Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

«Допустить к защите»_____

Заведующий кафедрой
д.ф.-м.н., профессор С.М. Ермаков

Учебное пособие для 1 курса факультета физической культуры и спорта

Алексеева Нина Петровна

Методы математической статистики

Санкт-Петербург – 2019

Оглавление

Список вопросов

1. Случайные события и их свойства. Достоверное, случайное, невозможное. Несовместность событий. Сумма и произведение событий. Противоположное событие.

2. Классическая вероятностная схема. Теорема о сложении двух несовместных и совместных событий. Вероятность суммы трех событий.

3. Условная вероятность. Независимость событий. Вероятность произведения независимых и зависимых событий.

4. Принцип маловероятного события. Размещения, перестановки, сочетания. Вычисление вероятности выигрыша в игре в кости.

5. Формула полной вероятности и теорема Байеса. Априорные и апостериорные вероятности.

6. Дискретные случайные величины. Независимость. Примеры распределений: Бернулли, биномиальное.

7. Характеристики распределений: энтропия, среднее, дисперсия.

8. Распределение Пуассона, его математическое ожидание и дисперсия.

9. Непрерывные случайные величины: равномерное и нормальное распределения.

10. Структура распределений хи-квадрат, Стьюдента и Фишера.

11. Проверка статистических гипотез. Принцип маловероятных событий. Ошибки 1 и 2-ого рода. Мощность критериев.

12. Критерий Пирсона хи-квадрат согласия эмпирического распределения с теоретическим.

13. Таблицы сопряженности. Критерий независимости качественных признаков.

14. Критерии Фишера и Стьюдента однородности независимых выборок. Критерии Вилкоксона и Манна-Уитни.

15. Критерий однородности для зависимых выборок. Критерии Стьюдента и Мак-Немара.

16. Проверка однородности для нескольких выборок при помощи однофакторного дисперсионного анализа.

17. Метод наименьших квадратов для оценки параметров линейной регрессии.

18. Коэффициент корреляции. Критерий равенства нулю коэффициента корреляции.

19. Прогнозирование и частные корреляции.

20. Классификация при помощи дискриминантного анализа

21. Основы факторного анализа. Главные компоненты и факторные нагрузки.

1. Случайные события и их свойства

Объектами исследования теории вероятностей являются события. Если при определенном комплексе условий событие обязательно происходит, то оно называется **достоверным**. Обычно его обозначают через Ω. Например, при подбрасывании игрального кубика, шесть граней которого пронумерованы от 1 до 6, достоверным является событие, которое заключается в том, что выпадет какое-то число очков, больше нуля.

Если при определенном комплексе условий событие не может произойти, то оно называется **невозможным**. Обычно его обозначают через знак пустого множества $\varnothing$. Например, невозможно при однократном подбрасывании кубика получить 7 очков.

Если событие может произойти или не произойти, то оно называется **случайным**. Случайные события принято обозначать большими латинскими буквами $A, B, C, \ldots$. Например, событие, которое заключается в том, что при однократном подбрасывании кубика число выпавших очков будет четно, является случайным.

События A и B называются **несовместными**, если появление одного события исключает появление другого. Например, выпадение четного и нечетного числа очков являются несовместными, а события "сегодня пойдет дождь" и "сегодня пойдет снег" несовместными не являются, так как иногда бывает дождь со снегом.

Событие $C = A + B$ называется **суммой** событий A и B, если происходит либо A, либо B, либо оба одновременно, если они не являются несовместными. Например, события A и B заключаются в том, что при подбрасывании кубика выпало три и шесть очков соответственно. Событие $C = A + B$ означает, что выпадет число очков, кратное трем.

Событие $\overline{A}$ **противоположно** событию A, если A и $\overline{A}$ несовместны,

а их сумма есть достоверное событие, то есть какое-то из A или $\overline{A}$ обязательно происходит. Например, если событие A соответствует выпадению четного количества очков, то событие $\overline{A}$ нечетного.

Событие $D = AB$ называется **произведением** двух событий A и B, если оно происходит, когда одновременно происходят оба этих события. Например, событие A соответствует выпадению четного количества очков, событие B означает, что число очков кратно трем, тогда AB означает, что число очков равно шести (и четное, и делится на три).

2. Классическая вероятностная схема

Пусть имеется конечное множество $\Omega = \{\omega_1, \ldots, \omega_n\}$ элементарных (неделимых) исходов. **Вероятность любого составного события $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{im}\}$, состоящего из m элементарных исходов $\omega_{i_k} \in \Omega$, определяется как отношение числа элементарных исходов, благоприятствующих событию A, к общему числу исходов**

$$P(A) = \frac{m}{n}\,. \tag{1}$$

Например, всего $n = 20$ экзаменационных вопросов, из которых студент знает только 15. Вероятность того, что достанется незнакомый вопрос, равна $\frac{5}{20} = 0.25$.

Событие $\Omega = \{\omega_1, \ldots, \omega_n\}$, образованное всей совокупностью элементарных событий, является достоверным. Согласно (1), вероятность достоверного события равна 1,

$$P(\Omega) = \frac{n}{n} = 1\,.$$

Теорема (сложение вероятностей). *Для любых несовместных составных событий $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{im}\}$ и $B = \{\omega_{j_1}, \omega_{j_2}, \ldots, \omega_{jk}\}$ справедливо*

$$P(A + B) = P(A) + P(B). \tag{2}$$

Для доказательства объединим события A и B в одно составное событие

$$C = A + B = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{im}, \omega_{j_1}, \omega_{j_2}, \ldots, \omega_{jk}\}\,.$$

По формуле (1) вычислим вероятность $P(C)$,

$$P(A + B) = P(C) = \frac{m + k}{n} = \frac{m}{n} + \frac{k}{n} = P(A) + P(B).\ \blacksquare$$

Из этой теоремы получаем два важных следствия, касающиеся вероятности невозможного и противоположного события. Вероятность противоположного события вычисляется как $P(\overline{A}) = 1 - P(A)$, так как $1 = P(\Omega) =$

$P(A+\overline{A}) = P(A)+P(\overline{A})$. Вероятность невозможного события равна нулю, так как невозможное событие противоположно полному событию Ω.

Теорема (формула сложения вероятностей для событий, которые не являются несовместными).

$$P(A+B) = P(A)+P(B)-P(AB). \tag{3}$$

Доказательство. Представим сумму $A+B$ в виде суммы двух несовместных событий A и $B\setminus A$ и используем формулу (2) вычисления вероятности суммы двух несовместных событий

$$P(A+B) = P(A)+P(B\setminus A).$$

Разность $B\setminus A$ означает событие, при котором событие B происходит, когда не происходит A. С другой стороны, $P(B) = P(AB)+P(B\setminus A)$, отсюда

$$P(A+B) = P(A)+P(B\setminus A) = P(A)+(P(B)-P(AB)).\blacksquare$$

В случае трех событий формула вероятности суммы событий имеет вид:

$$P(A+B+C) = P(A)+P(B)+P(C)-P(AB)-P(BC)-P(AC)+P(ABC) \tag{4}$$

3. Условная вероятность и независимость событий

Условной вероятностью $P(A|B)$ называется вероятность события A при условии, что событие B уже произошло и уменьшило количество элементарных исходов с n до $m \leq n$. Условную вероятность можно вычислить по классической формуле вероятности (1). Для этого нужно знать m – количество элементарных исходов, удовлетворяющих условному событию B, и r – количество элементарных исходов, удовлетворяющих событию AB:

$$P(A|B) = \frac{r}{m} = \frac{r/n}{m/n} = \frac{P(AB)}{P(B)} \tag{5}$$

Отсюда получаем формулу для вычисления вероятности произведения двух событий

$$P(AB) = P(A|B)P(B) = P(B|A)P(A). \tag{6}$$

Например, событие A означает четное число очков, B_1 число очков, большее 2, B_2 число очков, большее 3. Условные вероятности равны

$$P(A|B_1) = \frac{2}{4} = 0.5, \quad P(A|B_2) = \frac{2}{3} = 0.667.$$

Сравним эти вероятности с безусловной вероятностью $P(A) = \frac{3}{6} = 0.5$. В одном случае имеем совпадение условной и безусловной вероятностей, в другом нет.

Если условная вероятность совпадает с безусловной, т.е. $P(A|B) = P(A)$, то события A и B называются **независимыми.** Из (5) получаем, что для независимых событий вероятность произведения независимых событий равна произведению вероятностей этих событий:

$$P(A) = P(A|B) = \frac{P(AB)}{P(B)}\,, \quad \Rightarrow \quad P(AB) = P(A)P(B)\,. \tag{7}$$

Задача 1.(продолжение) *3). Будут ли независимыми пары событий E-F и E-D?*

Для проверки независимости событий E (знание английского языка), F (знание французского языка) сравним вероятность $P(EF) = 0.2$ с произведением вероятностей $P(E)P(F) = 0.5 \cdot 0.4 = 0.2$, вероятности совпали, следовательно, знание английского языка не зависит от знания французского.

Для проверки независимости событий E (знание английского языка), D (знание немецкого языка) сравним вероятность $P(ED) = 0.08$ с произведением вероятностей $P(E)P(D) = 0.5 \cdot 0.35 = 0.175 \neq 0.08$, вероятности не совпали, следовательно, знание немецкого языка зависит от знания английского.

$F \setminus E$	E	$\overline{E}$	сумма
F	20	20	40
$\overline{F}$	30	30	60
сумма	50	50	100

$D \setminus E$	E	$\overline{E}$	сумма
D	8	27	35
$\overline{D}$	42	23	65
сумма	50	50	100

Вычисляя условные вероятности $P(D|E) = 8/50 = 0.16 < 0.35 = P(D)$ и $P(D|\overline{E}) = 27/50 = 0.54 > 0.35 = P(D)$, получаем более высокую вероятность знания немецкого языка среди студентов, которые не знают английский язык.

4. Маловероятное событие

В математической статистике для проверки гипотез используется принцип маловероятных событий, а именно, событие, вероятность которого мала, считается невозможным.

4.1. Размещения, перестановки, сочетания

Множество, состоящее из n элементов, будем называть n-множеством. **Сочетанием** будем называть m-подмножество из n-множества. Количество сочетаний будем обозначать через C_n^m. **Размещение** – это упорядоченное m-подмножество из n-множества. Количество размещений A_n^m вычисляется по формуле

$$A_n^m = \underbrace{n \cdot (n-1) \cdot \ldots \cdot (n-m+1)}_{m \text{ сомножителей}} = \frac{n!}{(n-m)!} \tag{8}$$

Например, 5 спортсменов борются за первые три призовых места. Сколько вариантов исходов соревнования может быть? Первое место может занять любой из пяти спортсменов, второе место – любой из оставшихся четырех, третье – любой из оставшихся трех. Таким образом $A_5^3 = 5 \cdot 4 \cdot 3 = \frac{5!}{2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2} = 60$. Различные варианты упорядочивания n-множества называются **перестановками**. Число перестановок вычисляется по (8)

$$A_n^n = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$$

Перестановки 3-множества имеют вид $(123), (132), (213), (231), (312), (321)$. С другой стороны, число размещений можно выразить через число сочетаний $A_n^m = C_n^m \cdot m!$, отсюда и из (8) получаем формулу для вычисления числа сочетаний:

$$C_n^m = \frac{A_n^m}{m!} = \frac{n!}{(n-m)!m!} \tag{9}$$

Например, количество исходов результата соревнований, в котором нас интересует только тройка призеров из четырех претендентов, равно $C_4^3 = \frac{4 \cdot 3 \cdot 2}{3!} = 4$. Эти варианты можно перечислить: $(123), (124), (134), (234)$.

4.2. Вероятность выигрыша в игре в кости

Игральными костями служили кости животных – астрагалы, которые при бросании могли падать на четыре стороны: A, B, C, D. В одной из игр в древней Греции бросали одновременно четыре астрагала. Выигрышным броском считался тот, при котором выпадали разные стороны; такой бросок назывался „Венерой“. В археологических раскопках, начиная с V тысячелетия до н.э. (возможно, ранее) среди найденных костей астрагалы встречаются в несколько десятков раз чаще, чем другие кости. Частота выпадения широкой стороны A с углублением примерно равна 0.39, следующей по величине B равна 0.37; частоты выпадения двух оставшихся сторон C и D равны 0.12.

При подбрасывании четырех костей одновременно общее число исходов равно $4^4 = 256$, но вероятности их появления разные. Например, по формуле (7) вероятность получить комбинацию сторон $ABCD$ равна произведению $P(ABCD) = 0.39 \cdot 0.37 \cdot 0.12 \cdot 0.12 = 0.002$.

Всего броску „Венера“ соответствуют $24 = 4 \cdot 3 \cdot 2 \cdot 1$ варианта перестановок названных сторон на четырех астрагалах:

$$
\begin{array}{cccccc}
ABCD & ABDC & ACBD & ACDB & ADBC & ADCB \\
BACD & BADC & BCAD & BCDA & BDAC & BDCA \\
CABD & CADB & CBAD & CBDA & CDAB & CDBA \\
DABC & DACB & DBAC & DBCA & DCAB & DCBA
\end{array}
$$

Суммируя соответствующие вероятности, получаем, что вероятность броска „Венера“ равна $24 \cdot 0.002 \approx 0.05$.

Практическое занятие

Первое практическое занятие посвящено решению задач на тему классического определения вероятности.

1. *Пусть A, B, C три произвольных события одного испытания. Например, A высота взята первый раз, B второй, C третий. Выразите через них следующие события: а) произошли только два из указанных исхода, б) произошло не более двух из них, в) ни одно из событий не произошло.*

   Ответ.

   а) $\bar{A}BC + A\bar{B}C + AB\bar{C}$ — две успешные попытки.

   б) $\bar{A}\bar{B}\bar{C} + \bar{A}\bar{B}C + \bar{A}B\bar{C} + A\bar{B}\bar{C}$ — три или две неудачи.

   в) $\bar{A}\bar{B}\bar{C}$ — все неудачно.

2. *Из карточек с буквами Ф, И, З, К, У, Л, Ь, Т, У, Р, А выбираются наугад четыре и приставляются друг к другу. Чему равна вероятность того, что получится слово КУРА?*

   Решение. Эта задача на произведение зависимых событий. Слово КУРА получится, если на первое место мы выберем букву К – она одна из 11 букв, вероятность 1/11, на второе У – это событие произойдет с вероятностью 2/10 – букв осталось только 10, буква Р на третьем месте окажется с вероятностью 1/9 и буква А на четвертом с вероятностью 1/8. Если перемножить эти вероятности, то получится число $\frac{1}{11} \cdot \frac{2}{10} \cdot \frac{1}{9} \cdot \frac{1}{8} = 0.00025$. Формально это можно выразить так:

   $P(A_1A_2A_3A_4) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2) \cdot P(A_4|A_1A_2A_3)$

3. *В лотерее 100 билетов, среди которых выигрыши 1 – 500 рублей, 3*

по 250, 6 по 100 и 15 по 30. Найти вероятность выиграть не менее 250 рублей владельцу одного билета.

Решение. Выигрыш не менее 250 рублей осуществляется с вероятностью $4/100$.

4. *Десять книг на одной полке расставлены наудачу. Найти вероятность того, что три определенных тома окажутся рядом в определенном порядке.*

Общее число исходов равно числу перестановок 10!. Три тома, идущие подряд, могут начинаться с 1, 2,..., 8-го места, для остальных томов возможно расположение 7! способами. Таким образом, ответ $\frac{8\cdot 7!}{10!} = \frac{1}{9\cdot 10} = \frac{1}{90}$.

5. *Из колоды 52 карт наугад извлекаются три карты. Найти вероятность того, что это будут тройка, семерка и туз.*

Урновая схема. Предположим, что в урне N шаров, среди которых M красных. Из урны вынимают наугад n шаров без возвращения. Вероятность того, среди n шаров окажутся из m красных, вычисляется как отношение числа $C_M^m C_{N-M}^{n-m}$ событий, отвечающих этому условию, к общему числу возможных исходов C_N^n. Это можно обобщить на случай большего разнообразия шаров. Если в урне M_1 красных, M_2 белых и M_3 черных шаров, $N = M_1+M_2+M_3$, то вероятность того, что из $n = m_1+m_2+m_3$ наугад вынутых шаров окажутся соответственно m_1 красных, m_2 белых и m_3 черных шаров, равна

$$\frac{C_{M_1}^{m_1} C_{M_2}^{m_2} C_{M_3}^{m_3}}{C_{M_1+M_2+M_3}^{m_1+m_2+m_3}}.$$

Аналогично строится формула для большего разнообразия. Применим эту формулу для решения задачи про карты. В колоде $52 =$

$4+4+4+40$ четыре тройки, четыре семерки, четыре туза и 40 карт остального вида. Вынимаются три $3=1+1+1+0$.

$$\frac{C_4^1C_4^1C_4^1C_{40}^0}{C_{52}^3}=\frac{4^3\cdot 3!}{52\cdot 51\cdot 50}=0.002895\,.$$

6. *Номер автомашины состоит из 3 букв 30 букв алфавита и 4 цифр. 1) сколько различных номеров можно составить при условии, что цифры и буквы в отдельном номере не будут повторяться? 2) Какова вероятность того, что все буквы и цифры в номере различны?*

   Общее число номеров $M=30^3\cdot 10^4$, число вариантов без повторений равно $M_1=(30\cdot 29\cdot 28)\cdot(10\cdot 9\cdot 8\cdot 7)$,

   $$P=\frac{M_1}{M}=\frac{(30\cdot 29\cdot 28)\cdot(10\cdot 9\cdot 8\cdot 7)}{30^3\cdot 10^4}=0.455\,.$$

7. *Какова вероятность, что из выбранных наугад 7 карт из колоды, содержащей 32 карты, окажутся 1) три карты красной масти и четыре черной? 2) четыре туза, два короля и дама?*

   Общее число исходов $N=C_{32}^7$. Ответ первый $\frac{C_{16}^3\cdot C_{16}^4}{C_{32}^7}$. Ответ второй $\frac{C_4^4\cdot C_4^2\cdot C_4^1}{C_{32}^7}$.

8. *Три стрелка A, B, C независимо друг от друга выстрелили по мишени. Вероятности попадания для кадого стрелка равны $P(A)=1/2, P(B)=1/3, P(C)=1/4$. Каковы вероятности: 1) ни одна из пулей не попала в цель, 2) одна пуля попала в цель, 3) две пули попали в цель, 4) по крайней мере одна пуля попала в цель.*

   $$1)p_1=P(\bar{A})\cdot P(\bar{B})\cdot P(\bar{C})=\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}=\frac{1}{4},$$

$$2) p_2 = P(A\bar{B}\bar{C} + \bar{A}B\bar{C} + \bar{A}\bar{B}C) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4} + \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{3}{4} + \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{4} = \frac{11}{24},$$

$$3) p_3 = P(\bar{A}BC + A\bar{B}C + AB\bar{C}) = \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{3}{4} = \frac{1}{4},$$

$$3) p_4 = 1 - p_1 = \frac{3}{4}.$$

9. *Какова вероятность, что дни рождения 10 студентов приходятся на 1) разные дни невисокосного года, 2) разные дни января, на февраль?*

   Общее число исходов равно 365^{10}. Число исходов, соответствующих разным дням невисокосного года, равно $A_{365}^{10} = 365 \cdot 364 \cdot \ldots \cdot 356$. Разные дни января $\frac{A_{31}^{10}}{365^{10}}$, на февраль $\frac{28^{10}}{365^{10}}$.

10. *В игре в преферанс используется колода в 32 карты. Трем игрокам сдается по 10 карт и 2 карты идут в прикуп. Какова вероятность того, что обе карты в прикупе будут тузами, а одному игроку достанутся четыре семерки?*

    Общее число исходов $N = C_{32}^2 C_{30}^1 0 C_{20}^{10} C_{10}^1 0$, число благоприятствующих исходов равно $M = C_4^2 C_3^1 C_4^4 C_{26}^6 C_{20}^{10} C_{10}^1 0$. Ответ $P = \frac{M}{N} = \frac{1}{3496}$.

Задачи на самопроверку

1. Сколькими способами можно расставить 5 книг на полке?

   Варианты ответа: 1) 25, 2) 125, 3) 120, 4) 24.

2. Вероятность четкого броска у баскетболиста равна $p = 0.75$. Он сделал $n = 5$ бросков. Чему равны вероятности: p_1 - первый раз промахнуться, а потом все точно, p_2 промахнуться всего один раз, p_3 все броски точные, p_4 хоть раз, но промахнуться.

   Сопоставьте правильные ответы.

A) $1 - p^5$		p_1
B) p^5		p_2
C) qp^4		p_3
D) $5qp^4$		p_4

3. В группе из 10 человек половина перворазрядники, три мастера спорта и двое любителей. К доске вызваны три человека. Чему равна вероятность, что это все три мастера спорта? Отметьте правильные решения.

   1) $\frac{3}{7}$, 2) $\frac{3}{10}$, 3) $\frac{C_3^3 C_7^0}{C_{10}^3}$, 4) $\frac{1}{120}$.

4. В лотерейном билете, состоящем из 36 чисел, нужно отметить шесть чисел. Какова вероятность угадать 1) все шесть номеров, 2) пять из них, 3) четыре из них, 4) не более трех?

5. Априорные и апостериорные вероятности

5.1. Формула полной вероятности

События $A_1, \ldots, A_n$ образуют **полную группу событий**, если они несовместны ($A_i A_j = \varnothing,\ i \neq j$), и какое-то из них обязательно произойдет ($A_1 + \ldots + A_n = \Omega$). Например, событие A_1 заключается в том, что случайно выбранный человек имеет первую группу крови, A_2 – вторую, A_3 – третью, A_4 – четвертую. Известно, что 33.7% людей имеет первую группу крови, поэтому вероятность события A_1 равна $P(A_1) = 0.337$. Аналогично имеем остальные **априорные вероятности**: $P(A_2) = 0.375$, $P(A_3) = 0.209$, $P(A_4) = 0.079$.

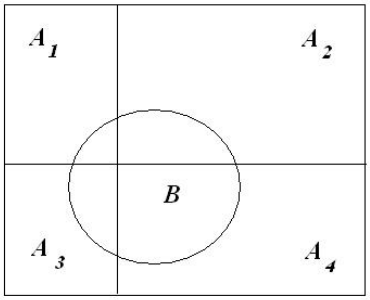

Рис. 1. $A_1, \ldots, A_4$ – полная группа событий. $B = BA_1 + BA_2 + BA_3 + BA_4$.

Обозначим через B событие – некоторому человеку неизвестно с какой группой крови перелили неизвестную кровь, и он остался жив. Для вычисления вероятности $P(B)$ заметим, что событие B складывается (рис.1) из своих пересечений BA_1, BA_2, BA_3, BA_4 с событиям $A_1, \ldots, A_4$, которые образуют полную группу:

$$B = BA_1 + \ldots + BA_n.$$

Вероятность этой суммы определяется как сумма вероятностей:

$$P(B) = P(BA_1) + \ldots + P(BA_n) =$$
$$= P(B|A_1)P(A_1) + \ldots + P(B|A_n)P(A_n).$$

Для вычисления вероятностей произведений $P(BA_k)$ мы воспользовались формулой (6) вида $P(BA_k) = P(B|A_k)P(A_k)$. Поскольку больному с первой группой можно переливать только первую группу крови, то вероятность его выживания равна $P(B|A_1) = 0.337$. Больному со второй группой можно переливать первую и вторую, то $P(B|A_2) = 0.337 + 0.375 = 0.712$. Больному с третьей группой крови можно переливать первую и третью, то $P(B|A_3) = 0.337 + 0.209 = 0.546$. Больному с четвертой группой крови можно переливать любую кровь, поэтому $P(B|A_4) = 1$, следовательно,

$$P(B) = P(B|A_1)P(A_1) + \ldots + P(B|A_4)P(A_4) = 0.5737.$$

Для полной группы событий **формула полной вероятности** имеет вид

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)\,. \tag{10}$$

5.2. Формула Байеса

Апостериорные (буквально, после опыта) вероятности $P(A_i|B)$ по формуле условной вероятности (5) имеют вид $P(A_i|B) = \frac{P(A_iB)}{P(B)}$. Отсюда получаем доли представителей каждой группы крови среди выживших после случайного переливания крови: $P(A_1|B) = 0.1980$, $P(A_2|B) = 0.4654$, $P(A_3|B) = 0.1999$, $P(A_4|B) = 0.1387$.

В общем виде для вычисления апостериорных вероятностей используется **формула Байеса**:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum\limits_{i=1}^{n} P(B|A_i)P(A_i)} = \frac{P(A_kB)}{P(B)}\,.$$

1. *В ящике лежат 20 теннисных мячей, в том числе 12 новых и 8 игранных. Из него извлекают случайно два мяча для игры и возвращают обратно. После этого снова случайно вынимают два мяча для следующей игры. Найти вероятность того, что последние два мяча будут неигранными.*

   Возможны три варианта: A_1 в первый раз были вынуты два игранных мача, поэтому к следующему разу число новых мячей так и осталось равным 12; A_2 в первый раз были вынуты два мача: один новый и один игранный, поэтому к следующему разу число новых мячей стало равным 11, а старых 9; A_3 в первый раз были вынуты два новых мача, поэтому к следующему разу число новых мячей стало равным 10, а старых 10. Вычислим вероятности этих событий.

$$P(A_1) = \frac{C_8^2 C_{12}^0}{C_{20}^2} = 0.1474, \quad P(A_2) = \frac{C_8^1 C_{12}^1}{C_{20}^2} = 0.5053,$$
$$P(A_3) = \frac{C_8^0 C_{12}^2}{C_{20}^2} = 0.3473,$$

   Можно непосредственно убедиться в том, что сумма этих вероятностей равна 1. Обозначим через B событие, которое заключается в том, что вынутые во второй раз два мяча окажутся неигранными. Для вычисления $P(B) = 0.2797$ используем формулу полной вероятности.

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3), \text{ где}$$

$$P(B|A_1) = \frac{C_8^0 C_{12}^2}{C_{20}^2} = 0.3474, \quad P(B|A_2) = \frac{C_9^0 C_{11}^2}{C_{20}^2} = 0.2895$$
$$P(B|A_3) = \frac{C_{10}^0 C_{10}^2}{C_{20}^2} = 0.2368.$$

2. *Из полного набора костей домино наугад выбираются две кости. Чему равна вероятность того, что вторую кость можно приставить к первой?*

   Обозначим искомое событие через B. В домино есть 7 костей вида (a,a), $a = 0,1,2,\ldots,6$, и $21 = \frac{7\cdot 6}{2}$ кость вида (a,b). Если первая кость вида (a,a) – событие A_1, которое осуществляется с вероятностью $P(A_1) = \frac{7}{28}$, то вторая кость должна иметь a – таких может быть шесть ($b = 0,1,\ldots,6$, но $b \neq a$) из оставшихся 27, следовательно, $P(B|A_1) = \frac{6}{27}$. В случае A_2, когда первая кость вида (a,b) – вероятность $P(A_2) = \frac{21}{28}$, ей подойдут 12 костей из 27, $P(B|A_2) = \frac{12}{27}$. Например, если $a = 1, b = 0$, то им подойдут шесть костей вида $(1,1),(1,2),(1,3,(1,4),(1,5),(1,6)$ и шесть костей вида $(0,0),(0,2),(0,3),(0,4),(0,5),(0,6)$. Итак

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) = \frac{7}{28}\cdot\frac{6}{27} + \frac{21}{28}\cdot\frac{12}{27} = \frac{7}{18}\,.$$

3. *На двух фермах A и B произошла вспышка заболевания ящуром. Доли зараженного скота равны соответственно $\frac{1}{6}$ и $\frac{1}{4}$. Из каждой фермы отбирают по одной корове. Какова вероятность того, что больна только одна изх них? Если заражена одна, то чему равна вероятность того, что эта корова с фермы A?*

   Обозначим через C событие, что из двух отобранных коров больна одна. Это возможно, если корова из первой фермы больная, а из второй здоровая, или наоборот, из первой здоровая, а из второй больная.

$$P(C) = \frac{1}{6}\cdot\frac{3}{4} + \frac{5}{6}\cdot\frac{1}{4} = \frac{3}{24} + \frac{5}{24} = \frac{1}{3} = P(CA) + P(CB)\,.$$

$$P(A|C) = \frac{P(AC)}{P(C)} = \frac{3/24}{8/24} = \frac{3}{8}\,.$$

4. *Стрелок A поражает мишень с вероятностью 0.6, а стрелок B с вероятностью 0.5. Стрелки одновременно выстрелили в мишень, и одна пуля попала в цель. Найти вероятность того, что стрелок B попал в цель.*

   Обозначим через C событие, что один из двух выстрелов успешен. Это возможно, если первый стрелок попал, а второй нет, или наоборот.

$$P(C) = 0.6 \cdot 0.5 + 0.4 \cdot 0.5 = 0.3 + 0.2 = 0.5 = P(CA) + P(CB)\,.$$

$$P(B|C) = \frac{P(BC)}{P(C)} = \frac{0.2}{0.5} = 0.4\,.$$

6. Дискретные случайные величины (Бернулли, биномиальные)

Случайная величина - это величина, которая в результате опыта со случайным исходом принимает то или иное числовое значение, причем заранее неизвестно, какое именно. В случае конечного или счетного количества исходов случайная величина ξ называется **дискретной**. Если известны значения $x_1, x_2, \ldots, x_N, \ldots$, которые принимает случайная величина ξ, а также вероятности $p_i = P\{\xi = x_i\},\ i = 1, \ldots, N,\ p_1 + \ldots + p_N + \ldots = 1$, то говорят, что задан ее **дискретный закон распределения**

$$\xi : \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ p_1 & p_2 & \ldots & p_N \end{pmatrix}. \tag{11}$$

Например, при подбрасывании монеты возможны два исхода: успех или неудача, которые кодируются соответственно 1 и 0. Успех обычно ассоциируется с выпадением монеты строной "орел", которая выпадает с вероятностью $p = 0.5$. Неудача ассоциируется с выпадением стороны "решка", которая выпадает с вероятностью $q = 1 - p = 0.5$. Вероятность успеха p необязательно равна 0.5. Если при подбрасывании игрального кубика успехом считать выпадении шести очков, то $p = \frac{1}{6}$, $q = \frac{5}{6}$. Случайная величина, принимающая два значения с вероятностями p и $q = 1 - p$, имеет **распределение Бернулли.**

Случайные величины ξ со значениями $x_1, x_2, \ldots$ и η со значениями $y_1, y_2, \ldots$ **независимы**, если для любых i и j

$$P\{\xi = x_i, \eta = y_j\} = P\{\xi = x_i\} P\{\eta = y_j\}. \tag{12}$$

Обозначим через ξ случайное число выпадения шести очков игрального кубика в $n = 4$ независимых испытаниях с вероятностью выпадения

шести очков (успех) $p = \frac{1}{6}$. Построим закон распределения этой случайной величины. Вычисление вероятностей $P\{\xi = k\},\ k = 0, 1, \ldots, 4$, отображено в таблице 1, где через $q = 1 - \frac{1}{6} = \frac{5}{6}$ обозначена вероятность невыпадения шести очков (неудача), через C_n^k число сочетаний по k из n элементов, которое вычисляется по (9). Проверим, что вероятности p_k соответствуют закону распределения:

$$\sum_{k=1}^{4} p_k = \frac{5^4 + 4 \cdot 5^3 + 6 \cdot 5^2 + 4 \cdot 5 + 1}{6^4} = \frac{(5+1)^4}{6^4} = 1.$$

Обозначим через ε_i, $i = 1, 2, \ldots, n$, независимые случайные величины, имеющие распределение Бернулли с параметром p. Случайная величина $\xi = \varepsilon_1 + \ldots + \varepsilon_n$, равная случайному числу успехов из n независимых испытаний, имеет **биномиальный закон распределения**:

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n. \tag{13}$$

число успехов k	варианты комбинаций	вероятность p_k
0	○ ○ ○○	$C_4^0 p^0 q^4 = \left(\frac{5}{6}\right)^4$
1	● ○ ○○ ○ ● ○○ ○ ○ ●○ ○ ○ ○●	$C_4^1 p^1 q^3 = 4 \cdot \frac{1}{6} \left(\frac{5}{6}\right)^3$
2	● ● ○○ ● ○ ●○ ● ○ ○● ○ ● ●○ ○ ● ○● ○ ○ ●●	$C_4^2 p^2 q^2 = 6 \cdot \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$
3	● ● ●○ ● ● ○● ● ○ ●● ○ ● ●●	$C_4^3 p^3 q^1 = 4 \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^1$
4	● ● ●●	$C_4^4 p^4 q^0 = \left(\frac{1}{6}\right)^4$

Таблица 1. Биномиальный закон распределения при $n = 4$ и $p = 1/6$.

1. *Игральная кость подбрасывается 16 раз. Найти наивероятнейшее число появления очков, кратного трем.*

В n независимых испытаниях с вероятностью успеха p определим наивероятнейшее число успехов. Обозначим через $P_n(k) = C_n^k p^k q^{n-k}$ вероятность k успехов в n испытаниях. Пусть k_0 наивероятнейшее число успехов. С одной стороны, должно выполняться неравенство $P_n(k_0-1) \le P_n(k_0)$, с другой стороны $P_n(k_0) \ge P_n(k_0+1)$. Из первого имеем $C_n^{k_0-1} p^{k_0-1} q^{n-k_0+1} \le C_n^{k_0} p^{k_0} q^{n-k_0} \iff$

$$\frac{n!}{(k_0-1)!(n-k_0+1)!} p^{k_0-1} q^{n-k_0+1} \le \frac{n!}{k_0!(n-k_0)!} p^{k_0} q^{n-k_0} \iff$$

$$\frac{q}{(n-k_0+1)} \le \frac{p}{k_0} \iff k_0 q \le p(n-k_0+1),$$

отсюда получаем первое условие $k_0 \le p(n+1)$. Из второго неравенства имеем $C_n^{k_0} p^{k_0} q^{n-k_0} \ge C_n^{k_0+1} p^{k_0+1} q^{n-k_0-1} \iff$

$$\frac{n!}{k_0!(n-k_0)!} p^{k_0} q^{n-k_0} \ge \frac{n!}{(k_0+1)!(n-k_0-1)!} p^{k_0+1} q^{n-k_0-1} \iff$$

$$\frac{q}{(n-k_0)} \ge \frac{p}{(k_0+1)} \iff (k_0+1)q \ge p(n-k_0),$$

отсюда получаем второе неравенство $k_0 \ge np - q$. Таким образом,

$$np - q \le k_0 \le np + p\,.$$

В задаче $n = 16, p = \frac{1}{3}$, получаем $k_0 = 5$, подставляя значения в неравенство,

$$\frac{16}{3} - \frac{2}{3} \le k_0 \le \frac{16}{3} + \frac{1}{3} \iff \frac{14}{3} \le k_0 \le \frac{17}{3}\,.$$

2. *Сколько раз необходимо бросить игральную кость, чтобы с вероятностью 0.9 шесть очков выпало хотя бы один раз?*

Вероятность того, что в n испытаниях шесть очков не выпадет ни разу равна $\left(\frac{5}{6}\right)^n$, поэтому вероятность того, что они выпадут хотя бы раз равна $1 - \left(\frac{5}{6}\right)^n$. Найдем n, при котором эта вероятность будет больше 0.9.

$$1 - \left(\frac{5}{6}\right)^n > 0.9 \iff \left(\frac{5}{6}\right)^n < \frac{1}{10} \iff \left(\frac{6}{5}\right)^n > 10,$$

отсюда $n > \frac{\log 10}{\log 6 - \log 5} = 12.63$.

7. Характеристики распределений

7.1. Математическое ожидание дискретной величины и ее линейного преобразования

Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n & \dots \\ p_1 & \dots & p_n & \dots \end{pmatrix} \tag{14}$$

случайную величину или эксперимент, а через $x_1, \dots, x_n, \dots$ возможные исходы, которые осуществляются с вероятностями $p_1, \dots, p_n, \dots$, $\sum\limits_{k=1}^{n} p_k = 1$. Математическое ожидание (МО) или среднее значение дискретной случайной величины (21) вычисляется по формуле

$$\mathbb{E}\xi = \sum_{k=1}^{\infty} x_k p_k$$

Например, МО случайной величины, имеющей распределение Бернулли, равно $0 \cdot q + 1 \cdot p = p$, а для биномиального распределения (13)

$$\mathbb{E}\xi = \sum_{k=0}^{n} \frac{kn!}{k!(n-k)!} p^k q^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-k} = np.$$

Математическое ожидание линейной комбинации $a\xi + b$ имеет вид

$$\mathbb{E}(a\xi + b) = \sum_{i=1}^{\infty} (ax_i + b)p_i = a \sum_{i=1}^{\infty} x_i p_i + \sum_{i=1}^{\infty} bp_i = a\mathbb{E}\xi + b. \tag{15}$$

МО суммы двух случайных величин ξ и η равно сумме их математических ожиданий.

$$\mathbb{E}(\xi + \eta) = \sum_i \sum_j (x_i + y_j) P\{\xi = x_i, \eta = y_j\} =$$

$$= \sum_i \sum_j x_i P\{\xi = x_i, \eta = y_j\} + \sum_i \sum_j y_j P\{\xi = x_i, \eta = y_j\} =$$

$$= \sum_i x_i \sum_j P\{\xi = x_i, \eta = y_j\} + \sum_j y_j \sum_i P\{\xi = x_i, \eta = y_j\} =$$

$$= \sum_i x_i P\{\xi = x_i\} + \sum_j y_j P\{\eta = y_j\} = \mathbb{E}\xi + \mathbb{E}\eta\,.$$

МО произведения двух независимых (12) случайных величин ξ и η равно произведению их математических ожиданий.

$$\mathbb{E}\xi\eta = \sum_i \sum_j x_i y_j P\{\xi = x_i, \eta = y_j\} = \sum_i \sum_j x_i y_j P\{\xi = x_i\} P\{\eta = y_j\} = \\ = \sum_i x_i P\{\xi = x_i\} \sum_j y_j P\{\eta = y_j\} = \mathbb{E}\xi\mathbb{E}\eta\,.$$

7.2. Дисперсия на примере дискретных величин. Дисперсия суммы величин

Характеристика вариабельности данных – **дисперсия** $\mathbb{D}\xi$, которую часто обозначают как σ^2, определяется как средний квадрат отклонения от среднего значения

$$\mathbb{D}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2 = \mathbb{E}(\xi^2 - 2\xi\mathbb{E}\xi + (\mathbb{E}\xi)^2) = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2\,. \tag{16}$$

Наряду с дисперсией σ^2 используют также характеристику **стандартного отклонения** σ, которое равно квадратному корню из дисперсии. Например, для случайной величины, имеющей распределение Бернулли, дисперсия может быть вычислена следующим образом:

$$\mathbb{E}\xi^2 = 0^2 \cdot q + 1^2 \cdot p = p,$$
$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = p - p^2 = pq\,.$$

Дисперсия суммы двух случайных величин ξ, η равна

$$D(\xi + \eta) = \mathbb{E}((\xi + \eta) - \mathbb{E}(\xi + \eta))^2 = \mathbb{E}((\xi - \mathbb{E}\xi) + (\eta - \mathbb{E}\eta))^2 = \\ = \mathbb{E}(\xi - \mathbb{E}\xi)^2 + \mathbb{E}(\eta - \mathbb{E}\eta)^2 + 2\underbrace{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}_{cov(\xi,\eta)}$$

Если величины ξ, η независимы, то ковариация

$$cov(\xi, \eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)\mathbb{E}(\eta - \mathbb{E}\eta) = 0,$$

и дисперсия суммы независимых величин равна сумме дисперсий. Отсюда получаем, что дисперсия биномиально распределенной случайной величины равна npq.

Дисперсия линейного преобразованияи $a\xi + b$ не зависит от сдвига и отличется на a^2

$$\mathbb{D}(a\xi + b) = \mathbb{E}(a\xi + b - \mathbb{E}(a\xi + b))^2 = a^2\mathbb{E}(\xi - \mathbb{E}\xi)^2 = a^2\mathbb{D}\xi$$

Заметим, что дисперсия константы равна нулю.

8. Распределение Пуассона

Отдельно рассматривается случай, когда число испытаний велико при малой вероятности успеха p. Например, вероятность вызова скорой помощи в определенном временном интервале чрезвычайно мала в каждом отдельном случае. Но при наличии большого числа испытаний количество вызовов может оказаться счетным, то есть равным $0, 1, 2, \ldots$. Для того чтобы вычислить при $p \to 0$ и $n \to \infty$ предел вероятности

$$p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

обозначим через $\lambda = np$ и умножим и разделим p_k на n^k.

$$p_k = \frac{n(n-1)\ldots(n-k+1)}{k! n^k} (np)^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \tag{17}$$

так как, по замечательному пределу, $e^{-1} = \lim\limits_{p\to 0}(1-p)^{\frac{1}{p}}$, при заданном k

$$\lim_{p\to 0}(1-p)^{n-k} = \lim_{p\to 0}\left((1-p)^{\frac{1}{p}}\right)^{pn}(1-p)^{-k} = e^{-\lambda},$$
$$\lim_{n\to\infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} = 1\,.$$

Случайная величина ξ, принимающая значения $k = 0, 1, 2, \ldots$ с вероятностями (17), имеет **распределение Пуассона.** Так как $e^{\lambda} = \sum\limits_{k=0}^{\infty} \frac{\lambda^k}{k!}$, то $\sum\limits_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = 1$. Математическое ожидание и дисперсия случайной величины ξ равны λ, так как

$$\mathbb{E}\xi = \sum_{k=0}^{\infty} k p_k = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda\,.$$
$$\mathbb{E}\xi^2 = \sum_{k=0}^{\infty} k^2 p_k = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} \overset{k=t+1}{=} \sum_{t=0}^{\infty} \frac{(t+1)\lambda^{t+1}}{t!} e^{-\lambda} = \lambda^2 + \lambda,$$
$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda\,.$$

9. Непрерывные случайные величины

9.1. Функция распределения и плотность. МО и дисперсия абсолюно непрерывных величин

Функцией распределения называется функция $F(x) = P\{\xi \leq x\}$. Из свойств вероятности эта функция неуюывающая, непрерывна справа и $F(-\infty) = 0$, $F(\infty) = 1$. Распределение случайной величины ξ называется абсолютно непрерывным, если существует неотрицательная функция $f(x)$, которая называется плотностью распределения, такая, что вероятность попадания случайной величины в интервал B определяется как

$$P\{\xi \in B\} = \int\limits_B f(x)dx\,. \tag{18}$$

Плотность может быть получена как производная от функции распределения $f(x) = F'(x)$, и наоборот, $F(x) = \int_{-\infty}^{x} f(u)du$. Основное свойство плотности заключается в том, что $\int_{-\infty}^{\infty} f(x)dx = 1\,.$

Плотность линейного преобразования $\eta = a\xi + b$ может быть выражена из плотности $f_\xi(x)$,

$$F_\eta(x) = P\{\eta \leq x\} = P\{a\xi + b \leq x\} = P\left\{\xi \leq \frac{x-b}{a}\right\} = F_\xi\left(\frac{x-b}{a}\right),$$
$$f_\eta(x) = F'_\eta(x) = \frac{1}{a} f_\xi\left(\frac{x-b}{a}\right) \tag{19}$$

МО и дисперсия абсолютно непрерывной величины ξ с плотностью $f(x)$ вычисляются по формулам

$$\mathbb{E}\xi = \int\limits_{-\infty}^{\infty} x f(x)dx\,, \quad \mathbb{D}\xi = \int\limits_{-\infty}^{\infty} x^2 f(x)dx - (\mathbb{E}\xi)^2\,,$$

Квадратный корень из дисперсии носит название стандартного отклонения.

9.2. Равномерное распределение.

Случайная величина ξ равномерно распределена на интервале $[a, b]$, если ее плотность распределения имеет вид $f(x) = \frac{1}{b-a}$ на интервале $[a, b]$ и равна нулю вне этого интервала. МО равно середине интервала, дисперсия пропорциональна длине интервала.

$$\mathbb{E}\xi = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2}|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

$$\mathbb{E}\xi^2 = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3}|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3},$$

$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

9.3. Нормальное распределение. МО и дисперсия.

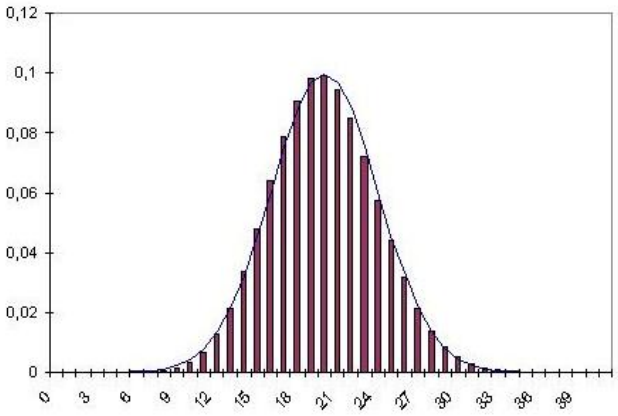

Рис. 2. Биномиальное распределение с параметрами $n = 100$ и $p = 0.2$ и нормальное распределение с параметрами $\mu = 20$, $\sigma^2 = 16$.

Нормальное или гауссовское распределение было получено как предельное биномиальное распределение при увеличиении числа испытаний.

Плотность его распределения зависит от двух параметров μ, σ^2, смысл

которых заключается в среднем и дисперсии.

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{20}$$

При $\mu = 0$, $\sigma^2 = 1$ распределение $f(x|0,1)$ называется стандартным нормальным.

Для доказательства основного свойства плотности нужно вычислить двойной интеграл

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = 4\int\limits_{0}^{\infty}\int\limits_{0}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = \left(\int\limits_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx\right)\left(\int\limits_{-\infty}^{\infty} e^{-\frac{y^2}{2}}dy\right),$$

который может быть вычислен переходом к полярным координатам $x = r\cos\phi, y = r\sin\phi$ с якобианом преобразования

$$dxdy = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{vmatrix} drd\phi = \begin{vmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{vmatrix} drd\phi = rdrd\phi,$$

$$4\int\limits_{0}^{\infty}\int\limits_{0}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = 4\int\limits_{0}^{\frac{\pi}{2}} d\phi \int\limits_{0}^{\infty} e^{-\frac{r^2}{2}} rdr = 2\pi\left(-e^{-\frac{r^2}{2}}\right)|_0^{\infty} = 2\pi\,.$$

Отсюда $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx = \sqrt{2\pi}$, и справедливо свойство плотности. Пусть $\xi_0 \sim \mathcal{N}(0,1)$. МО $\mathbb{E}\xi_0 = 0$ из-за симметричности функции относительно нуля:

$$\mathbb{E}\xi_0 = \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}}dx = 0\,,$$

$$\mathbb{E}\xi_0^2 = \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}}dx = \frac{2}{\sqrt{2\pi}}\int\limits_{0}^{\infty} x\cdot x e^{-\frac{x^2}{2}}dx =$$

Воспользуемся интегрированием по частям $\int uv'dx = uv - \int u'vdx$, где $u = x$, $v' = xe^{-\frac{x^2}{2}}$, $v = -e^{-\frac{x^2}{2}}$.

$$= -\frac{2}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}}|_0^{\infty} + \frac{2}{\sqrt{2\pi}}\int\limits_{0}^{\infty} e^{-\frac{x^2}{2}}dx = 1\,.$$

Следовательно, $\mathbb{D}\xi_0 = 1$. Отсюда для $\xi = \sigma\xi_0 + \mu$ имеем МО $\mathbb{E}\xi = \mathbb{E}(\sigma\xi_0 + \mu) = \mu$ и дисперсию $\mathbb{D}\xi = \mathbb{D}(\sigma\xi_0 + \mu) = \sigma^2\mathbb{D}\xi_0 = \sigma^2$.

Функцию и плотность нормального распределения $\mathcal{N}(\mu, \sigma)$ можно использовать для вычисления вероятностей биномиального распределения $\beta(x|n, p)$ при больших значениях n. При этом между параметрами распределений существуют определенные соотношения, связанные с тождеством математических ожиданий и дисперсий: $\mu = np$, $\sigma^2 = npq$. Например, значения $\beta(10|n = 30, p = 0.2) = 0.03547089, f(10|\mu = np = 6, \sigma = \sqrt{npq} = 2.19) = 0.03439262$ оказываются достаточно близкими.

Пусть случайная величина ξ имеет биномиальное распределение $\beta(x|n, p)$. Для вычисления вероятности $P\{a < \xi < b\}$ используют так называемую интегральную теорему Муавра-Лапласа.

$$P\{a < \xi < b\} = F(b|np, \sqrt{npq}) - F(a|np, \sqrt{npq}),$$

где через $F(x|np, \sqrt{npq})$ обозначена функция нормального распределения. Например, при $n = 30, p = 0.2$ имеем $np = 6, \sqrt{npq} = 2.19$,

$$P\{4 < \xi < 8\} = F(8|np, \sqrt{npq}) - F(4|np, \sqrt{npq}) = 0.6386.$$

Для вычисления $F(x|\mu, \sigma)$ и $f(x|\mu, \sigma)$ можно использовать встроенную функцию в электронных таблицах НОРМРАСП, в которой нужно указать соответствующие параметры среднего и дисперсии, и в первом случае параметр ИНТЕГРАЛЬНАЯ=1, во втором ИНТЕГРАЛЬНАЯ=0.

1. *В коллективе 100 человек и начальник. Начальник выбирает правильное решение с вероятностью 0.95, а каждый член коллектива с вероятностью 0.8. Что лучше диктат или демократия?*

   Вычислим среднее $\mu = np = 80$, стандартное отклонение $\sigma = \sqrt{npq} = 4$ и вероятность правильного решения при демократии, то есть когда

число правильных решений более половины, то есть

$$P\{50 < \xi < +\infty\} = F(+\infty|80, 4) - F(50|80, 4) = 1 - 3 \cdot 10^{-14}\,.$$

Можно показать, что при вероятности привильного решения одного члена коллектива 0.585 уравниваются шансы диктатора и демократии.

2. *В страховой компании застраховано 10000 лиц. Вероятность страхового случая равна 0.006. Каждый застрахованный вносит 12 рублей в год, а в страховом случае компания платит 1000. Найти 1) вероятнсть того, что компания потерпит убыток, 2) компания получит прибыль менее 60000.*

   Поскольку число испытаний велико, а вероятность страхового случая мала, то применяем теорему Пуассона. Вычисляем интенсивность $\lambda = np = 10000 \cdot 0.006 = 60$. Обозначим ξ случайное число страховых случаев, а через $K = 12 \cdot 10000/1000 = 120$ число страховых случаев, при котором компания потерпит убыток.

$$P\{\xi > 120\} = 1 - F_p(120|\lambda = 60) = 3 \cdot 10^{-12}.$$

   При $K < 60$ прибыль окажется более 60000. Вероятность этого события равна

$$P\{\xi < 60\} = F_p(60|\lambda = 60) = 0.53.$$

   Для вычислений также можно использовать встроенные функции электронных таблиц.

3. *Сколько необходимо провести испытаний, чтобы вероятность того, что отклонение относительной частоты появления успеха от его вероятности $p = \frac{2}{3}$ меньше 0.01, будет равна 0.995?*

Пусть $x_1, \ldots, x_n$ случайные величины, принимающие значение 1 с вероятностью $p = 2/3$, $\mathbb{E}x_i = p$ и $\mathbb{D}x_i = pq$. Относительная частота имеет вид $\eta = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ и согласно центральной предельной теореме имеет нормальное распределение с параметрами

$$\mu = \mathbb{E}\eta = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x_i = p,$$

$$\sigma^2 = \mathbb{D}\eta = \mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}x_i = \frac{pq}{n} = \frac{2}{9n}\,.$$

Для стандартно нормально распределенной величины $\varepsilon = \frac{\eta - \mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}}$ справедливо $P\{-2.807 < \varepsilon < 2.807\} = 0.995$, поскольку $P\{\varepsilon < -2.807\} = 0.0025$, поэтому

$$P\left\{\left|\frac{\eta - \mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}}\right| < 2.807\right\} = 0.995$$

$$P\left\{|\eta - p| < 2.807\sqrt{\mathbb{D}\eta}\right\} = 0.995$$

$$2.807\sqrt{\mathbb{D}\eta} = 0.01 \quad\Longleftrightarrow\quad \frac{2\cdot 2.807^2}{9n} = 0.01^2$$

$$n^2 = \frac{2\cdot 2.807^2}{9\cdot 0.01^2} = 17422.22, \quad n \approx 132.$$

10. Энтропия и коэффициент неопределенности

Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n & \dots \\ p_1 & \dots & p_n & \dots \end{pmatrix} \tag{21}$$

случайную величину или любой другой эксперимент, а через $x_1, \dots, x_n, \dots$ возможные исходы, которые осуществляются с вероятностями $p_1, \dots, p_n, \dots$, $\sum\limits_{k=1}^{n} p_k = 1$. Количество информации $H(\xi)$, которое нужно получить, чтобы узнать об исходе эксперимента, вычисляется по формуле Шеннона

$$H(\xi) = \sum_{k=1}^{\infty} p_k \log_2 \frac{1}{p_k}. \tag{22}$$

Если число исходов конечно и равно n, то $H(\xi) = \sum\limits_{k=1}^{n} p_k \log_2 \frac{1}{p_k}$. Непосредственным вычислением можно убедиться, что в случае распределения Бернулли с вероятностью успеха $p = 0.5$ энтропия равна 1. Это так называемая единица информации или бит. Если $p \neq 0.5$, то $H(\xi) < 1$, что соответствует увеличению определенности, соответственно большей предсказуемости эксперимента. В случае вырожденного распределения, когда случайная величина принимает одно значение с вероятностью 1, энтропия равна 0.

Энтропия условного распределения называется условной энтропией. Если условные распределения случайного эксперимента ξ задаются при помощи некоторого случайного эксперимента η, то вычисляется средняя условная энтропия $H(\xi|\eta)$. Разность между энтропией $H(\xi)$ и средней условной энтропией $H(\xi|\eta)$ называется общим количеством информации $I(\xi, \eta)$, для которого есть формула $I(\xi, \eta) = H(\xi) + H(\eta) - H(\xi, \eta)$, где через $H(\xi, \eta)$ обозначена энтропия совместного распределения. Коэффициенты неопределенности

$$J(\xi|\eta) = \frac{I(\xi, \eta)}{H(\eta)}, \quad J(\eta|\xi) = \frac{I(\xi, \eta)}{H(\xi)}, \quad J = \frac{2I(\xi, \eta)}{H(\xi) + H(\eta)}.$$

Пример задачи. Обозначим через $p_{ij} = P\{\xi = x_i, \eta = y_j\}$ вероятности совместного распределения. В каком случае коэффициенты неопределенности равны нулю?

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.15 & 0.15 \\ 0.35 & 0.35 \end{bmatrix} \qquad \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.15 & 0.35 \\ 0.35 & 0.15 \end{bmatrix}$$

10.1. Распределения, связанные с нормальным. Квантили распределений

Из последовательности $\xi_1, \xi_2, \ldots, \xi_n$ независимых стандартно распределенных случайных величин были построены новые случайные величины η, ζ и ϑ,

Распределения	хи-квадрат	Стьюдента	Фишера
Степени свободы	n	n	n и m
	$\eta = \sum\limits_{i=1}^{n} \xi_i^2$	$\zeta = \frac{\xi_0}{\sqrt{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}}$	$\vartheta = \frac{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}{\frac{1}{m}\sum\limits_{k=1}^{m} \eta_k^2}$

распределения которых были вычислены и названы соответственно хи-квадрат, Стьюдента и Фишера. Степени свободы являются параметрами этих распределений.

Для того чтобы использовать эти распределения в статистическом анализе, нужно ввести понятие P-квантиль распределения, которому соответствует некоторое число F_p такое, что $P\{\xi \leq F_p\} = P$. Например, 0.95-квантиль распределения $\chi^2(1)$ со степенью свободы 1 равна 3.841459, то есть $P\{\chi^2(1) < 3.841459\} = 0.95$, $P\{\chi^2(1) > 3.841459\} = 0.05$. В R можно использовать функцию

$$qchisq(0.95, 1),$$

значение которой равно 3.841459. Для проверки использование обратной функции $pchisq(3.841459, 1)$ приведет к значению 0.95. Другой пример, 0.975-квантиль нормального распределения можно найти при помощи функции $qnorm(0.975, mean = 0, sd = 1)$, значение которой равно 1.959964, то есть вероятность получить это значение или даже больше равна 0.025. Обратная функция $pnorm(1.96, 0, 1)$ приводит к значению 0.9750021.

Для распределения Стьюдента с числом степеней свободы, равным 10, 0.975-квантиль можно найти при помощи функции $qt(0.975, 10)$, значение которой равно 2.228139, то есть вероятность получить это значение или даже больше равна 0.025. Обратная функция $pt(2.228139, 10)$ приводит к значению 0.975. При увеличении числа параметров (степени свободы) распределение Стьюдента становится все более похожим на стандартное нормальное распределение, кривая которого на рисунке 3 изображена тонкой линией.

P-квантили распределения Фишера с n_1, n_2 степенями свободы ищутся при помощи функции $qf(P, n_1, n_2)$. Если по некоторой x квантили нужно найти соответствующую вероятность, то используется функция $pf(x, n_1, n_2)$, например, функция $pf(4.102821, 2, 10)$ приводит к значению 0.95.

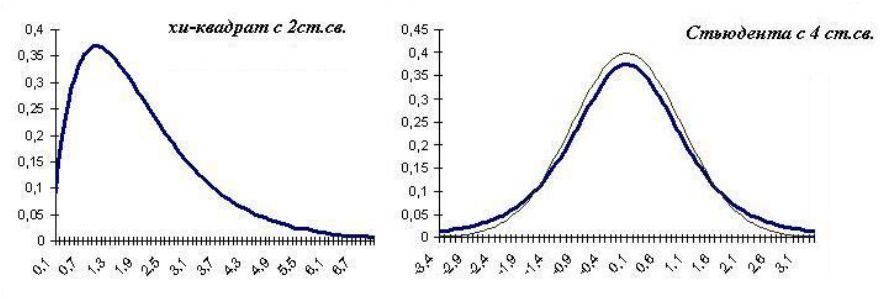

Рис. 3. Плотности распределения хи-квадрат и Стьюдента соответственно с 2-мя и 4-ми степенями свободы.

11. Проверка статистических гипотез

11.1. Принцип маловероятных событий

Для проверки статистической гипотезы H_0 относительно параметров θ или других свойств генеральной совокупности с функцией распределения $F(x|\theta)$ случайной величины ξ используются выборочные наблюдения $x_1, \ldots, x_n$. Альтернативную гипотезу будем обозначать через H_1.

Гипотеза называется простой, если она однозначно определяет функцию распределения, например, $H_0 : \theta = \theta_0$, иначе сложной, например, $H_0 : \theta > \theta_0$.

Правило, согласно которому отвергается гипотеза, называется *статистическим критерием,* а используемая для проверки гипотезы функция от выборочных наблюдений $x_1, \ldots, x_n$ называется *статистикой критерия.*

Проверка гипотез основана на *принципе маловероятных событий*, согласно которому события, вероятность которого мала, считаются невозможными. Сложилось так, что в качестве „малой“ вероятности принимают вероятность $\alpha = 0.05$ выигрыша в игре в кости. Гипотезу отвергают тогда, когда наблюдаемое значение статистики попадает в *критическую область* V, которую выбирает так, чтобы вероятность $P(V|H_0)$ была мала,

$$P(V|H_0) \leq \alpha\,.$$

ПРИМЕР. На $n = 25$ автомобилях с усовершенствованным двигателем средний расход бензина составил $\bar{x} = 9.3$ л на 100 км. Считая выборку нормальной $\mathcal{N}(\mu, \sigma)$ с дисперсией $\sigma^2 = 4$, выясним, нельзя ли наблюдаемое улучшение технической характеристики считать случайным. Проверим гипотезу $H_0 : \mu = 10$ о том, что расход топлива не изменился, В качестве альтернативной рассмотрим гипотезу $H_1 : \mu < 10$.

Используя свойства выборочного среднего $\bar{x}$, рассмотрим статистику критерия $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ с критической областью $V = (-\infty; -1.645]$, о которой известно, что $P(V) = 0.05$. Подставляя значения параметров и $\bar{x} = 9.3$, получаем значение статистики $Z = -1.75 \in V$, следовательно, гипотеза H_0 отвергается в пользу альтернативной H_1, уменьшение топлива значимо и не может быть объяснено случайностью.

11.2. Ошибки первого и второго рода, мощность, p-значение

Ошибка первого рода возникает, если отвергается верная гипотеза, $\alpha = P(V|H_0)$, *ошибка второго рода*, если не отвергается ложная гипотеза $\beta = P(\bar{V}|H_1)$. *Мощностью* называется вероятность $P(V|H_1) = 1 - \beta$.

Критерий для проверки H_0 называется *несмещенным,* если для любой альтернативы вероятность попадания в критическую область будет больше, чем вероятность попадания в нее при справедливости нулевой гипотезы, то есть $\forall H_1 \; P(V|H_0) \leq P(V|H_1)$.

Например, для проверки гипотезы $H_0 : \mu = 10$, при $\sigma = 2$, $n = 25$ с критической областью V, соотвествующей неравенству $\bar{x} < \mu - Z_{0.95}\frac{\sigma}{\sqrt{n}} = 9.342$, где $Z_{0.95}$ — квантиль $\mathcal{N}(0,1)$, ошибка первого рода равна

$$P\{\bar{x} < 9.342|\mu = 10\} = P\left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{9.342-\mu}{\sigma/\sqrt{n}}|\mu = 10\right\} = \Phi(-1.645) = 0.05\,,$$

через $\Phi(x)$ обозначена функция стандартного нормального распределения. Рассмотрим еще два критерия: $9.975 < \bar{x} < 10.025$ и $\bar{x} > 10.658$ с ошибками первого рода, также равными 0.05, и вычислим мощности при альтернативной гипотезе $H_1 : \mu = 9$.

$$\alpha_1 = \Phi\left(\frac{10.025-10}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{9.975-10}{\sigma/\sqrt{n}}\right) = 0.05,$$
$$1-\beta_1 = \Phi\left(\frac{10.025-9}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{9.975-9}{\sigma/\sqrt{n}}\right) = 0.0022,$$
$$\alpha_2 = 1 - \Phi\left(\frac{10.658-10}{\sigma/\sqrt{n}}\right) = 0.05, \quad 1-\beta_2 = 1 - \Phi\left(\frac{10.658-9}{\sigma/\sqrt{n}}\right) = 1.7\cdot 10^{-5}\,.$$

Итак, критерий $\bar{x} < 9.342$ оказывается наиболее мощным, но смещенным, так как $P(V|H_1) < 0.05$ при $H_1 : \mu > 10$. Несмещенным является критерий $|\bar{x} - \mu| > Z_{0.975}\frac{\sigma}{\sqrt{n}}$ с критической областью $V = (-\infty; 9.216] \cup [10.784; +\infty)$. Например, при $H_1 : \mu = 9$ и $H_1 : \mu = 11$ имеем $P(V) = 0.705$.

Кроме построения критических областей, для проверки гипотез используют p-значения. В случае односторонней альтернативы, например, $H_1 : \Theta > \Theta_0$ по заданному значению x_* статистики X считается доверительный уровень вероятности или p-значение, равное вероятности получить случайное значение больше наблюдаемого, $p = P\{X > x_*\}$. Нулевая гипотеза отвергается в пользу альтернативной в случае, когда p меньше заданного уровня значимости α, то есть при $p < \alpha$. В качестве уровня значимости обычно берется значение 0.05.

12. Критерий Пирсона хи-квадрат согласия эмпирического распределения с теоретическим.

12.1. Теорема о согласии распределений

Нулевая гипотеза H_0 заключается в том, что выборка наблюдений $x_1, \ldots, x_n$ относится к случайной величине ξ с генеральной функцией распределения $F(x)$. Множество S значений случайной величины разбито на r непересекающихся интервалов $S_1, \ldots, S_r$.

$$\cup_{i=1}^{r} S_i = S, \quad S_i \cap S_j = \emptyset,\ i \neq j,$$
$$p_i = P\{\xi \in S_i\} > 0, \quad \sum_{i=1}^{r} p_i = 1.$$

Соответствующие эмпирические частоты равны ν_i, $\sum\limits_{i=1}^{r} \nu_i = n$.

Теорема 1. *При справедливости H_0 статистика*

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{23}$$

имеет распределение $\chi^2(r-1)$ хи-квадрат с числом степеней свободы, равным $r-1$. Если параметры распределения оцениваются по выборке, то число степеней свободы выбирается равным $r-s-1$, где s число оцениваемых по выборке параметров.

12.2. Проверка согласия с нормальным распределением

На рис. 4 представлена гистограмма относительных частот выборки, смоделированной по нормальному закону с параметрами $\mu = 50$, $\sigma = 20$. Выясним, насколько согласовано эмпирическое распределение с нормальным $\mathcal{N}(\mu, \sigma)$. В качестве оценок рассмотрим $\hat{\mu} = \bar{x} = 48.72$, $\hat{\sigma} = S = 20.47$.

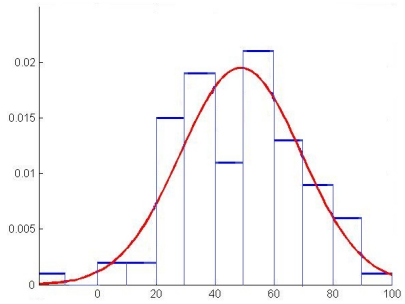

Рис. 4. Выборочная гистограмма и плотность $\mathcal{N}(\mu, \sigma)$.

Обозначим через $(z_{i-1}; z_i]$ интервал S_i, $i = 1, \ldots, r$, $z_0 = -\infty$, $z_r = +\infty$, через ν_i количество элементов выборки x_k, таких что $z_{i-1} < x_k \leq z_i$. Для вычисления вероятностей p_i используем функцию $\Phi(x)$ стадартного нормального распределения[1] $p_i = \Phi\left(\frac{z_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{z_{i-1} - \hat{\mu}}{\hat{\sigma}}\right)$. Объединим наблюдения в крайних ячейках так, чтобы $np_i > 5$, $n = 100$. Суммируя элементы в последнем столбце, получаем значение статистики $\chi^2_* = 8.356$. Число степеней свободы равно $df = 8 - 1 - 2 = 5$. Наблюдаемое значение статистики равно $\chi^2_* = 8.356$, p-значение имеет вид $p = P\{\chi^2(5) > \chi^2_*\}$. В R считаем $pchisq(8.356, 5) = 0.8623235 > 0.05$, следовательно, гипотеза о согласии эмпирического распределения с нормальным не отвергается с уровнем значимости $\alpha = 0.05$. При $p < \alpha$ мы бы сказали, что эмпирическое распределение плохо согласуется с нормальным.

i	$(z_{i-1}; z_i]$	ν_i	$\Phi(\frac{z_i - \hat{\mu}}{\hat{\sigma}})$	p_i	$\frac{(\nu_i - np_i)^2}{np_i}$
1	$(-\infty; 20]$	5	0.080	0.080	1.144
2	$(20; 30]$	15	0.180	0.100	2.510
3	$(30; 40]$	19	0.335	0.155	0.799
4	$(40; 50]$	11	0.525	0.190	3.360
5	$(50; 60]$	21	0.709	0.184	0.359
6	$(60; 70]$	13	0.851	0.142	0.094
7	$(70; 80]$	9	0.937	0.086	0.018
8	$(80; +\infty)$	7	1	0.063	0.072

[1] В электронных таблицах и в R значения функции $\Phi(x)$ можно получить при помощи функций НОРМСТРАСП(x) и $pnorm(x, mean, sd)$ соответственно

13. Таблицы сопряженности.

Предположим, что у нас имеется n наблюдений, характеризующихся значениями двух качественных признаков X и Y с градациями $x_1, \ldots, x_r$ и $y_1, \ldots, y_s$. Обозначим через

$$p_{ij} = P(X = x_i, Y = y_j), \quad p_{i*} = P(X = x_i), \quad p_{*j} = P(Y = y_j).$$

ПРИМЕР 1.1. X – это случайный ответ на вопрос „У меня бывают периоды плача и близости к слезам“, допускающий четыре градации x_1 – „совершенно не согласен“, x_2 – „скорее не согласен“, x_3 – „скорее согласен“, x_4 – „совершенно согласен“. Признак Y указывает на пол: y_1 – женский, y_2 – мужской. Из $n = 36$ студентов, участвующих в анкетировании, 14 студентов „совершенно не согласны“, из них 6 девушек и 8 юношей, „скорее не согласны“ пятеро (4 девушки и 1 юноша), 7 человек „скорее согласны“ (1 юноша и 6 девушек) и 10 „совершенно согласны“ (один юноша и 9 девушек).

$X \setminus Y$	y_1	y_2	$\ldots$	y_s	сумма
x_1	n_{11}	n_{12}	$\ldots$	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	$\ldots$	n_{2s}	n_{2*}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	n_{r*}
сумма	n_{*1}	n_{*2}	$\ldots$	n_{*s}	n

$X \setminus Y$	y_1	y_2	сумма
x_1	6	8	14
x_2	4	1	5
x_3	6	1	7
x_4	9	1	10
сумма	25	11	36

Если X и Y *независимы*, тогда для любых $i = 1, 2, \ldots, r$ и $j = 1, 2, \ldots, s$ справедливо

$$p_{ij} = p_{i*}p_{*j}. \tag{24}$$

При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{25}$$

и крайне редко удается получить равенства вида $\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \cdot \frac{n_{*j}}{n}$, например $\widehat{p}_{11} = \frac{6}{36} \neq \frac{14}{36} \cdot \frac{25}{36} = \widehat{p}_{1*}\widehat{p}_{*1}$. Важно определить, не является ли это нарушение случайным.

Для этого используется статистика χ^2 Пирсона (теор. 1), в выражении которой в качестве вероятностей гипотетического распределения рассматриваются произведения $\widehat{p}_{i*}\cdot\widehat{p}_{*j}$, соответствующие независимости признаков.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n\frac{n_{i*}}{n}\frac{n_{*j}}{n})^2}{n\frac{n_{i*}}{n}\frac{n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2 - 2n_{ij}\frac{n_{i*}n_{*j}}{n} + \left(\frac{n_{i*}n_{*j}}{n}\right)^2}{\frac{n_{i*}n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\frac{n\cdot n_{ij}^2}{n_{i*}n_{*j}} - 2n_{ij} + \frac{n_{i*}n_{*j}}{n}\right) = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right)$$

и при $n\to\infty$ имеет распределение хи-квадрат с числом степеней свободы, равным

$$rs - 1 - (r-1) - (s-1) = (r-1)(s-1)\,,$$

так как $\sum\limits_{i=1}^{r} p_{i*} = \sum\limits_{j=1}^{s} p_{*j} = 1$, и количество оцениваемых параметров по выборке равно $r+s-2$.

Получаем значение статистики $\chi^2 = 7.79$ с числом степеней свободы, равным $(4-1)(2-1) = 3$. В электронных таблицах для получения p-значения можно воспользоваться встроенной функцией ХИ2РАСП$(7.79;3)$ или в R функцией $pchisq(7.79,3) = 0.05055712$. Следует обратить внимание, что гипотеза отвергается при $p < 0.05$, не отвергается при $p > 0.2$, и в случае $0.05 \le p \le 0.2$ говорят, что имеет место ситуация неопределенности, при которой небольшие изменения в составе выборки могут изменить ситуацию в ту или другую сторону. Следовательно, связать этот вопрос с гендерным признаком не представляется возможным.

В частном случае $r = 2$, $s = 2$ и частотах $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$ в качестве статистики хи-квадрат независимости используется выражение

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)\,. \tag{26}$$

14. Критерии однородности двух независимых выборок.

14.1. Выборочные моменты

Пусть имеется выборка случайных наблюдений $x_1, \ldots, x_n$, под которыми мы понимаем независимые одинаково распределенные случайные величины. *Распределением выборки* называется распределение дискретной случайной величины, принимающей значения $x_1, \ldots, x_n$ с вероятностями $1/n$. Соответствующая функция распределения называется выборочной или *эмпирической функцией распределения*

$$F_n(x) = \frac{\mu_n(x)}{n},$$

где $\mu_n(x)$ равно количеству элементов выборки, не больших x. Эмпирическая функция $F_n(x)$ может служить приближенным значением (оценкой) функции распределения $F(x)$.

Характеристики выборочного распределения называются выборочными или эмпирическими. *Выборочное среднее* и *выборочные смещенная и несмещенная дисперсии* имеют соответственно вид

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

выражения для начальных и центральных выборочных моментов

$$a_\nu = \frac{1}{n}\sum_{i=1}^{n} x_i^\nu, \qquad m_\nu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^\nu.$$

Выборочной *модой* унимодального распределения является элемент выборки mod, встречающийся с наибольшей частотой. Выборочной *медианой* является число, которое делит вариационный ряд на две части, содержащие одинаковое число элементов. Если $n = 2k$, то $med = (x^{(k)} + x^{(k+1)})/2$. Если $n = 2k + 1$, то $med = x^{(k+1)}$.

14.2. Доверительный интервал для генерального среднего при известной дисперсии

Определение 1. Интервал, накрывающий истинное значение параметра распределения с заданной вероятностью P называется $P \cdot 100\%$ **- доверительным интервалом.**

В случае нормально распределенной выборки $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma)$ при известной дисперсии σ выборочное среднее как сумма одинаково нормально распределенных случайных величин имеет нормальное распределение с МО μ и дисперсией $\frac{\sigma^2}{n}$, так как

$$\mathbb{E}\bar{x} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x_i = \mu,$$
$$\mathbb{D}\bar{x} = \mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}x_i = \frac{\sigma^2}{n}.$$

Доверительный интервал для генерального среднего μ можно построить μ на основании

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \iff z = \frac{\bar{x}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1).$$

Обозначим через z_p квантили стандартного нормального распределения, соответствующие вероятности p, и выберем уровень значимости α.

$$P\{-z_{1-\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

и построим доверительный интервал для μ

$$P\left\{\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha.$$

В случае произвольного распределения это выражение можно использовать при достаточно большом объеме выборки.

14.3. Статистики Фишера и Стьюдента

Теорема 2. *Пусть* $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma)$, $\bar{x}$ *и* m_2 *– выборочные среднее и второй центральный момент. Тогда 1)* $\bar{x} \sim \mathcal{N}(\mu; \frac{\sigma}{\sqrt{n}})$*; 2) статистика* $\frac{nm_2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$*; 3)* $\bar{x}$ *и* m_2 *независимы.*

Пусть $x_1, \ldots, x_{n-1} \sim \mathcal{N}(\mu_1; \sigma)$ и $y_1, \ldots, y_{n_2} \sim \mathcal{N}(\mu_2; \sigma)$ две независимые нормально распределенные выборки с одинаковыми дисперсиями, S_1^2 и S_2^2 – несмещенные оценки их дисперсий. Тогда статистики

$$F = \frac{S_1^2}{S_2^2} \sim \mathbf{F}(n_1 - 1; n_2 - 1), \tag{27}$$

$$T = \frac{(\bar{x} - \bar{y} - (\mu_1 - \mu_2))\sqrt{n_1 + n_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sim \mathbf{T}(n_1 + n_2 - 2), \tag{28}$$

где через $\mathbf{F}(n_1; N-2)$ и $\mathbf{T}(n_1 + n_2 - 2)$ обозначены распределения Фишера и Стьюдента с соответствующим числом степеней свободы.

Действительно, по теор.2 имеем $(n_i - 1)S_i^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_i - 1})$, то

$$F = \frac{\frac{1}{n_1 - 1}(n_1 - 1)S_1^2/\sigma^2}{\frac{1}{n_2 - 1}(n_2 - 1)S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim \mathbf{F}(n_1 - 1; n_2 - 1).$$

Далее $\bar{x} - \bar{y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$, откуда

$$\xi = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0; 1).$$

С другой стороны, из аддитивности распределения $\chi^{\mathbf{2}}$

$$\eta = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^{\mathbf{2}}(\mathbf{n_1 + n_2 - 2}), \text{ так как}$$

$$(n_1 - 1)S_1^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_1 - 1}), \quad (n_2 - 1)S_2^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_2 - 1}),$$

$$\text{следовательно, } T = \frac{\xi}{\sqrt{\frac{1}{n_1 + n_2 - 2}\eta}} \sim \mathbf{T}(n_1 + n_2 - 2)\,.$$

14.4. Проблема неодинаковых дисперсий

В случае известных дисперсий σ_1^2 и σ_2^2 (возможно неодинаковых), учитывая нормальность средних

$$\bar{x} \sim \mathcal{N}(\mu_1, \sigma_1/\sqrt{n_1}), \quad \bar{y} \sim \mathcal{N}(\mu_2, \sigma_2/\sqrt{n_2}),$$

вычисляем дисперсию разности

$$\mathbf{D}(\bar{x} - \bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

откуда получаем

$$Z = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0; 1). \tag{29}$$

В случае неизвестных неодинаковых дисперсий $\sigma_1^2 \neq \sigma_2^2$ статистика T имеет приближенно распределение Стьюдента с нецелым числом степеней свободы.

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathbf{T}(k), \tag{30}$$

$$\text{где } k = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}.$$

14.5. Непараметрический аналог критерия Стьюдента

В R функция для использования параметрического критерия Стьюдента для проверки равенства средних имеет вид $t.test(x, y)$. По умолчанию равенство дисперсий не предполагается, иначе нужно изменить опцию $var.equal = FALSE$ на $var.equal = TRUE$.

В случае произвольного распределения выборок предпочтительнее использовать точный критерий Вилкоксона (в случае выборок объема менее 20), или его обобщение критерий Манна-Уитни при помощи функции $wilcox.test(x, y)$.

15. Критерий однородности для зависимых выборок

15.1. Критерий Стьюдента для зависимых выборок

Часто встречаются данные, в которых за одними и теми же индивидами наблюдают в разные моменты времени, например, у $n = 67$ больных с воспалением легких логарифм показателя скорости оседания эритроцитов (СОЭ) в начале заболевания в среднем равен 3.82 со стандартным отклонением 0.13, а при выписке из стационара 3.28 со стандартным отклонением 0.19. Требуется выяснить, можно ли считать уменьшение СОЭ значимым, или это уменьшение можно объснить случайностью.

Итак, имеется парная выборка (x_i, y_i), $i = 1, \ldots, n$, зависимых наблюдений объема n над случайными величинами ξ_1 и ξ_2 с нормальным распределением $\mathcal{N}(\mu_1, \sigma_1)$ и $\mathcal{N}(\mu_2, \sigma_2)$, относительно которых нет предположения о независимости — скорее всего, чем меньше был уровень СОЭ в начале заболевания, тем меньше он при выздоровлении. Рассмотрим разность $\eta = \xi_1 - \xi_2$ и соответствующую выборку $z_i = x_i - y_i$ с выборочным средним $\bar{z}$ и несмещенной оценкой дисперсии S_z^2. Если справедлива нулевая гипотеза $H_0 : \mu_1 = \mu_2$, то для случайной выборки $z_i = x_i - y_i$ должна быт справедлива гипотеза о равенстве нулю ее математического ожидания.

$$t = \frac{\bar{z}}{S_z}\sqrt{n} \ \sim \ \mathbf{T}(n-1) \text{ при справедливости } H_0 : \mu_1 = \mu_2 \,.$$

. Например, $\bar{z} = 0.54$, $S = 0.19$, $t = \frac{\bar{z}-0}{S}\sqrt{n} = \frac{0.54}{0.19}\sqrt{66} = 23.45 = t_*$, двусторонний доверительный уровень вероятности равен

$$p = P\{|t| > |t_*|\} = \text{СТЮДРАСП}(23.45; 66; 2) = 10^{-33} < 0.05,$$

поэтому гипотеза $H_0 : \mu_1 = \mu_2$ отвергается,[2] и уменьшение уровня СОЭ в процессе выздоровления нельзя объяснить случайностью.

[2] Значимость критерия в R считается как $p = 2 * (1 - pt(t_*, n-1))$ или используется общая функция $t.test$ с опцией $paired = TRUE$.

Критерий знаков

	1	2	3	4	5	6	7	8	9	10
Масса до диеты	68	80	92	81	70	79	78	66	57	76
Масса после диеты	60	84	87	79	74	71	72	67	57	70
Потерянный вес	8	−4	5	2	−4	8	6	−1	0	6

Имеется парная выборка (x_{1j}, x_{2j}), $j = 1, \ldots, n$. Рассмотрим $z_j = x_{1j} - x_{2j}$, $j = 1, \ldots, n$, в виде выборки наблюдений случайной разности ζ. Обозначим случайное число плюсов ($\zeta > 0$) и минусов ($\zeta < 0$) через ν_+ и ν_- соответственно, а через n_+ и n_- число наблюдаемых разностей со знаком + и −, $n_+ + n_- = n$. Для определенности будем считать успехом событие $\zeta > 0$, вероятность которого равна $p = P\{\zeta > 0\}$. Нулевая гипотеза имеет вид $H_0 : p = 0.5$, то есть выборки одинаково распределены. Вследствие непрерывности случайных величин, $\xi = 0$ с нулевой вероятностью, поэтому одинаковые наблюдения следует исключить.

При $H_1 : p > 0.5$ число неудач должно быть невелико, поэтому вычисляется вероятность менее n_- неудач, а при гипотезе $H_1 : p < 0.5$ вероятность более n_+ успехов соответственно:

$$\alpha_* = P\{\nu_- \le n_-\} = \sum_{k=0}^{n_-} P\{\nu_- = k\} = \frac{1}{2^n} \sum_{k=0}^{n_-} C_n^k,$$

$$\alpha_* = P\{\nu_+ \ge n_+\} = \sum_{k=n_+}^{n} P\{\nu_+ = k\} = \frac{1}{2^n} \sum_{k=n_+}^{n} C_n^k.$$

При $H_1 : p \ne 0.5$ $\alpha_* = \frac{1}{2^{n-1}} \sum\limits_{k=0}^{\min(n_-, n_+)} C_n^k$. Если $\alpha_* < \alpha$, то гипотеза о случайности в колебаниях знаков разностей отвергается с уровнем значимости α, иначе нет оснований отвергнуть эту гипотезу. [3]

[3] Считая успехом снижение массы, $n_=3$, $n = 9$, при помощи $pbinom(3, prob = 1/2, size = 9)$ получаем $\alpha_* = 0.2539063$. Поэтому H_0 не отвергается в пользу альтернативной $H_1 : p > 0.5$, число неудач мало, но это можно объяснить случайностью, эффект диеты есть, но незначим.

16. Однофакторный дисперсионный анализ

Пусть имеется выборка наблюдений, которая разбивается на r групп, Каждая группа содержит n_i величин $\mathcal{N}(\mu_i, \sigma)$ с одинаковой дисперсией, $n_1 + \ldots + n_r = n$, x_{ij} — j-я величина $(j = 1, \ldots, n_i)$ в i-й группе $(i = 1, \ldots, r)$.

- $\bar{x}_i = \frac{1}{n_i} \sum\limits_{j=1}^{n_i} x_{ij}$ — среднее в i-й группе;
- $\bar{x} = \frac{1}{n} \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{n_i} x_{ij}$ — общее среднее.

Покажем, что общий источник вариации $Q = \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ можно разложить на сумму источников вариации, обусловленных различием между группами Q_1 и внутри групп Q_2.

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 =$$

$$= \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}_{Q_2} + \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2}_{Q_1}, \text{ так как}$$

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) = \sum_{i=1}^{r} (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0\,.$$

Для проверки гипотезы $H_0 : \mu_1 = \ldots = \mu_r = \mu$ используется статистика

$$F = \frac{\frac{1}{r-1} Q_1}{\frac{1}{n-r} Q_2} \;\sim\; \mathbf{F}(r-1, n-r)\,.$$

Для проверки гипотез $H_0 : \mu_i = \mu_k$ значимости отклонений внутригрупповых средних используем статистки

$$t = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{Q_2 \sqrt{\frac{1}{n_i} + \frac{1}{n_k}}} \sqrt{n-r} \;\sim\; T(n-r)\,.$$

В R используется процедура *aov*. Непараметрическим аналогом этого метода является критерий Краскела-Уоллиса *kruskal.test*.

17. Оценка параметров линейной регрессии.

Кривую регрессии для среднего величины η по переменной ξ определяют как траекторию точки $(x, \mathbf{E}(\eta|\xi = x))$. При $g(\xi) = \mathbf{E}(\eta|\xi = x)$ достигается минимум выражения $\mathbb{E}(\eta - g(\xi))^2$. Если в качестве функций $g(\xi)$ рассматривать выражения вида $\alpha + \beta\xi$, то прямая, при которой достигается минимум, называется *линейной средней квадратической регрессией.*

Рассмотрим двумерную выборку наблюдений (x_i, y_i) случайных величин ξ и η, $i = 1, \ldots, n$. Предположим, что

$$y_i = \alpha + \beta x_i + \delta_i \tag{31}$$

где ошибки δ_i такие, что $\mathbf{E}\delta_i = 0, \quad \mathbf{D}\delta_i = \sigma^2, \quad \mathbf{E}\delta_i\delta_j = 0$. Из всего семейства прямых вида $g(x) = \alpha + \beta x$ выбирается „наилучшая“ в том смысле, что

$$Q = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

принимает наименьшее значение.

$$\begin{cases} -\frac{1}{2}\frac{\partial Q}{\partial \alpha} = \sum\limits_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ -\frac{1}{2}\frac{\partial Q}{\partial \beta} = \sum\limits_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \end{cases} \iff$$

$$\iff \begin{cases} \sum\limits_{i=1}^{n} y_i - \alpha n - \beta \sum\limits_{i=1}^{n} x_i = 0 \\ \sum\limits_{i=1}^{n} y_i x_i - \alpha \sum\limits_{i=1}^{n} x_i - \beta \sum\limits_{i=1}^{n} x_i^2 = 0 \end{cases}$$

Учитывая то, что $\sum\limits_{i=1}^{n} x_i = n\bar{x}$ и $\sum\limits_{i=1}^{n} y_i = n\bar{y}$, получаем оценки параметров линейной регрессии:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{32}$$

18. Коэффициент корреляции и его значимость

Если две случайные величины ξ и η независимы, то ковариация между ними равна нулю,

$$cov(\xi,\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = (\mathbb{E}\xi - \mathbb{E}\xi)(\mathbb{E}\eta - \mathbb{E}\eta) = 0\,.$$

Нормированная ковариация или **корреляция** определяется как

$$\varrho = \frac{cov(\xi,\eta)}{\sqrt{\mathbb{D}\xi\mathbb{D}\eta}},$$

где $\mathbb{D}\xi$ и $\mathbb{D}\eta$ дисперсии, поэтому в случае независимых величин $\varrho = 0$. Если между величинами имеется линейная зависимость $\eta = \alpha + \beta\xi$, то

$$cov(\xi,\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\alpha + \beta\xi - \mathbb{E}(\alpha + \beta\xi)) = \beta\mathbb{D}\xi,$$

$$\mathbb{D}\eta = \mathbb{D}(\alpha + \beta\xi) = \beta^2\mathbb{D}\xi, \qquad \varrho = \frac{\beta\mathbb{D}\xi}{\sqrt{\mathbb{D}\xi\beta^2\mathbb{D}\xi}} = \frac{\beta}{|\beta|} = \begin{cases} 1, & \beta > 0, \\ -1, & \beta < 0 \end{cases}$$

Выборочный коэффициент корреляции, полученный по двумерной выборке $(x_1, y_1), \ldots, (x_n, y_n)$, вычисляется по формуле:

$$\hat{\varrho} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}} \tag{33}$$

Таким образом получаем статистику для проверки значимости коэффициента корреляции

$$T = \frac{\hat{\varrho}}{\sqrt{1 - \hat{\varrho}^2}}\sqrt{n-2} \;\sim\; \mathbf{T}(n-2), \tag{34}$$

которая имеет распределение Стьюдента с $n-2$ степенями свободы в случае справедливости нулевой гипотезы $H_0 : \varrho = 0$.

19. Множественная регрессия

19.1. Значимость прогноза

Множественной регрессией переменной Y в зависимости от значений переменных $X_1, \ldots, X_s$ называют условное среднее $E(Y|X_1, \ldots, X_s)$. Модель множественной регрессии имеет вид уравнения $Y = E(Y|X_1, \ldots, X_s) + \varepsilon$, где ε – случайная несистематическая ошибка с нулевым математическим ожиданием и дисперсией σ^2. Для практических целей удобной оказывается линейная аппроксимация множественной регрессии в виде

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_s X_s + \varepsilon.$$

Пусть имеется выборка $\{y_j\}$, $j = 1, \ldots, n$ для зависимой переменной и многомерная выборка $\{x_{1j}, \ldots . x_{sj}\}$ независимых переменных. Частные коэффициенты регрессии $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_s$ вычисляются из условия минимальности выражения $\sum\limits_{j=1}^{n}(y_j - \beta_0 - \beta_1 x_{1j} - \ldots - \beta_s x_{sj})^2$ и выражаются через алгебраические дополнения ковариационной матрицы переменных. Значения наилучшего предсказания переменной Y по $X_1, \ldots, X_s$ обозначены через

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \ldots + \hat{\beta}_s x_{sj}\,.$$

Источником вариации переменной Y является

$$SS_T = \sum_{j=1}^{n}(y_j - \bar{y})^2,$$

Можно показать, $SS_T = SS_R + SS_E$, где

$$SS_R = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2, \quad SS_E = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

– источники вариации, обусловленные соответственно регрессией и ошибкой. Выражение $\hat{\sigma}^2 = \frac{SS_E}{n-s-1}$ используют в качестве несмещенной оценки.

Коэффициентом детерминации называется доля дисперсии, объясняемая регрессией, равная отношению

$$R^2 = \frac{SS_R}{SS_T}.$$

Чем ближе коэффициент детерминации к единице, тем ближе точки наблюдений к наилучшему линейному предсказанию, тем значимее прогноз. Корень из коэффициента детерминации называется **множественным коэффициентом корреляции**. Он равен коэффициенту корреляции между переменной Y и ее наилучшим предсказанием $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_s X_s$. При $s = 1$ множественный коэффициент корреляции равен обычному коэффициенту корреляции.

Значимость коэффициента детерминации, т.е. гипотеза равенства нулю R^2 проверяется при помощи статистики Фишера вида:

$$F = \frac{SS_R/s}{SS_E/(n-s-1)},$$

которая при справедливости $H_0 : R^2 = 0$ имеет распределение Фишера с s и $n - s - 1$ степенью свободы. [4]

19.2. Частный коэффициент корреляции

Разность $Y - \hat{Y}$ между признаком Y и его наилучшим линейным предсказанием $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_s X_s$ по переменным $X_1, \ldots, X_s$ называется *остатком*. Важным свойством остатка является его некоррелированность ни с одной из переменных $X_1, \ldots, X_s$, входящих в предиктор $\hat{Y}(X_1, \ldots, X_s)$.

Это свойство позволяет устранять влияние одних переменных на другие. Рассмотрим наилучшие линейные предсказания переменных Y и X_1 по

[4] Доверительный уровень вероятности $P = P\{F > F_*\}$ В R значимость можно вычислить при помощи функции $1 - pf(F_*, s, n - s - 1)$ или использовать более общую функцию lm. указывает на то, насколько можно доверять прогнозу. Если $P < 0.05$, то гипотеза о равенстве нулю R^2 отвергается, и прогнозу можно доверять с вероятностью, большей 0.95.

переменным $X_2, \ldots, X_s$ $\hat{Y}(X_2, \ldots, X_s)$ и $\hat{X}_1(X_2, \ldots, X_s)$. Коэффициент корреляции $\rho_{yx_1|x_2 \ldots x_s}$ между остатками $Y - \hat{Y}(X_2, \ldots, X_s)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_s)$ называется **частным коэффициентом корреляции**.

В случае трех переменных X_1, X_2, X_3 для вычисления частных коэффициентов корреляции $r_{ij|k}$ по коэффициентам корреляции $r_{ij} = \varrho(X_i, X_j)$ можно использовать непосредственно формулы

$$r_{12|3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13^2}} \sqrt{1 - r_{23}^2}}, \quad r_{13|2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12^2}} \sqrt{1 - r_{23}^2}}.$$

Значимость частного коэффициента корреляции проверяется также как и в случае обычного коэффициента корреляции при помощи статистики Стьюдента вида:

$$T = \frac{\rho_{yx_1|x_2 \ldots x_s}}{\sqrt{1 - \rho^2_{yx_1|x_2 \ldots x_s}}} \sqrt{n - s - 1}, \tag{35}$$

которая в случае $H_0 : \varrho = 0$ имеет распределение Стьюдента с числом степеней свободы, равным $n - s - 1$.[5]

Рассмотрим три признака: X_1 – детская смертность в городах США, X_2 процент афроамериканцев, X_3 – уровень безработицы. Все коэффициенты корреляции значимо отличаются от нуля: $r_{12} = 0.84$ ($p = 2.45 \cdot 10^{-21}$), $r_{13} = 0.32$ ($p = 0.004$), $r_{23} = 0.37$ ($p = 0.001$). Не может ли одна зависимость являться следствием другой? Вычисляем частные коэффициенты корреляции $r_{12|3} = 0.82$ ($p < 0.0001$), $r_{13|2} = 0.03$ ($p = 0.808$). Высокая детская смертность среди афроамериканцев не зависит от того, высокая или низкая безработица в городе, а наблюдаемая корреляция между безработицей и детской смертностью, скорее всего обусловлена тем, что высокая безработца имеет место в городах с высоким процентом афроамериканцев и соответственно с более высоким уровенем детской смертности.

[5] Для вычисления P-значения при полученном наблюдаемом значении статистики Стьюдента T_* можно использовать функцию в электронных таблицах *Excel* вида: $P = P\{|T| > |T_*|\} =$ СТЬЮРАСП($|T_*|; n - s - 1; 2$) или в R функцию $2 * (1 - pt(T_*, n - s - 1))$.

20. Классификация при помощи дискриминантного анализа

20.1. Эвристическая классификация в случае двух популяций

Пусть имеется две нормально распределенные популяции W_1 и W_2 с одинаковой известной ковариационной матрицей Σ и векторами средних μ_1 и μ_2. Задача дискриминантного анализа состоит в том, чтобы решить, к какой из этих популяций отнести вектор наблюдений $X = (x_1, \ldots, x_p)^T$. Для этого ищется так называемая *дискриминантная функция* вида

$$Z = \alpha_1 x_1 + \ldots + \alpha_p x_p = \alpha^T X,$$

где $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ – некоторый вектор коэффициентов, и граничное значение c так, что если $Z \geq c$, то наблюдение X относится к популяции W_1, и иначе в противоположном случае.

Коэффициенты α_i выбираются таким образом, чтобы было максимальным расстояние Махаланобиса между популяциями

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{\sigma^2} = \frac{(\alpha^T(\mu_1 - \mu_2))^2}{\alpha^T \Sigma \alpha}.$$

Фишером было показано, что *максимальное расстояние Махаланобиса доставляет вектор α, который является решением системы уравнений*

$$\Sigma\alpha = \mu_1 - \mu_2. \tag{36}$$

Если $X \in W_2$, но $z \geq c$, то имеет место ошибка классификации $P(1|2)$, если $X \in W_1$, но $z \leq c$, то $P(2|1)$. Для того чтобы эта ошибка была минимальной, выбирается $c = \frac{\zeta_1 + \zeta_2}{2}$. Эта процедура классификации, не зависящая от априорной вероятности i-ой популяции, называется эвристической. [6]

[6] В *R* для того чтобы выполнить эвристическую процедуру, нужно использовать функцию *lda* с опцией $prior = c(1,1)/2$.

20.2. Байесовская классификация в случае k популяций

Задача состоит в том, чтобы отнести некоторый вектор наблюдений $\mathbf{x} = (x_1, \ldots, x_p)^T$ к одной из популяций $W_1, \ldots, W_k$ с произвольным распределением. Обозначим через $q_i = P(W_i)$ априорную вероятность того, что случайный вектор $\mathbf{X} \in W_i$, $f_i(\mathbf{x})$ — плотность распределения популяции W_i, P_{ij} — вероятность отнесения к популяции W_i элемента из популяции W_j. Будем считать, что все параметры известны. Апостериорная вероятность $P(W_j|\mathbf{x})$ выражается по формуле Байеса

$$P(W_j|\mathbf{x}) = \frac{q_j f_j(\mathbf{x})}{\sum\limits_{i=1}^{k} q_i f_i(\mathbf{x})}. \tag{37}$$

Байесовская процедура относит наблюдение $\mathbf{x}$ к той популяции, у которой больше апостериорная вероятность. Если популяция W_i имеет нормальное распределение с параметрами $\mathcal{N}(\mu_\mathbf{i}, \mathbf{\Sigma})$, то вектор $\mathbf{x}$ относится к популяции с максимальным значением дискриминанта

$$\delta_i = \alpha_i^T \mathbf{x} + \gamma_i + \ln q_i,$$

где $\alpha_i = \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$, $\gamma_i = -\frac{1}{2}\mu_\mathbf{i}{}^T \mathbf{\Sigma}^{-1}\mu_\mathbf{i}$.

21. Основы факторного анализа

21.1. Главные компоненты

Рассмотрим ситуацию с большим числом признаков. В большинстве случаев эти признаки коррелированы, поскольку являются результатами частных измерений некоторых интегральных характеристик. Для их обнаружения предназначен метод главных компонент. Идея этого метода заключается в том, чтобы заменить k-мерную случайную величину при наименьшей потере информативности на m- мерную ($m < k$), под которой понимается наилучшее восстановление.

Пусть признаки $X = (X_1, \ldots, X_k)^T$ центрированы, $\mathbf{E}X_i = 0$ и имеют ковариационную матрицу $\Sigma = \mathbf{E}XX^T$. Обозначим через A_i собственные векторы матрицы Σ, соответствующие собственным числам λ_i,

$$\Sigma A_i = \lambda_i A_i, \quad A_i^T A_j = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad i = 1, \ldots, k. \tag{38}$$

Упорядочим собственные числа по убыванию, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. Соберем собственные векторы в одну ортогональную матрицу

$$\mathcal{A} = [A_1, \ldots, A_k] = \begin{bmatrix} a_{11}, & \ldots, & a_{1k} \\ . & . & . \\ a_{k1}, & \ldots, & a_{kk} \end{bmatrix}, \text{ где } \mathcal{A}^T\mathcal{A} = I = \begin{bmatrix} 1, & \ldots, & 0 \\ . & . & . \\ 0, & \ldots, & 1 \end{bmatrix},$$

и определим главную компоненту как линейную комбинацию признаков

$$Y_j = A_j^T X = \sum_{i=1}^{k} a_{ij} X_i. \tag{39}$$

Для всех главных компонент справедливо выражение

$$Y = \mathcal{A}^T X, \text{ откуда } X = \mathcal{A}Y. \tag{40}$$

Теорема 3. *(о некоррелированности главных компонент и инвариантности суммарной дисперсии).*

$$1.\ \ \mathbf{E}Y_iY_j = \begin{cases} \lambda_j, & \textit{при } i=j, \\ 0, & \textit{при } i\neq j \end{cases}, \qquad 2.\ \ \sum_{i=1}^{k}\mathbf{D}X_i = \sum_{j=1}^{k}\mathbf{D}Y_j.$$

Доказательство.

$$1.\ \mathbf{E}Y_iY_j = \mathbf{E}Y_iY_j^T \overset{(39)}{=} \mathbf{E}A_i^TX(A_j^TX)^T = A_i^T\Sigma A_j \overset{(38)}{=}$$

$$= A_i^T\lambda_jA_j = \begin{cases} \lambda_j, & \text{при } i=j, \\ 0, & \text{при } i\neq j. \end{cases}$$

Таким образом $\mathbf{D}Y_j = \lambda_j$ и $\mathbf{D}Y_1 \geq \mathbf{D}Y_2 \geq \ldots \geq \mathbf{D}Y_k$.

$$2.\ \mathbf{D}X_1 + \ldots + \mathbf{D}X_k = \sum_{i=1}^{k}\mathbf{E}X_i^2 = \mathbf{E}X^TX \overset{(40)}{=}$$

$$= \mathbf{E}Y^T\mathcal{A}^T\mathcal{A}Y = \mathbf{E}Y^TY = \sum_{j=1}^{k}\mathbf{E}Y_j^2 = \mathbf{D}Y_1 + \ldots + \mathbf{D}Y_k.$$

21.2. Факторные нагрузки и формула восстановления

Корреляция $\beta_{ij} = cor(X_i, Y_j)$ между признаком X_i и главной компонентой Y_j называется *факторной нагрузкой*. Пусть σ_i — стандартное отклонение X_i, λ_j — дисперсия главной компоненты Y_j. Тогда факторная нагрузка имеет вид

$$\beta_{ij} = cor(X_i, Y_j) = \frac{\mathbf{E}X_iY_j}{\sigma_i\sqrt{\lambda_j}} \overset{(40)}{=} \frac{\mathbf{E}\sum\limits_{t=1}^{k}\alpha_{it}Y_tY_j}{\sigma_i\sqrt{\lambda_j}} = \frac{\alpha_{ij}\sqrt{\lambda_j}}{\sigma_i}\,. \tag{41}$$

Формулу восстановления X_i по главным компонентам можно выразить через факторы $G_j = Y_j/\sqrt{\lambda_j}$:

$$X_i = \sum_{j=1}^{k}\alpha_{ij}Y_j = \sigma_i\sum_{j=1}^{k}\frac{\beta_{ij}}{\sqrt{\lambda_j}}G_j\sqrt{\lambda_j} = \sigma_i\sum_{j=1}^{k}\beta_{ij}G_j\,.$$

ПРИМЕР 1. Признаки: L – средняя продолжительность жизни; M – количество чиновников; A – количество автомобилей; P – доходы бедных; V – объемы продажи водки.

Вклад первого фактора равен 72%. Значения в i-й строке и j-м столбце соответствуют коэффициенту корреляции между i-м признаком и j-й главной компонентой. Чем больше первый фактор, тем больше продолжительность жизни и доходы бедных, меньше чиновников и автомобилей и не много водки – фактор благополучия. С первым фактором имеет значимую корреляцию количество заключаемых браков.

факторные нагрузки

	Factor 1	Factor 2
L	0.896	-0.398
M	-0.815	-0.564
P	0.905	-0.045
A	-0.847	-0.486
V	-0.772	0.613
Дисп.гл.комп. λ	3.60	1.06
вклад в дисп.	72%	21.8%

Значения факторов

годы	f_1	f_2
1970	0.600	0.465
1975	0.540	0.710
1980	0.375	0.790
1985	0.724	-0.339
1990	0.662	-1.301
1995	-1.587	1.019
1998	-1.314	-1.345

Второй фактор менее значим, его вклад равен 22%. Большие значения второго фактора соответствуют ситуации, когда пьют много, автомобилей мало, соответственно малые наоборот: автомобилей много, пьют мало. Фактор имеет значимо положительную корреляцию с числом разводов.[7]

[7] В R для того чтобы выполнить факторный анализ, нужно использовать функцию *princomp*.