alpha) = 1 - \alpha$$

Квантили Стюдента

$$S_n(t_\alpha) = 1 - \alpha$$

## 110 Интервальные оценки.

Пусть мы имеем выборку  $\xi_1, \dots, \xi_n$ , знаем  $F_\xi(x) = F(x; \theta)$ , но не знаем значения параметра  $\theta$ . Пусть  $\hat{\theta} = f(\xi_1, \dots, \xi_n)$  - оценка неизвестного параметра, а  $F_{\hat{\theta}}(x) = F_{\hat{\theta}}(x; \theta)$  - теоретическая функция распределения оценки. И пусть она убывающая функция параметра  $\theta$ . Например:  $F(x, \theta) = F(x - \theta)$ .

Квантиль удовлетворяет уравнению  $F(x; \theta) = 1 - \alpha$  следовательно  $x_\alpha = x_\alpha(\theta)$ . Если  $F$  - убывающая функция  $\theta$ , то  $x_\alpha(\theta)$  - возрастающая функция. Зададим малым  $\alpha$  (например:  $\alpha = 0.05$  или  $\alpha = 0.01$ ). При каждом  $\theta$  неравенство

$$P(x_{1-\alpha}(\theta) < \hat{\theta} < x_\alpha(\theta)) = 1 - 2\alpha$$

выполняется с вероятностью близкой к 1.  $\hat{\theta} < x_\alpha(\theta)$  из этого следует, что  $x_\alpha^{-1}(\hat{\theta}) < \theta$  и  $P(x_\alpha^{-1}(\hat{\theta}) < \theta < x_{1-\alpha}^{-1}(\hat{\theta})) = 1 - 2\alpha$ . Обозначим  $x_\alpha^{-1}(\hat{\theta}) = \hat{\theta}_1$  и  $x_{1-\alpha}^{-1}(\hat{\theta}) = \hat{\theta}_2$

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - 2\alpha$$

$(\hat{\theta}_1, \hat{\theta}_2)$  - доверительный интервал,  $1 - 2\alpha$  - доверительная вероятность.



## 111 Интервальные оценки среднего и дисперсии в нормальной выборке.

Пусть имеется  $\xi_1, \dots, \xi_n$  - нормальная  $(a, \sigma)$  выборка.

1.  $a$  неизвестно,  $\sigma$  - известно.

$\hat{a}_n = \frac{1}{n} \sum_{i=1}^n \xi_i$  - нормальная  $(a, \frac{\sigma}{\sqrt{n}})$

$$\frac{\hat{a}_n - a}{\sigma/\sqrt{n}} \stackrel{d}{=} \eta \quad (01)$$

$$P\left(u_{-\alpha} < \frac{\hat{a}_n - a}{\sigma/\sqrt{n}} < u_{\alpha}\right) = 1 - 2\alpha$$

следовательно

$$P\left(\underbrace{\hat{a}_n - u_{\alpha}\sigma/\sqrt{n}}_{\hat{\theta}_1} < a < \underbrace{\hat{a}_n + u_{\alpha}\sigma/\sqrt{n}}_{\hat{\theta}_2}\right) = 1 - 2\alpha$$

2.  $a$  известно,  $\sigma$  - неизвестно.

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2 \stackrel{d}{=} \frac{\sigma^2}{n} \chi_n^2$$

$$\chi_n^2 \stackrel{d}{=} n\hat{\sigma}_n^2/\sigma^2 < q_{\alpha}$$

$$P\left(q_{-\alpha} < \frac{n\hat{\sigma}_n^2}{\sigma^2} < q_{\alpha}\right) = 1 - 2\alpha$$

$$P\left(\underbrace{\sqrt{n\hat{\sigma}_n^2 q_{\alpha}^{-1/2}}}_{\hat{\theta}_1} < \sigma < \underbrace{\sqrt{n\hat{\sigma}_n^2 q_{1-\alpha}^{-1/2}}}_{\hat{\theta}_2}\right) = 1 - 2\alpha$$

3. Оба неизвестны:

Вместо  $u_{-\alpha} < \frac{\hat{a}_n - a}{\sigma/\sqrt{n}} < u_{\alpha}$ , следовательно

$$-t_{\alpha} < \frac{\hat{a}_n - a}{\sigma/\sqrt{n}} < t_{\alpha}$$

$$\underbrace{\hat{a}_n - \frac{t_{\alpha} S}{\sqrt{n}}}_{\hat{\theta}_1} < a < \underbrace{\hat{a}_n + \frac{t_{\alpha} S}{\sqrt{n}}}_{\hat{\theta}_2}$$

## 112 Доверительный интервал для среднего при известной дисперсии (нормальное значение).

Пусть  $\xi_1, \dots, \xi_N$  распределение нормального  $(a, \sigma^2)$  и  $\sigma^2$  известно. Найдём доверительный интервал для  $a$ .

$\hat{a} = \frac{1}{N} \sum_{i=1}^N \xi_i$  - нормальное распределение с  $a$  и  $D\hat{a} = \sigma^2/N$ .

Следовательно  $\frac{\hat{a}-a}{\sigma/\sqrt{N}}$  - нормальное  $\eta(0,1)$

$$P\{-u < \frac{\hat{a}-a}{\sigma/\sqrt{N}} < u\} = 2\frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{x^2}{2}} dx = 2\frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{x^2}{2}} dx - 2\frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-\frac{x^2}{2}} dx = 1 - 2\alpha$$

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-\frac{x^2}{2}} dx$$

это уравнение относительно  $u$ . Задаем  $\alpha$  и находим  $u = u_\alpha$ .

Теперь:

$$P\{-\frac{\sigma}{\sqrt{N}}u_\alpha - \hat{a} < -a < \frac{\sigma}{\sqrt{N}}u_\alpha - \hat{a}\} = P\{\hat{a} - \frac{\sigma}{\sqrt{N}}u_\alpha < a < \hat{a} + \frac{\sigma}{\sqrt{N}}u_\alpha\} = 1 - 2\alpha$$

$$(\hat{\theta}_1, \hat{\theta}_2) = \left(\hat{a} - \frac{\sigma}{\sqrt{N}}u_\alpha, \hat{a} + \frac{\sigma}{\sqrt{N}}u_\alpha\right)$$

## 113 Доверительный интервал для дисперсии при известном среднем (нормальный закон).

Пусть  $\xi_1, \dots, \xi_N$  - нормальная  $(a, \sigma^2)$  выборка и среднее известно. Если среднее известно, то

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \xi_i^2 - a^2$$

будет несмещенной:

$$M\hat{\sigma}_N^2 = \overline{\xi^2} - \bar{\xi}^2$$

Смещенной является оценка

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N \xi_i^2 - \hat{a}^2$$

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^N (\xi_i - a)^2 = \frac{1}{N} \sum_{i=1}^N \eta_i^2 \sigma^2 = \frac{\sigma^2}{N} \sum_{i=1}^N \eta_i^2 \stackrel{d}{=} \frac{\sigma^2}{N} \chi_N^2$$

или

$$\frac{N\hat{\sigma}_N^2}{\sigma^2} \stackrel{d}{=} \chi_N^2$$

Есть таблицы для  $\alpha, q_\alpha$ .

При заданном  $\alpha$  найдем  $c_1$  и  $c_2$ , тогда

$$\begin{aligned} P\{c_1 < \chi_N^2 < c_2\} &= P\{c_1 < \frac{N\hat{\sigma}_N^2}{\sigma^2} < c_2\} = P\{\frac{c_1}{N\sigma_N^2} < \frac{1}{\sigma^2} < \frac{c_2}{N\sigma_N^2}\} = P\{\frac{N\hat{\sigma}_N^2}{c_2} < \sigma^2 < \frac{N\hat{\sigma}_N^2}{c_1}\} = \\ &= 1 - 2\alpha \end{aligned}$$

так что  $\left(\frac{N\hat{\sigma}_N^2}{q_\alpha}, \frac{N\hat{\sigma}_N^2}{q_{1-\alpha}}\right)$  - искомый интервал.

## 114 Метод наибольшего правдоподобия.

Пусть  $P_\xi = P_\xi(x; \theta)$ , как найти  $\hat{\theta}_n(\xi_1, \dots, \xi_n)$ ?

Рассмотрим сначала  $n = 1$ .

При  $\theta = \theta'$  наиболее вероятно  $\xi = x'$

При  $\theta = \theta''$  наиболее вероятно  $\xi = x''$

При  $\theta = \theta'''$  наиболее вероятно  $\xi = x'''$

Обратная формулировка:

При  $\xi = x'$  наиболее вероятно  $\theta = \theta'$

При  $\xi = x''$  наиболее вероятно  $\theta = \theta''$

При  $\xi = x'''$  наиболее вероятно  $\theta = \theta'''$

$P_\xi(x; \theta)$  как функция от  $\theta$  при фиксированном  $x$  называется функцией правдоподобия.

$$L(\theta; x) = P_\xi(x; \theta)$$

Для выборки объема  $n$

$$L(\theta; x_1, \dots, x_n) = P_\xi(x_1; \theta) P_\xi(x_2; \theta) \dots P_\xi(x_n; \theta)$$

По методу наибольшего правдоподобия за оценку параметра  $\theta$  принимается такое значение, при котором  $L$  является наибольшим.

$$l = L_{max} = \ln L = \ln L_{max}$$

$$\frac{\partial \ln L}{\partial \theta} = 0$$

Так как  $\theta = \theta_1, \dots, \theta_k$ , то

$$\frac{\partial \ln L}{\partial \theta_j} = 0$$

## 115 Оценка параметров нормального распределения.

Пусть  $\xi_1, \dots, \xi_n$  имеют нормальное распределение  $(a, \sigma)$ . Найдём их оценки методом наибольшего правдоподобия:

$$\sigma^2 = b$$

$$L = L(a, b; x_1, \dots, x_n) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x_1-a)^2}{2b}} \dots \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x_n-a)^2}{2b}} = \frac{1}{(2\pi b)^{n/2}} e^{-\sum_{k=1}^n \frac{(x_k-a)^2}{2b}}$$

$$\ln L = -\frac{1}{2b} \sum_{k=1}^n (x_k - a)^2 - \frac{n}{2} \ln(2\pi b)$$

константы можно опускать

$$\left\{ \begin{array}{l} \frac{\partial \ln L}{\partial a} = -\frac{1}{b} \sum_{k=1}^n (x_k - \hat{a}), \\ \frac{\partial \ln L}{\partial b} = \frac{1}{2b^2} \sum_{k=1}^n (x_k - \hat{a})^2 - \frac{n}{2} \hat{b}^{-1} = 0, \end{array} \right. \Rightarrow \begin{array}{l} \sum_{k=1}^n \hat{a} = \sum_{k=1}^n x_k \Rightarrow \hat{a} = \frac{1}{n} \sum_{k=1}^n x_k \\ \hat{b} = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{a})^2 \end{array}$$

## 116 Распределение выборочной дисперсии в нормальной выборке.

Пусть  $\xi_1, \dots, \xi_n$  имеют нормальное распределение  $(a, \sigma)$  - нормальная выборка.  $\frac{\xi - a}{\sigma} = \eta$  -  $(0, 1)$  - нормальная случайная величина.

Рассмотрим два случая:

Случай 1  $a$  известно

$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{j=1}^N (\xi_j - a)^2 = \frac{\sigma^2}{N} \sum_{j=1}^N \eta_j^2$$

$$\hat{\sigma}_N^2 =^d \frac{\sigma^2}{N} \chi_N^2$$

$$\hat{a}_N - a = \frac{\sigma}{\sqrt{N}} \eta$$

Случай 2  $a$  неизвестно

$$\hat{a}_N = \frac{1}{N} \sum_{j=1}^N \xi_j = \frac{1}{N} \sum_{j=1}^N (a + \eta'_j) =^d a + \frac{\sqrt{N\sigma^2}}{N} \eta = a + \frac{\sigma}{\sqrt{N}} \eta$$

$$S_N^2 = \frac{\sigma^2}{N-1} \chi_{N-1}^2$$

Независимы:  $P_{\hat{a}_N, S_N^2}(x, y) = P_{\hat{a}_N}(x) P_{S_N^2}(y)$

Ошибка:

$$\hat{a}_N - a = \frac{\sigma}{\sqrt{N}} \eta$$

Отношение:

$$\frac{\hat{a}_N - a}{S} =^d \frac{\frac{\sigma}{\sqrt{N}} \eta}{\frac{\sigma}{\sqrt{N-1}} \chi_{N-1}} = \frac{\eta}{\chi_{N-1}} = \frac{\tau_{N-1}}{\sqrt{N-1}}$$

## 117 Эмпирические распределения.

Конечно, экспериментально (эмпирически) можно исследовать не только  $a$  и  $\sigma$ , но и  $F_\xi(x)$ ,  $P_\xi(x)$  - одним словом распределение случайной величины.

Прежде всего вспомним модель Бернулли:  $n$  испытаний и  $\nu$  успехов.

$$P_k = P(\nu = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$M\nu = np \quad M\frac{\nu}{n} = p$$

$$D\nu = npa \quad D\frac{\nu}{n} = \frac{pq}{n} \rightarrow \infty$$

$\frac{\nu}{n} = \hat{p}$  - несмещенная и состоятельная оценка  $p$ .

Пусть имеем  $\xi_i \in (-\infty, \infty)$

$$P(\xi \leq x) \approx \hat{P}(\xi \leq x) = \frac{\nu(\xi \leq x)}{n} = \hat{F}_\xi(x)$$

$$P(x_1 < \xi \leq x_2) \approx \frac{\nu(x_1 < \xi \leq x_2)}{n} = \hat{F}_\xi(x_2) - \hat{F}_\xi(x_1)$$

$$x_2 = x_1 + \Delta x_1$$

$$F_\xi(x_2) - F_\xi(x_1) = \int_{x_1}^{x_2} P_\xi(x) dx$$

$$\frac{\Delta F_\xi}{\Delta x_1} \approx \frac{1}{\Delta x} \int_{x_1}^{x_2} P_\xi(x) dx$$

Разобьем ось на  $x_1, x_2, \dots, x_k$  - получим гистограмму - прямоугольники, которой будут иметь площадь  $\nu(x_1 < \xi \leq x_2)$

## 118 Критерий .

Рассмотрим гистограмму  $\nu_1 + \nu_2 + \dots + \nu_k = n$  и гипотезу  $P_1, P_2, \dots, P_k$ . Каждое из  $\nu_i$  распределено по биномиальному закону  $a = np$ ,  $\sigma^2 = npq$ . Биномиальное распределение следует из распределения Пуассона с  $a = np$  и  $\sigma^2 = np$ . При  $a \geq 5$ , распределение Пуассона становится схожим с нормальным распределением.

Значит, при  $np_i \geq 10$  случайная величина распределена примерно по нормальному закону со средним  $np_i$  и дисперсией  $np_i$ ;

$$\zeta_i = \frac{\nu_i - np_i}{\sqrt{np_i}}$$

- по нормальному закону  $(0, 1)$ , а

$$\sum_{i=1}^n \zeta_i^2 = \frac{(\nu_i - a_i)^2}{a_i}$$

- по закону  $\chi^2$ .

Из распределения  $\chi^2$  найдем такие  $z_\alpha$ , чтобы

$$\int_0^{z_\alpha} P_{\chi^2}(z) dz = 1 - \alpha$$

Значит, событие  $(\chi^2 > z_\alpha)$  может произойти лишь с малой вероятностью  $\alpha$  (например:  $\alpha = 0.05$ ). Поэтому, если  $\chi^2 > z_\alpha$ , то гипотеза неверна, если  $\chi^2 \leq z_\alpha$  - то гипотеза верна (точнее не противоречит).\*/

## 119 Обобщенные функции (распределения)

Один из способов введения дельта-функции состоит в следующем:

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \quad \text{функция X}$$

$\varphi(x)$  - дифференцируемая функция, обращающаяся в нуль вне некоторого интервала (пробная функция), непрерывная  $\varphi(x) = 0$ ,  $|x| \geq a$ .

Рассмотрим

$$\int_{-\infty}^{\infty} \varphi'(x) H(x) dx = \varphi(x) H(x) \Big|_{-a}^a - \int_{-\infty}^{\infty} \varphi(x) H'(x) dx$$

Таким образом

$$-\int_{-\infty}^{\infty} \varphi(x) H'(x) dx = \int_0^{\infty} \varphi'(x) dx = \varphi(x) \Big|_0^{\infty} = -\varphi(0)$$

Если  $\varphi(x)$  - дважды дифференцируемая функция, то

$$\int_{-\infty}^{\infty} \varphi''(x) H(x) dx = - \int_{-\infty}^{\infty} \varphi'(x) H'(x) dx = \int_{-\infty}^{\infty} \varphi(x) \delta'(x) dx$$

$$H''(x) = \delta'(x)$$

$$\int_0^{\infty} \varphi''(x) dx = -\varphi'(0)$$

Обобщенные функции распределения.

## 120 Критерий $\chi^2$ для сравнения распределений.

Рассмотрим  $N$  измерений,  $N_i$  измерений попало в  $i$ -ый интервал.

Построим гистограмму:  $\frac{N_i}{N\Delta x}$ ,  $\frac{N_i}{N} = \hat{P}_i$  сравним с

$$P_i = \int_{\Delta x_i} P_{\xi}(x) dx$$

1. Задают уровни значимости  $\alpha$ .
2. Из распределения  $\chi^2_{n-1}$  находят квантиль  $x_{\alpha}$ :  $P\{\chi^2 \geq x_{\alpha}\} = \alpha$
3. Находят

$$\chi^2 = \sum_{i=1}^n \frac{(N_i - Np_i)^2}{Np_i}$$

4. Если  $\chi^2 \geq x_{\alpha}$ , то эксперимент не подтверждает теорию. Если  $\chi^2 < x_{\alpha}$ , то подтверждает.

## 121 Анализ оценок $\hat{a}$ и $\hat{b}$ .

$M\hat{a}_n = \dots a$  - несмещенная.

$$P(|\hat{a}_n - a| \leq \varepsilon) = 1 - P(|\hat{a}_n - a| > \varepsilon)$$

Если  $D\hat{a}_n \rightarrow 0$ , то состоятельная:

$$\leq \frac{D\hat{a}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty$$

по неравенству Чебышева.

Значит  $P(|\hat{a}_n - a| \leq \varepsilon) \rightarrow 1$  при  $n \rightarrow \infty$  состоятельная.

Найдем чему равно  $M\hat{b}_n$ ?

Пусть  $a = 0$ .

$$\hat{b}_n = \frac{1}{n} \sum_{k=1}^n (\xi_k^2 - 2\xi\hat{a}_n + \hat{a}_n^2) = \dots$$

1.

$$M\left(\frac{1}{n}\sum_{k=1}^n\xi_k^2\right)=\frac{1}{n}nM\xi^2=\sigma^2$$

2.

$$M\left(2\xi_k\hat{a}_k^2\right)=2\frac{1}{n}\sum_{k=1}^nM\xi_k\xi_i=\frac{2}{n}\sigma^2$$

3.

$$M(\hat{a}_n^2)=\frac{1}{n^2}\sum_{lm}\xi_l\xi_m=\frac{1}{n}\sigma^2$$

$$M\hat{b}_n=\left(1-\frac{2}{n}+\frac{1}{n}\right)\sigma^2=\frac{n-1}{n}\sigma^2$$

$\hat{b}_n$  не является несмещенной! Она смещена. Несмещенной является

$$\frac{n}{n-1}\hat{b}_n=\frac{1}{n-1}\sum_{k=1}^n(\xi_k-\hat{a})^2\equiv S^2$$

$$MS^2=\sigma^2$$

Можно показать, что  $S^2$  состоятельная оценка  $\sigma^2$ .

## 122 Формула Стирлинга.

$$\int_1^n \ln x dx = x \ln x \Big|_1^n - \int_1^n dx = n \ln n - n + 1$$

$$\int_1^n \ln x dx \approx \text{по формуле трапеций} \approx \frac{\ln 1 + \ln 2}{2} + \ln 2 + \dots + \ln(n-1) = \ln 1 * 2 * \dots * n - \frac{1}{2} \ln n =$$

$$= \ln n! - \frac{1}{2} \ln n$$

приравниваем:

$$\ln n! \approx n \ln n - n + 1 + \frac{1}{2} \ln n = \left(n + \frac{1}{2}\right) \ln n - n$$

$$n! \approx n^{n+1/2} e^{-n}$$

или более точно,

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \quad \text{формула Стирлинга.}$$

А ели еще точнее, то

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \left[1 + \frac{1}{12n}\right]$$

## 123 Центральная предельная теорема.

Если  $\xi_1, \dots, \xi_n$  - независимые случайные величины, имеющие один и тот же закон распределения с мат. ожиданием  $a$  и дисперсией  $\sigma^2 < \infty$ , то при  $n \rightarrow \infty$  закон распределения суммы  $\sum_{k=1}^n \xi_k$  стремится к нормальному.

**Доказательство.** Обозначим  $\sum_{k=1}^n \xi_k = \eta_n$ . Характеристическая функция

$$\varphi_{\eta_n}(t) = [\varphi_{\xi}(t)]^n$$

В окрестности  $t = 0$

$$\varphi_{\xi}(t) = \varphi_{\xi}(0) + \varphi'_{\xi}(0)t + \frac{1}{2}\varphi''_{\xi}(0)t^2 + \dots = 1 + iat - \frac{\sigma^2}{2}t^2$$

Обозначим

$$\zeta_n = \frac{\eta_n - na}{\sigma\sqrt{n}} = \underbrace{\frac{\xi_1 - a}{\sigma\sqrt{n}}}_{\xi'_1} + \underbrace{\frac{\xi_2 - a}{\sigma\sqrt{n}}}_{\xi'_2} + \dots + \underbrace{\frac{\xi_n - a}{\sigma\sqrt{n}}}_{\xi'_n}$$

$$\varphi_{\zeta_n}(t) = [\varphi_{\xi'}(t)]^n$$

$$\begin{aligned}\varphi_{\xi'}(t) &= \varphi_{-\frac{a}{\sigma\sqrt{n}} + \frac{1}{\sigma\sqrt{n}}\xi} = e^{-\frac{iat}{\sigma\sqrt{n}}} \varphi_{\xi}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left(1 - \frac{iat}{\sigma\sqrt{n}} + \dots\right) \left(1 + \frac{iat}{\sigma\sqrt{n}} - \frac{\sigma^2 t^2}{2\sigma^2 n} + \dots\right) = \\ &= 1 - \frac{iat}{\sigma\sqrt{n}} + \frac{iat}{\sigma\sqrt{n}} - \frac{t^2}{2n} + \dots\end{aligned}$$

## 124 Другие теоремы

**Теорема 1.** Некоррелированные нормальные случайные величины являются независимыми.

**Теорема 2.** Выборочное среднее  $\bar{X}$  и дисперсия  $s^2$  нормальной генеральной совокупности являются независимыми. **Доказательство:** Для  $n = 2$ .

$$\bar{X} = \frac{X_1 + X_2}{2}, \quad s^2 = \frac{(X_1 - X_2)^2}{2}$$

Достаточно доказать, что  $X_1 + X_2$  и  $X_1 - X_2$  являются независимыми. Поскольку они распределены нормально, то достаточно доказать, что они некоррелированы. Мы можем легко это сделать:

$$\text{Cov}(X_1 + X_2, X_1 - X_2) = \text{Cov}(X_1, X_1) - \text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_1) - \text{Cov}(X_2, X_2) = 0.$$

Допустим теперь, что теорема справедлива для выборки объёма  $n$  и покажем, что это влечет справедливость для выборки объёма  $n + 1$ .



## 124.1 Проверка на двух средних

Независимые случайные выборки объёмов  $n_1$  и  $n_2$  соответственно получены из двух генеральных совокупностей со средними  $\mu_1$  и  $\mu_2$  и дисперсиями  $\sigma_1$  и  $\sigma_2$ . Поскольку дисперсии известны, статистика

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

имеет нормальное распределение. Тестируя гипотезы

$$H_0 : \mu_1 - \mu_2 = d, \quad H_1 : \mu_1 - \mu_2 \neq d$$

мы *отвергаем*  $H_0$ , если

$$Z \equiv \frac{\bar{X}_1 - \bar{X}_2 - d}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -z_{\alpha/2}, \quad \text{or if} \quad Z > z_{\alpha/2}.$$

Если дисперсии неизвестны и объёмы выборок невелики, то мы должны использовать  $t$ -статистику:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - d}{S_p \sqrt{1/n_1 + 1/n_2}},$$

где

$$S_p = \sqrt{S_p^2}, \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Таким образом, двухсторонняя гипотеза  $H_0$  *не отвергается* при

$$-t_{\alpha/2}(n_1 + n_2 - 2) < t < t_{\alpha/2}(n_1 + n_2 - 2).$$