Н. П. Алексеева

ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

лекции для студентов 2 курса восточного факультета

Математическая статистика – это естественно научная дисциплина, основной задачей которой является разработка методов получения научно-обоснованных выводов о массовых явлениях и процессах на основе данных наблюдений и экспериментов. В последнее время в области приложений математической статистики наметился большой прогресс, обусловленный распространением использования вычислительной техники и повышением естественно-научного уровня ее пользователей.

При повсеместном использовании результатов статистического анализа важно уметь оценивать степень их неточности, а также иметь представление о том, каким образом они были получены.

1 Основные сведения из теории вероятностей

1.1 Свойства событий

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, что выпадет 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 случайым.

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 $\overline{A}$ называется событием, **противоположным** к A, если события A и $\overline{A}$ несовместны, а их сумма есть достоверное событие. То есть какое-то из A и $\overline{A}$ обязательно происходит: либо A, либо $\overline{A}$. 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).

1.2 Классическая вероятностная схема

Вероятность некоторого события A определяется как отношение числа элементарных исходов, благоприятствующих событию A, к общему числу исходов. Например, всего $n = 20$ экзаменационных вопросов, из которых студент не знает пять. Вероятность того, что достанется незнакомый вопрос, равна $\frac{5}{20} = 0.25$.

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 исходов, благоприятствующих событию A, к общему числу элементарных исходов

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

Именно по этой формуле была вычислена вероятность "плохого" билета. 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). $P(A + B) = P(A) + P(B) - P(AB)$.

Доказательство. Представим сумму $A+B$ в виде суммы двух несовместных событий A и $B\backslash A$ 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:

$$P(A+B+C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC) \quad (3)$$

1.3 Примеры решения задачи

Задача 1. *Подкидываем одновременно два игрального кубика. Найти вероятности следующих событий:*

- *A число очков одинаково на обоих кубиках;*
- *B число очков на одном больше, чем на другом;*
- *C сумма очков четна;*
- *D сумма очков больше двух;*
- *E сумма очков не меньше пяти;*
- *F хотя бы на одном шесть очков;*
- *G произведение очков равно шести.*

Прежде всего вычисляется общее число элементарных исходов. Поскольку каждому из шести элементарных исходов, имеющих место при подбрасывании одного кубика, соответствует шесть элементарных исходов, имеющих место при подбрасывании второго кубика, то общее число элементарных исходов равно $6^2 = 36$.

A	1	2	3	4	5	6
1	*					
2		*				
3			*			
4				*		
5					*	
6						*

C	1	2	3	4	5	6
1	*		*		*	
2		*		*		*
3	*		*		*	
4		*		*		*
5	*		*		*	
6		*		*		*

D	1	2	3	4	5	6
1		*	*	*	*	*
2	*	*	*	*	*	*
3	*	*	*	*	*	*
4	*	*	*	*	*	*
5	*	*	*	*	*	*
6	*	*	*	*	*	*

E	1	2	3	4	5	6
1				*	*	*
2			*	*	*	*
3		*	*	*	*	*
4	*	*	*	*	*	*
5	*	*	*	*	*	*
6	*	*	*	*	*	*

F	1	2	3	4	5	6
1						*
2						*
3						*
4						*
5						*
6	*	*	*	*	*	*

G	1	2	3	4	5	6
1						*
2			*			
3		*				
4						
5						
6	*					

Изобразим элементарные исходы в виде таблицы 6x6 отметим на ней исходы, благоприятствующие соответствующему событию. Используя формулу (1), получаем:

$$P(A) = \frac{6}{36} = \frac{1}{6}, \quad P(B) = 1 - P(A) = \frac{5}{6}, \quad P(C) = \frac{18}{36} = \frac{1}{2},$$

$$P(D) = \frac{35}{36}, \qquad P(E) = \frac{30}{36} = \frac{5}{6}, \qquad P(F) = \frac{11}{36},$$

$$P(G) = \frac{4}{36}.$$

Задача 2. *В аудитории 100 человек. 50 из них знают английский язык (E), 40 французский (F), 35 немецкий (D), 20 английский и французский (EF), 8 английский и немецкий (ED), 10 французский и немецкий (FD), 5 немецкий,английский и французский (EFD). Сколько человек не знает ни одного языка ($\overline{E}\ \overline{D}\ \overline{F}$)?*

Событие $\overline{E}\ \overline{D}\ \overline{F}$ противоположно событию $E + F + D$, которое означает, что человек знает хотя бы один иностранный язык. Воспользуемся формулой (3).

$$P(E + F + D) = P(E) + P(F) + P(D) - P(EF) - P(ED) - P(FD) + P(EFD) =$$
$$0.5 + 0.4 + 0.35 - 0.2 - 0.08 - 0.1 + 0.05 = 0.92.$$

Отсюда $P(\overline{E}\ \overline{D}\ \overline{F}) = 1 - P(EFD) = 1 - 0.92 = 0.08$. Следовательно, 8 человек не знают ни одного иностранного языка.

1.4 Перестановки, размещения, 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. Для того чтобы получить формулу для вычисления

числа сочетаний, требуется определить еще одно понятие. **Размещение** – это упорядоченное m-подмножество из n-множества. Количество размещений A_n^m вычисляется по формуле

$$A_n^m = \underbrace{n \cdot (n-1) \cdot \ldots \cdot (n-m+1)}_{m \text{ сомножителей}} = \frac{n!}{(n-m)!} \tag{4}$$

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**. Очевидно, что число перестановок можно вычислить при помощи частного случая формулы (4)

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

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 (4) получаем формулу для вычисления числа сочетаний:

$$C_n^m = \frac{A_n^m}{m!} = \frac{n!}{(n-m)!m!} \tag{5}$$

Например, количество исходов результата соревнований, в котором нас интересует только тройка призеров, равно $C_5^3 = \frac{60}{3!} = 10$. Эти варианты можно перечислить: $\{1,2,3\}$, $\{1,2,4\}$, $\{1,2,5\}$, $\{1,3,4\}$, $\{1,3,5\}$, $\{1,4,5\}$, $\{2,3,4\}$,$\{2,3,5\}$, $\{2,4,5\}$, $\{3,4,5\}$. Каждому такому сочетанию соответствует сочетание аутсайдеров: $\{4,5\}$, $\{3,5\}$, $\{3,4\}$, $\{2,5\}$, $\{2,4\}$, $\{2,3\}$, $\{1,5\}$,$\{1,4\}$, $\{1,3\}$, $\{1,2\}$. Этот факт есть проявление симметрического числа сочетаний $C_n^m = C_n^{n-m}$. Кроме этого, к свойствам числа сочетаний относятся существенные частные случаи $C_n^0 = C_n^n = 1$ и рекуррентное соотношение $C_n^m = C_{n-1}^m + C_{n-1}^{m-1}$, позволяющие свести количества сочетаний C_n^m в треугольник Паскаля:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & 1 & & & \\
 & & & & 1 & & 2 & & 1 & & \\
 & & & 1 & & 3 & & 3 & & 1 & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

1.5 Конечная схема с неравновозможными исходами

Рассмотрим задачу вычисления вероятности, в которой классической вероятностной схемы с равновозможными исходами оказывается недостаточно.

Это задача вычисления вероятности выигрыша в игре в кости. Азартные игры существовали в древности повсеместно. Об этом свидетельствуют многочисленные

археологические находки. Игральными костями служили кости животных – астрагалы, которые при бросании могли падать на четыре стороны. Эти стороны как-то нумеровались, но единой системы не было. 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. После многократных исследований различных астрагалов были получены следующие частоты выпадения различных сторон. Частота выпадения широкой стороны A с углублением примерно равна 0.39, следующей по величине B – 0.37; частоты выпадения двух оставшихся сторон C и D – по 0.12.

Общее число исходов равно $4^4 = 256$, но вероятности их появления разные. Например, вероятность того, что на четырех астрагалах одна и та же сторона A, по формуле (8) равна $P(AAAA) = 0.39^4 = 0.0231$, а 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 вероятность броска "Венера" , равную $24 \cdot 0.002 = 0.05$.

В общем виде эта задача решается таким образом: имеется конечное число неравновозможных элементарных исходов $\Omega = \{\omega_1, \ldots, \omega_n\}$ с известными вероятностями $p_i = P(\omega_i)$, $i = 1, \ldots, n$, $\sum\limits_{i=1}^{n} p_i = 1$. Вероятность составного события $A = \{\omega_{i_1}, \ldots, \omega_{i_m}\}$ вычисляется суммированием соответствующих вероятностей $P(A) = \sum\limits_{j=1}^{m} p_{i_j}$.

1.6 Счетная схема с неравновозможными исходами

В случае счетного количества элементарных исходов $\Omega = \{\omega_1, \ldots, \omega_n, \ldots\}$ с известными вероятностями $p_i = P(\omega_i)$, $i = 1, \ldots, \infty$, $\sum\limits_{i=1}^{\infty} p_i = 1$, вероятность составного события $A = \{\omega_{i_1}, \ldots, \omega_{i_m}, \ldots\}$ также вычисляется суммированием соответствующих вероятностей $P(A) = \sum\limits_{j=1}^{\infty} p_{i_j}$.

Задача 3. *Играют двое, по очереди вынимая наугад игральную карту из полной колоды. Тот, кто первый вытащит красную карту, выиграл, в противном случае карта возвращается в колоду, и право хода передается другому игроку. Нужно найти вероятность выигрыша игрока, который начинает первым. Обозначим через p вероятность события, которое можно назвать успехом – в данном случае это*

карта красной масти. Если колода полная, то $p = 0.5$. *Вероятность события, противоположного успеху, обозначим через* $q = 1 - p$.

Игра может закончиться на первом, втором, третьем и т.д. шаге. Когда она закончится, неизвестно. Такое множество исходов $\Omega = \{\omega_1, \ldots, \omega_n, \ldots\}$ и называется счетным. Вычислим вероятности элементарных исходов.

$$\begin{aligned} p_1 = P(\omega_1) &= p, \\ p_2 = P(\omega_2) &= qp, \\ p_3 = P(\omega_3) &= q^2 p, \\ &\ldots \\ p_n = P(\omega_n) &= q^{n-1} p, \\ &\ldots \end{aligned}$$

Проверим равенство единице суммы вероятностей элементарных исходов, используя выражение для суммы элементов геометрической прогрессии.

$$p + qp + q^2 p + \ldots = p(1 + q + q^2 + \ldots) = \frac{p}{1-q} = 1.$$

Составное событие $A = \{\omega_1, \omega_3, \ldots, \omega_{2k+1}, \ldots\}$, означающее выигрыш игрока, начинающего первым, состоит из нечетных элементарных исходов:

$$\begin{aligned} P(A) = p + q^2 p + q^2 p + \ldots = p(1 + q^2 + q^4 + \ldots) = \\ = \frac{p}{1-q^2} = \frac{1-q}{(1-q)(1+q)} = \frac{1}{1+q} = \frac{1}{2-p}. \end{aligned}$$

При равенстве вероятности успеха $p = 0.5$ вероятность выигрыша игрока, начинающего первым, равна 2/3. При увеличении вероятности успеха p увеличивается вероятность выигрыша первого игрока, при уменьшении – уменьшается. Например, при $p = 1/6$ и $p = 1/9$ имеем соответственно $P(A) = 6/11$ и $P(A) = 9/17$. Таким образом, при уменьшении вероятности успеха шансы выигрыша первого и второго игроков выравниваются.

1.7 Геометрическая вероятность

Инструментом одной из азартных игр является рулетка – круг, поделенный на сектора, по которому благодаря вращательному механизму движется стрелка. Для вычисления вероятности событий, связанных с остановкой стрелки в некотором заданном секторе, оказываются недостаточными перечисленные ранее схемы. В таком случае используется схема геометрической вероятности, в которой множество элементарных исходов Ω представляет собой множество точек на плоскости, имеющее площадь (для рулетки – множество точек на окружности, длина которой может быть измерена). На множестве всех подмножеств S множества Ω, также имеющих площадь, определяем вероятность попадания в A как числовую функцию, равную отношению площадей $P(A) = \frac{mes(A)}{mes(\Omega)}$, $A \in S$, со значениями на интервале $[0; 1]$,

$P(\Omega) = 1$, для которой вероятность попадания в множество суммы $A + B$ двух непересекающихся множеств равна $P(A + B) = P(A) + P(B)$.

Задача 4. *Чему равна вероятность того, что при случайном прокручивании часового механизма минутная стрелка остановится между 1 и 2 или 5 и 7?*

Длина окружности вычисляется по формуле $2\pi R$, где R – радиус циферблата. Длина сектора между 1 и 2 равна $\frac{2\pi R}{12} = \frac{\pi}{6}$, между 5 и 7 равна $\frac{\pi R}{3}$. Таким образом искомая вероятность равна сумме отношений $\frac{\pi R/6}{2\pi R} + \frac{\pi R/3}{2\pi R} = \frac{1}{12} + \frac{1}{6} = \frac{1}{4}$.

Задача 5. *Двое договорились о встрече в условленном месте следующим образом. Каждый приходит в случайный момент времени с 11 до 12 часов, находится там 20 минут и уходит. Чему равна вероятность того, что встреча состоится?*

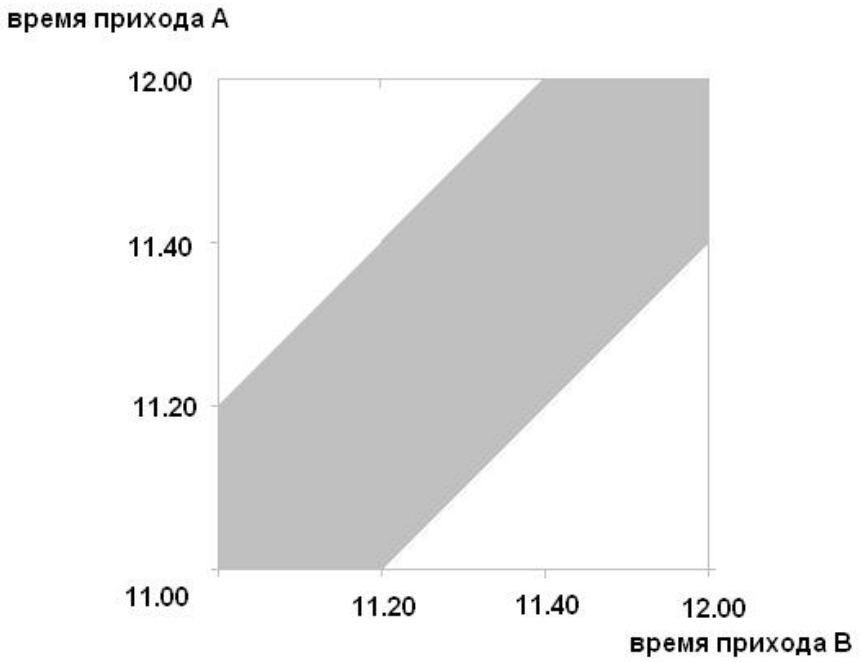

Рис. 1: Геометрическое решение задачи 5.

На рис.1 точкам с координатами (x, y) соответствуют моменты времени прихода каждого из участников встречи. Затемненная область соответствует событию, заключающемуся в том, что встреча состоялась. Вероятность встречи равна $1 - \frac{4}{9} = \frac{5}{9}$, так как площадь единичного квадрата равна единице, а площадь не затененных треугольников равна $2 \cdot \frac{(2/3)^2}{2} = \frac{4}{9}$.

1.8 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:

$$P(A|B) = \frac{r}{m} = \frac{r/n}{m/n} = \frac{P(AB)}{P(B)} \tag{6}$$

Отсюда получаем формулу для вычисления вероятности произведения

$$P(AB) = P(A|B)P(B). \tag{7}$$

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 (6) получаем, что для независимых событий вероятность произведения независимых событий равна произведению вероятностей этих событий:

$$P(A) = P(A|B) = \frac{P(AB)}{P(B)}\,, \quad \Rightarrow \quad P(AB) = P(A)P(B)\,. \tag{8}$$

Задача 2.(продолжение) *Являются 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 $P(E)P(D) = 0.5 \cdot 0.35 = 0.175 = 0.08$, 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 $P(D|E) = 8/50 = 0.16 < 0.35$ и $P(D|\overline{E}) = 27/50 = 0.54 > 0.35$, получаем, что более высокую вероятность знания немецкого языка среди студентов, которые не знают английский язык.

1.9 Формула полной вероятности и формула Байеса

События $A_1, \ldots, A_n$ образуют **полную группу**, если появление одного события исключает появление другого ($A_i A_j = \emptyset,\ i \neq j$), 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**: $P(A_2) = 0.375$, $P(A_3) = 0.209$, $P(A_4) = 0.079$.

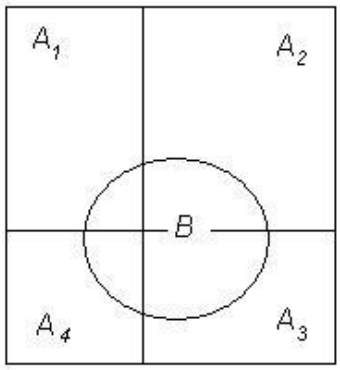

Рис. 2: $A_1, \ldots, A_4$ – полная группа событий. $B = BA_1 + BA_2 + BA_3 + BA_4$.

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.2) 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 вероятности умножения событий (7) вида $P(BA_k) = P(B|A_k)P(A_k)$.

Поскольку больному с первой группой можно переливать только первую группу крови, то $P(B|A_1) = 0.337$. 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, при случайном переливании крови

$$\begin{aligned} P(B) = P(B|A_1)P(A_1) + \ldots + P(B|A_4)P(A_4) = \\ = 0.337 \cdot 0.337 + 0.712 \cdot 0.375 + 0.546 \cdot 0.209 + 1 \cdot 0.079 = \\ = 0.1136 + 0.2671 + 0.1141 + 0.0790 = 0.5737. \end{aligned}$$

В общем виде для любой полной группы событий $A_1, \ldots, A_n$ **формула полной вероятности** имеет вид

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)\,. \tag{9}$$

Апостериорные (буквально, после опыта) вероятности $P(A_i|B)$ из формулы условной вероятности (6) 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:

$$\begin{aligned} P(A_1|B) = \frac{P(A_1B)}{P(B)} = \frac{0.1136}{0.5737} = 0.1980, \\ P(A_2|B) = \frac{P(A_2B)}{P(B)} = \frac{0.2671}{0.5737} = 0.4654, \\ P(A_3|B) = \frac{P(A_3B)}{P(B)} = \frac{0.1141}{0.5737} = 0.1999, \\ P(A_4|B) = \frac{P(A_4B)}{P(B)} = \frac{0.0790}{0.5737} = 0.1387. \end{aligned}$$

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

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

Задача 6. *Надежность выявления туберкулеза при флюрофотометрии равна 90%. Вероятность ошибочного диагноза равна 1%. Распространенность туберкулеза равна 0.1%. Чему равна вероятность того. что человек с "плохой" флюрофотометрией на самом деле болен?* Выделим полную группу событий: A_1 – человек здоров, A_2 – человек болен. Априорные вероятности: $P(A_1) = 0.999$, $P(A_2) = 0.001$. Событие B – "плохая" флюрофотометрия. Если человек здоров, то вероятность ошибки ("плохой" флюрофотометрии) равна 1%, то есть $P(B|A_1) = 0.01$. Если человек болен, то вероятность правильной диагностики равна $P(B|A_2) = 0.9$. Сначала вычислим вероятность "плохой" флюрофотометрии о формуле полной вероятности (9):

$$P(B) = 0.999 \cdot 0.01 + 0.001 \cdot 0.9 = 0.00999 + 0.0009 = 0.01089.$$

Отсюда получаем вероятность того. что человек с "плохой" флюрофотометрией на самом деле болен, $P(A_2|B) = \frac{0.0009}{0.0189} = 0.0826$. Так что только 8% лиц оказываются больными на самом деле.

1.10 Задача де Мере

Существует исторический анекдот, связанный с постижением теории вероятностей шевалье де Мере, который предлагал заключить пари на то, что за 4 подбрасывания игральной кости (обычный шестигранный кубик) хотя бы один раз кубик выпадет стороной с шестью очками (событие B).

Обозначим через A_t событие, которое заключается в том, что шесть очков выпадет t раз, $t = 0, 1, 2, 3, 4$. Событие $\Omega = A_0 + A_1 + A_2 + A_3 + A_4$ является достоверным, оно обязательно произойдет, так кубик либо ни разу не упадет шестиочковой стороной, либо это произойдет один, два, три или 4 раза. $P(\Omega) = 1$. Событие B (хотя бы раз 6 очков) можно выразить как сумму $B = A_1 + A_2 + A_3 + A_4$. Вероятность $P(B)$ можно сосчитать двумя способами: либо найти вероятности $P(A_t)$ при $t = 1, 2, 3, 4$ и их сложить, либо как вероятность противоположного события $P(B) = 1 - P(A_0)$. Событие A_0 означает, что за четыре попытки сторона с шестью очками так и не появится. Вероятность того, что при одном испытании кубик упадет на любую из пяти сторон кроме шестиочковой, равна $\frac{5}{6}$, тогда $P(A_0) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = 0.48$. Следовательно, $P(B) = 1 - 0.48 = 0.52$. Эта вероятность не намного, но больше половины.

Как гласит анекдот, де Мере решил, что нашел путь к обогащению, но с ним довольно скоро никто не стал играть. Тогда он придумал другое пари и для двух кубиков увеличил число испытаний до 24. Ошибочно он вычислил, что при 24-х подбрасываниях двух кубиков с вероятностью $\frac{2}{3}$ хотя бы раз выпадут две "шестерки", однако экспериментально установил вероятность, близкую к половине. Об этом несовпадении он сообщил при встрече Блезу Паскалю, который в свою очередь обсуждал эту задачу в переписке с Пьером де Ферма. Оба эти ученые подтвердили правильность экспериментального результата шевалье де Мере. Поскольку вероятность невыпадения двух "шестерок" одновременно равна $\frac{35}{36}$, то вероятность того, что за 24 раза "две шестерки" так и не выпадут ни разу, равна $\left(\frac{35}{36}\right)^{24} \approx 0.51$, следовательно, хотя бы один раз они выпадут с вероятностью 0.49.

1.11 Биномиальный закон распределения

Обозначим через ξ случайное число выпадения шести очков игрального кубика в $n = 4$ независимых испытаниях, $p_k = P\{\xi = k\}$, $k = 0, 1, 2, 3, 4$ через $p = \frac{1}{6}$ вероятность выпадения шести очков (успех), через $q = 1 - \frac{1}{6} = \frac{5}{6}$ вероятность невыпадения шести очков (неудача), через C_n^k обозначено число сочетаний по k из n

элементов, которое вычисляется по (5).

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	● ● ●●	$C_4^4 p^4 q^0 = \left(\frac{1}{6}\right)^4$

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

Если известны всевозможные исходы $x_1, x_2, \ldots, x_N$ случайного эксперимента ξ, а также вероятности $p_i = P(\xi = x_i),\ i = 1, \ldots, N$, то говорят, что задан **закон распределения** случайного эксперимента

$$\xi : \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ p_1 & p_2 & \ldots & p_N \end{pmatrix}. \tag{10}$$

Случайная величина ξ, равная случайному числу успехов, имеет биномиальный закон распределения:

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n. \tag{11}$$

На рисунке 3 один графиков (столбики) представляет собой зависимость между числом успехов k и вероятностями $P\{\xi = k\}$ в случае $n = 100$ и $p = 0.4$.

1.12 Нормальный закон распределения

При увеличении числа испытаний n с вероятностью успеха p биномиальнй закон распределения может быть выражен аналитически в виде функции

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{12}$$

где $\mu = np$, $\sigma = \sqrt{np(1-p)}$. Это означает, что если число испытаний велико, то для случайного числа успехов ξ из n независимых испытаний вероятность $P\{\xi =$

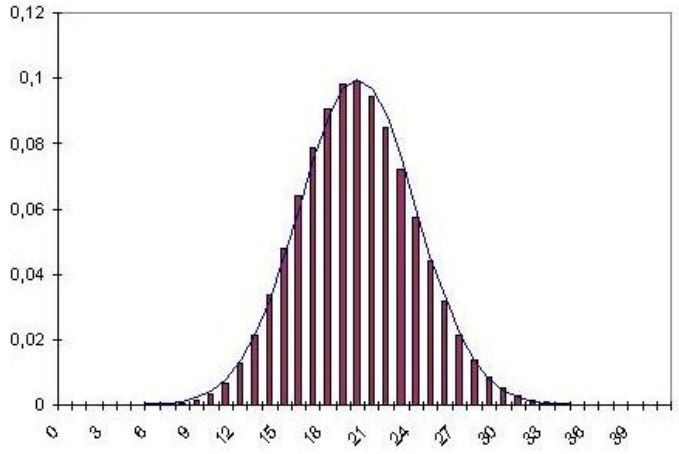

Рис. 3: Биномиальное распределение с параметрами $n = 100$ и $p = 0.2$ и нормальное распределение с параметрами $\mu = 20$, $\sigma^2 = 16$.

$k\}$ может быть вычислена двумя способами: либо при помощи формулы (11) биномиального закона распределения, либо при помощи формулы (12) нормального закона распределения, где $x = k$. Этот факт в теории вероятностей называется **теоремой Муавра-Лапласа** (рис.3). Например, $p = 0.2$ – вероятность того, что молодой человек к 20 годам не курит, ξ – случайное число некурящих из 100 случайно выбранных двадцатилетних. Вычислим вероятность того, что из них 15 окажется некурящими. По формуле (11) эта вероятность равна 0.048, по формуле (12) равна 0.046.

Графики функции (12), которая называется **плотностью** нормального распределения, при различных значениях параметров μ и σ изображены на рисунке 4. Максимальное значение плотности достигается в точке μ, а в точки $\mu - \sigma$ и $\mu + \sigma$ являются точками перегиба. Чем меньше σ, тем менее вариабельнее случайные наблюдения и тем плотнее они концентрируются около μ.

Для любой случайной величины ξ функция $F(x) = P\{\xi \leq x\}$ называется **функцией распределения**. Если она известна, то можно вычислить вероятности попадания случайной величины в любой интервал,

$$P\{a < \xi \leq b\} = F(b) - F(a). \tag{13}$$

Таблица значений функции $\Phi(x)$ **стандартного нормального распределения**, которое имеет место при $\mu = 0$ и $\sigma = 1$, представлена в таблице 7. В электронных таблицах Excel эта функция вызывается через НОРМРАСП.

Чаще всего таблицей 7 пользуются для вычисления квантилей распределения. **P-квантиль** – это число, вероятность не превысить которое случайной величине Z равна P (рис. 5). $P\{Z \leq Z_P\} = P$ равно площади, ограниченной плотностью распределения и вертикальной прямой, проходящей через точку Z_P. Например, 0.975-квантилью стандартного нормального распределения является число 1.96. В таблице 7 это число находится на пересечении строки 1.9 и столбца 0.06. Для вычисления

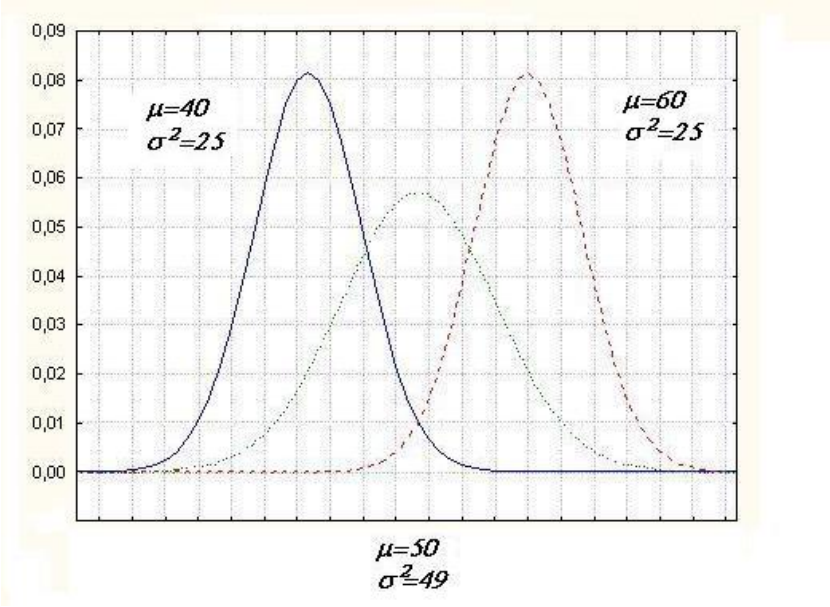

Рис. 4: Кривые плотностей нормального закона распределения с параметрами $\mu = 40$, $\sigma^2 = 25$; $\mu = 60$, $\sigma^2 = 25$; $\mu = 50$, $\sigma^2 = 49$.

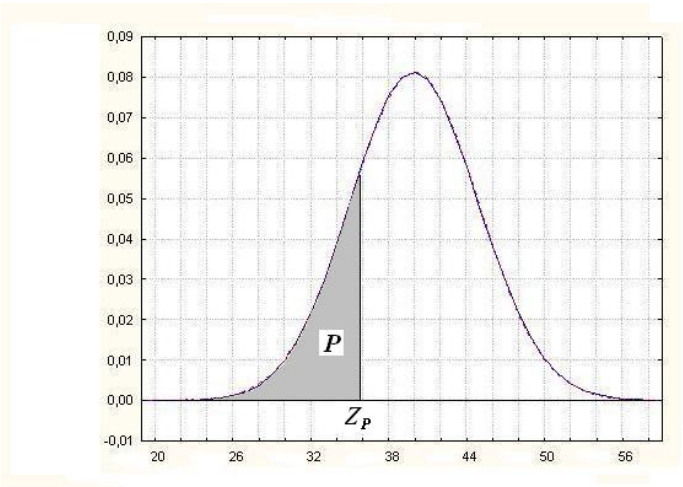

Рис. 5: Иллюстрация P-квантили распределения.

P-квантили при $P < 0.5$ можно воспользоваться тем, что плотность стандартного нормального распределения является функцией, симметричной относительно нуля, и тем, что вся площадь, ограниченная плотностью распределения равна 1, то есть $\Phi(-x) = 1 - \Phi(x)$. В частности, $\Phi(-1.96) = 1 - \Phi(1.96) = 1 - 0.975 = 0.025$. Согласно

(13), для стандартно нормально распределенной случайной величины Z справедливо

$$P\{-1.96 < Z <\leq 1.96\} = \Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95.$$

Это означает, что с вероятностью 0.95 значения случайной величина сосредоточены на интервале от -1.96 до 1.96.

Для вычисления вероятностей функции распределения $F(x|\mu,\sigma)$ с параметрами $\mu \neq 0$ и $\sigma \neq 1$ можно воспользоваться формулой

$$F(x|\mu,\sigma) = P\{\xi \leq x\} = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{14}$$

В частности, на практике иногда важнее бывает определить не вероятность того, что число некурящих точно равно 15, а вероятность того, некурящих будет не больше 15. Эта вероятность равна площади, ограниченной функцией (12) при $\mu = np = 100 \cdot 0.2 = 20$, $\sigma = \sqrt{np(1-p)} = \sqrt{16} = 4$ слева до значения $x = 15$. Обозначим через X случайное число курящих студентов.

$$P\{X \leq 15\} = \Phi\left(\frac{15-20}{\sqrt{16}}\right) = \Phi(-1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056.$$

Для вычисления $\Phi(1.25)$ мы воспользовались таблицей 7, значение 0.8944 лежит на пересечении строки 1.2 и столбца 0.05. В электронных таблицах Excel можно воспользоваться функцией $P\{X \leq 15\}$=НОРМРАСП(15;20;4;1). Число 15 является 0.1056-**квантилью** нормального закона с параметрами $\mu = 20,\ \sigma^2 = 16$.

Для того чтобы сравнить, насколько велико расхождение между биномиальным и нормальным распределением при числе испытаний, равном $n = 100$, нужно вычислить вероятности $P\{\xi = k\}$ для $k = 0, 1, \ldots, 15$ при помощи формулы (11) и затем их сложить. Без использования вычислительной техники вычисления оказываются очень трудоемкими, ответ 0.128.

Задача 7. *В некотором курортном городке с двумя одинаковыми ресторанами в разгар сезона проживают* $n = 1000$ *туристов. Чему равна вероятность того, что при наличии 500 мест туриста не смогут обслужить? Сколько мест нужно иметь в ресторане, чтобы с вероятностью* 0.99 *случайно зашедший турист смог пообедать?*

Обозначим через ξ случайное число посетителей, через M число мест. $p = 0.5$, $np = 500$, $np(1-p) = 1000 \cdot 0.5 \cdot 0.5 = 250$, $\sqrt{250} \approx 16$. Пусть $M = 500$.

$$P\{\xi > 500\} = 1 - P\{\xi \leq 500\} = 1 - \Phi\left(\frac{500-500}{16}\right) = 1 - \Phi(o) = 0.5.$$

Для ответа на второй вопрос используем выражение (14).

$$P\{\xi \leq M\} = 0.99 \quad \Leftrightarrow \quad \Phi\left(\frac{M-np}{\sqrt{np(1-p)}}\right) = 0.99\,.$$

По таблице 7 находим, что вероятности 0.9901 соответствует квантиль, равная 2.33, следовательно, $\frac{M-np}{\sqrt{np(1-p)}} = 2.33$, откуда $M = np + 2.33\sqrt{np(1-p)} = 500 + 2.33 \cdot 16 = 537.28$. Таким образом дополнительные 37 стульев могут существенно улучшить в городе обслуживание туристов.

1.13 Распределения, связанные с нормальным

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
	$\eta = \sum\limits_{i=1}^{n} \xi_i^2$	$\zeta = \dfrac{\xi_0}{\sqrt{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}}$	$\vartheta = \dfrac{\frac{1}{n}\sum\limits_{k=1}^{n} \xi_k^2}{\frac{1}{m}\sum\limits_{k=1}^{m} \eta_k^2}$

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

Современное программное обеспечение позволяет найти вероятность попадания случайной величины η в любой интервал. Например, для любого $x_0 > 0$

$$P(\eta > x_0) = \text{ХИ2РАСП}(x_0; n),$$
$$P(\zeta > x_0) = \text{СТЬЮДРАСП}(x_0; n; 1),$$
$$P(\vartheta > x_0) = F\text{РАСП}(x_0; n; m).$$

При увеличении числа степеней свободы распределение Стьдента становится все более похожим на стандартное нормальное распределение, кривая которого на рисунке 6 изображена тонкой линией.

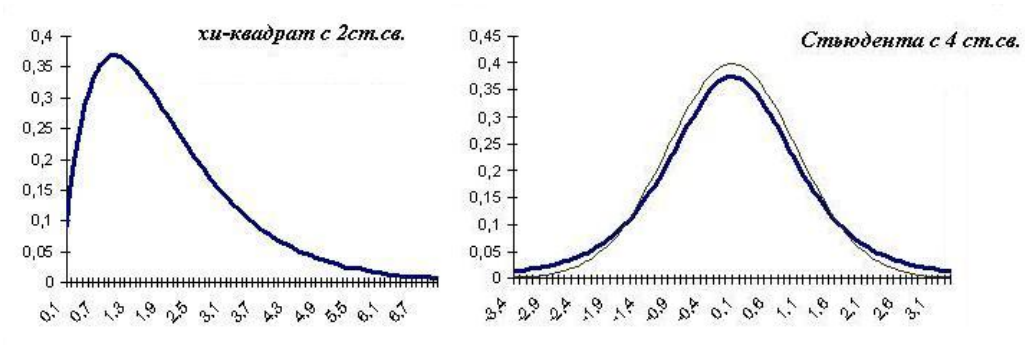

Рис. 6: Плотности распределения хи-квадрат и Стьюдента соответственно с 2-мя и 4-ми степенями свободы.

Для вычисления квантилей этих распределений используются другие функции. По определению, P-квантиль $\chi^2_P(n)$ распределения хи-квадрат с n степенями свободы удовлетворяет соотношению $P\{\chi^2 < \chi^2_P(n)\} = P$, следовательно,

$\chi^2_P(n) = \text{ХИ2ОБР}(1-P; n)$, $\quad \text{ХИ2ОБР}(0.05; 1) = 3.84 = \chi^2_{0.95}(1)$,
$\eta_P(n) = \text{СТЬЮДРАСПОБР}(2(1-P); n)$, $\quad \text{СТЬЮДРАСПОБР}(0.1; 10) = 1.81 = \eta_{0.95}(10)$,
$\vartheta_P(n,m) = F\text{РАСПОБР}(1-P; n; m)$, $\quad F\text{РАСПОБР}(0,05; 10; 10) = 2.98 = \vartheta_{0.95}(10, 10)$.

2 Статистические методы

2.1 Построение гистограммы

Математическая статистика позволяет получать обоснованные выводы о свойствах случайных величин по конечной совокупности наблюдений над ними – **выборке**. Закон распределения наблюдаемой случайной величины называется **генеральным**.

Выборка представляет собой последовательность чисел $x_1, \ldots, x_n$, полученных на практике при n-ратном повторении эксперимента в неизменных условия. Например, в результате измерений депрессии по шкале Зунга у студентов дневных отделений университета была получена выборка объема $n = 83$.

Перейдем к упорядоченной выборке, которую называют **вариационным рядом**. Оказывается, что один студент имеет оценку депрессии 24, двое 27, еще двое 28, еще двое 29, один 29,5, трое 30 и т.д. Если количество **градаций** невелико, то эту процедуру можно продолжить и представлять данные в виде вариационного ряда. Однако в данном случае имеется 34 градации, поэтому есть смысл их сгруппировать по некоторым интервалам, лучше одинаковым.

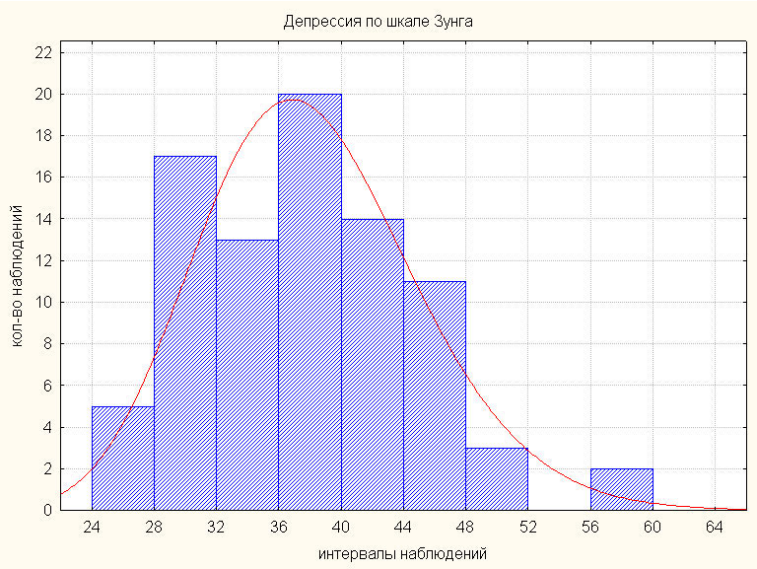

Рис. 7: Гистограмма признака депрессии по шкале Зунга.

Длина **интервала группировки** является управляющим параметром, поэтому выбор подходящей длины и количества интервалов определяется чаще всего здравым смыслом или опытом работы с данными. Если материал не знакомый, то вычисляют **размах выборки** R, равный расстоянию между максимальной и минимальной градациями, и длину интервала группировки получают как отношение размаха к количеству интервалов порядка десяти и выше в зависимости от объема выборки. Длину интервала для удобства можно округлить, а крайние интервалы ("хвосты") объединить, если количество наблюдений в них не превышает, например, пяти.

Рассмотрим интервалы длиной 4, начиная с минимального значения, равного 24, и сосчитаем количество наблюдений, попадающих в каждый из интервалов, и построим **гистограммы**: исходную и с ученным "хвостом", описывающие закон распределения исследуемой характеристики. График гистограммы представлен на рисунке 7.

интервалы	кол-во набюдений
≤ 28	5
$(28; 32]$	17
$(32; 36]$	13
$(36; 40]$	20
$(40; 44]$	14
$(44; 48]$	11
$(48; 52]$	3
$(52; 56]$	0
$(56; 60]$	2
всего	83

интервалы	кол-во набюдений
≤ 28	5
$(28; 32]$	17
$(32; 36]$	13
$(36; 40]$	20
$(40; 44]$	14
$(44; 48]$	11
> 48	5
всего	83

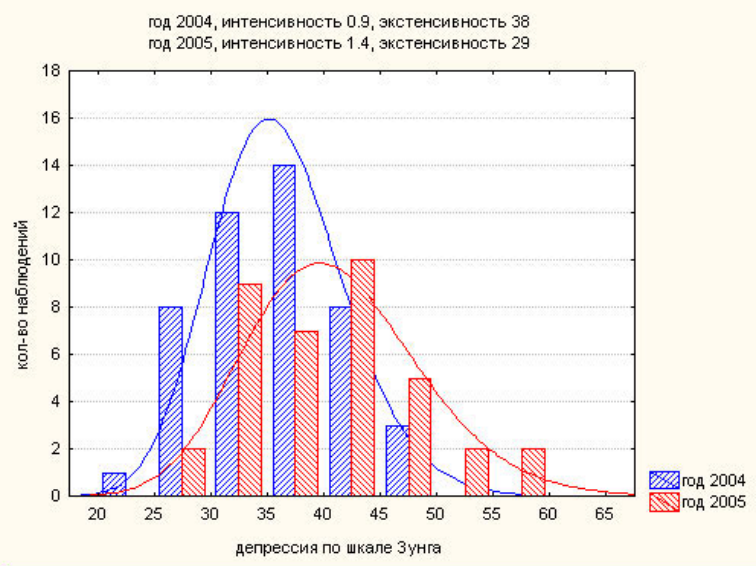

Рис. 8: Гистограммы признака депрессии по шкале Зунга, соответствующего разным годам.

Построим две гистограммы (рис.8), соответствующие одному и тому же признаку депрессии по шкале Зунга, но в разные годы: 2004 и 2005. Обращает на себя факт, что одна гистограмма сдвинута правее, и ее размах больше. Это соответствует тому, что в среднем оценка депрессии выше в 2005 году, и среди этих студентов наблюдается большее разнообразие. Численно эти различия можно измерить при помощи характеристик среднего и дисперсии.

2.2 Среднее, дисперсия

Основной характеристикой центра положения случайной величины ξ, распределенной по закону

$$\xi : \begin{pmatrix} x_1 & x_2 & \dots & x_N \\ p_1 & p_2 & \dots & p_N \end{pmatrix},$$

является **среднее значение** μ, которое вычисляется при помощи выражения

$$\mu = x_1 p_1 + \ldots + x_N p_N. \tag{15}$$

Поясним это выражение на примере. Всем хорошо знакома ситуация, когда в школе в конце четверти делаются прогнозы относительно того, какая оценка будет выведена по тому или иному предмету. Пусть случайные оценки по литературе ξ_1 и физике ξ_2 подчинены законам

$$\xi_1 : \begin{pmatrix} 2 & 3 & 4 & 5 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{pmatrix} \qquad \xi_2 : \begin{pmatrix} 2 & 3 & 4 & 5 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix},$$

которым могут соответствовать наборы оценок $2, 3, 3, 4, 4, 4, 5, 5, 5, 5$ и $2, 2, 2, 2, 3, 3, 3, 4, 4, 5$ со средними арифметическими 4 и 3. Непосредственными вычислениями убеждаемся, что средние значения случайных величин действительно равны

$$\mu_1 = 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.3 + 5 \cdot 0.4 = 0.2 + 0.6 + 1.2 + 2 = 4,$$
$$\mu_2 = 2 \cdot 0.4 + 3 \cdot 0.3 + 4 \cdot 0.2 + 5 \cdot 0.1 = 0.8 + 0.9 + 0.8 + 0.5 = 3\,.$$

Если имеется выборка наблюдений $x_1, \ldots, x_n$, то **выборочное среднее** обозначается через $\bar{x}$ или $\hat{\mu}$ и вычисляется как среднее арифметическое:

$$\bar{x} = \sum_{i=1}^{n} x_i. \tag{16}$$

Характеристика вариабельности данных – **дисперсия** σ^2 – определяется как средний квадрат отклонения от среднего значения

$$\sigma^2 = (x_1 - \mu)^2 p_1 + \ldots + (x_N - \mu)^2 p_N. \tag{17}$$

В качестве оценки дисперсии по выборке $x_1, \ldots, x_n$ обычно используют выражение

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right). \tag{18}$$

Наряду с дисперсией σ^2 используют также характеристику **стандартного отклонения** σ, которое равно квадратному корню из дисперсии. Оценкой стандартного отклонения по выборке $x_1, \ldots, x_n$ является S.

Для вычисления дисперсии случайных оценок по литературе и физике ξ_1 и ξ_2 рассмотрим квадраты отклонения от среднего $(\xi_1 - \mu_1)^2$ и $(\xi_2 - \mu_2)^2$

$$(\xi_1 - \mu_1)^2 : \begin{pmatrix} (-2)^2 & (-1)^2 & 0 & 1^2 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{pmatrix} \quad (\xi_2 - \mu_2)^2 : \begin{pmatrix} -1)^2 & 0 & 1^2 & 2^2 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}$$

и вычислим их средние значения, которые являются дисперсиями σ_1^2 и σ_2^2 случайных величин ξ_1 и ξ_2:

$$\sigma_1^2 = 4 \cdot 0.1 + 1 \cdot 0.2 + 0 \cdot 0.3 + 1 \cdot 0.4 = 0.4 + 0.2 + 0 + 0.4 = 1,$$
$$\sigma_2^2 = 1 \cdot 0.4 + 0 \cdot 0.3 + 1 \cdot 0.2 + 4 \cdot 0.1 = 0.4 + 0 + 0.2 + 0.4 = 1\,.$$

Итак, оценки по физике и литературе отличаются по средним значениям и имеют одинаковые дисперсии. Приведем пример распределения с меньшей дисперсией. Предположим, что случайная оценка по физкультуре ξ_3 имеет распределение

$$\xi_3 : \begin{pmatrix} 2 & 3 & 4 & 5 \\ 0 & 0.1 & 0.8 & 0.1 \end{pmatrix}.$$

Среднее значение равно $\mu_3 = 2{\cdot}0{+}3{\cdot}0.1{+}4{\cdot}0.8{+}5{\cdot}0.1 = 0{+}0.3{+}3.2{+}0.5 = 4$, дисперсия равна $\sigma_3^2 = (2{-}4)^2{\cdot}0{+}(3{-}4)^2{\cdot}0.1{+}(4{-}4)^2{\cdot}0.8{+}(5{-}4)^2{\cdot}0.1 = 0{+}0.1{+}0{+}0.1 = 0.2$. То есть вариабельность возможных оценок по физкультуре $5, 4, 4, 4, 4, 4, 4, 4, 4, 3$ в пять раз меньше вариабельности оценок по физике и литературе.

2.3 Доверительный интервал и ошибка среднего

Выборочное среднее $\bar{x}$ по наблюдениям $x_1, \ldots, x_n$ со средним значением μ и дисперсией σ^2, как нетрудно заметить, само является случайной величиной. Рассмотрим динамику изменения показателей уровня депрессии по шкале Зунга студентов восточного факультета.

	2004	2005	2006	2007
$\bar{x}$	35.9	40.9	38.1	41.0
S^2	49.8	59.5	36.0	85.2
n	15	38	31	14
$S/\sqrt{n}$	1.8	1.3	1.1	2.5

В математической статистике известно, что среднее значение случайного выборочного среднего $\bar{x}$ равняется генеральному средниму μ, а его дисперсия равна $\frac{\sigma^2}{n}$. По данным динамики уровня депрессии хорошо заметно значительное снижение вариабельности выборочного среднего по сравнению с вариабельностью всей выборки. Стандартное отклонение случайного выборочного среднего $\bar{x}$, равное $\frac{\sigma}{\sqrt{n}}$ называют **ошибкой среднего**. В качестве его оценки используют отношение $\frac{S}{\sqrt{n}}$. В статистических отчетах принято представлять средние значения в сопровождении ошибки среднего: $\bar{x} \pm \frac{S}{\sqrt{n}}$, например, средний уровень депрессии в 2004 году равен 35.9 ± 1.8, в 2005 году 40.9 ± 1.3.

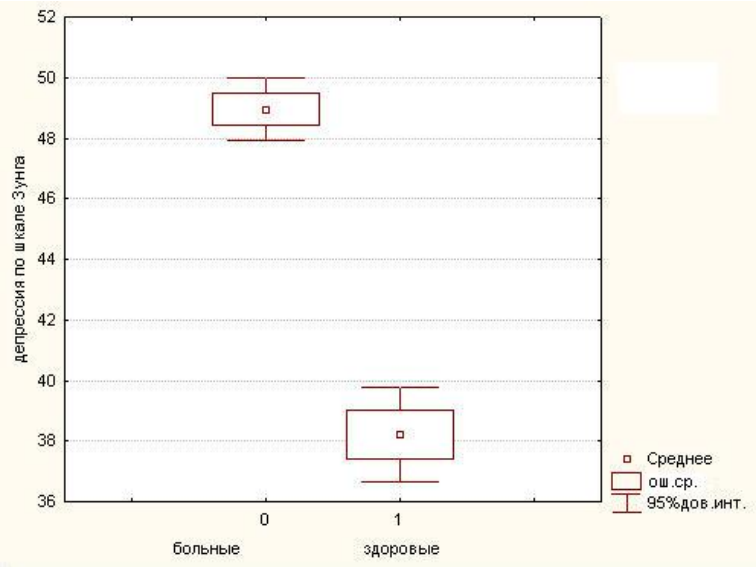

Рис. 9: Уровни депрессии в разных по состоянию здоровья группах.

Поскольку при помощи выборочного среднего $\bar{x}$ мы только примерно указываем, где может находится генеральное среднее μ, и при изменении объема выборки этот показатель может измениться, то есть смысл указать границы, за пределы которых с определенной вероятностью не может попасть генеральное среднее μ. Интервал, накрывающий истинное значение параметра распределения с заданной вероятностью P называется $P \cdot 100\%$**-доверительным интервалом.**

Для построения 95%-доверительного интервала для генерального среднего при достаточно большом объеме выборки и известной дисперсии используется следующая формула:

$$\bar{x} - 1.96 \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{S}{\sqrt{n}}. \tag{19}$$

Если дисперсия σ^2 неизвестна, то для нормально распределенной выборки используется выражение:

$$\bar{x} - \eta^{n-1}_{0.975} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + \eta^{n-1}_{0.975} \cdot \frac{S}{\sqrt{n}}, \tag{20}$$

где $\eta^m_{0.975}$ обозначена 0.975-квантиль распределения Стьюдента с m степенями свободы. Понятие квантили подробно рассматривается в разделе 1.12. Верхний индекс означает число степеней свободы. Вычисление квантили описано в разделе 1.13.

На рис. 9, 106 графически представлены средние значения депрессии, ошибки среднего и 95%-доверительные интервалы для среднего. Сразу видно, что более выражено различие между здоровыми и больными людьми, чем различие между представителями одной возрастной и социальной группы, но в разные годы.

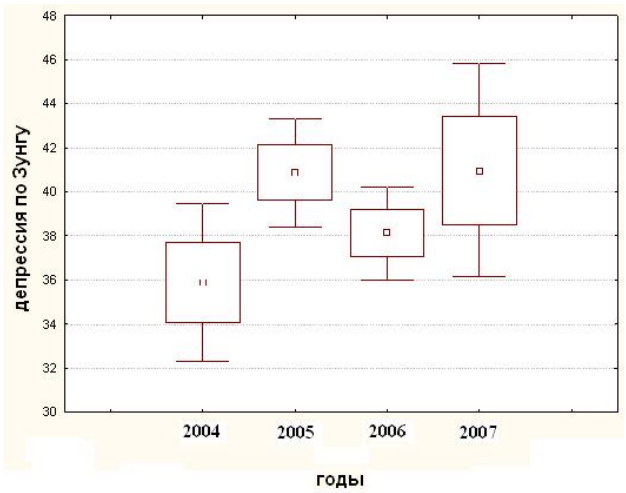

Рис. 10: Уровни депрессии по Зунгу у студентов.

2.4 Проверка гипотез

Принцип маловероятных событий состоит в том, что событие, вероятность которого мала, считается невозможным. Степень "маловероятности"определяется заранее задаваемым **уровнем значимости** α. Обычно рассматривают $\alpha = 0.05$ – вероятность броска Венеры в игре в кости.

Основная задача в построении **критерия** или правила, согласно которому принимается или отвергается гипотеза, заключается в нахождении подходящей **статистики** – функции от выборочных наблюдений, которая при справедливости проверяемой (**нулевой**) гипотезы имеет некоторое известное распределение.

Будем рассматривать случай простой гипотезы о равенстве некоторого параметра распределения $H_0 : \theta = \theta_0$. Альтернативная гипотеза обозначается через H_1. Статистика, на основе которой принимается или отвергается гипотеза, называется статистикой критерия. Для нее строится критическая область, вероятность попадания в которую равна заданному уровню значимости. Например, если статистика критерия имеет стандартное нормальное распределение, то для $\alpha = 0.05$ критическая область имеет вид $(-\infty; -1.96] \cup [1.96; +\infty)$, так как 0.025-квантиль стандартно нормально распределенной случайной величины Z равна $Z_{0.025} = -1.96$, то есть справедливо $P\{Z < -1.96\} = 0.025$, и 0.975-квантиль равна $Z_{0.975} = 1.96$, то есть $P\{Z > 1.96\} = 0.025$, следовательно, $P\{|Z| > 1.96\} = 0.05$. Если наблюдаемое значение статистики Z_* попадает в критическую область, то есть $|Z_*| > 1.96$, то H_0 отвергается с уровнем значимости $\alpha = 0.05$, в противном случае нет оснований отвергнуть гипотезу.

Чаще вместо построения критической области вычисляется P-значение (доверительный уровень вероятности) или значимость α_*, с которой сравнивается α. Для любой статистики Z при альтернативной гипотезе $H_1 : \theta \neq \theta_0$ значимость равна $\alpha_* = P\{|Z| > Z_*\}$, при $H_1 : \theta > \theta_0$ равна $\alpha_* = P\{Z > Z_*\}$, при $H_1 : \theta < \theta_0$ равна $\alpha_* = P\{Z < Z_*\}$. Нулевая гипотеза отвергается случае, когда $\alpha_* < \alpha$.

2.5 Критерии Фишера и Стьюдента

Выборки считаются **однородными**, если они одинаково распределены. Следствием однородности являются одинаковые характеристики среднего и дисперсии. Гистограммы на рисунке 8 показывают, что данные уровней депрессии по шкале Зунга, относящиеся к разным годам наблюдения, распределены неодинаково. То, насколько то различие значимо, можно выяснить при помощи статистических критериев Фишера (равенство дисперсий) и Стьюдента (равенство средних). Оба этих критерия предполагают, что данные распределены по нормальному закону. Гистограммы на рисунке 8 свидетельствуют об унимодальности распределения, мало выраженной асимметрии, поэтому предположение о нормальности выборок допустимо.

По (16) и (17) в группах 2004 и 2005 годов вычисляем выборочные средние $\bar{x} = 36$ и $\bar{y} = 41$, оцениваем дисперсии $S_1^2 = 5.8^2 = 33.6$ и $S_2^2 = 7.8^2 = 60$. Объемы выборок равны $n_1 = 46$ и $n_2 = 37$.

Сначала проверяется гипотеза о равенстве дисперсий $H_0 : \sigma_1^2 - \sigma_2^2 = 0$. **В случае равенства дисперсий статистика Фишера $F = \frac{S_1^2}{S_2^2}$ имеет распределение Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы.** Вычисляем значение статистики F.

$$\frac{S_1^2}{S_2^2} = \frac{33.6}{60.8} = 0.5556 = F_*.$$

При альтернативной гипотезе $H_1 : \sigma_2^2 - \sigma_1^2 \neq 0$ вычисляем P-значение, которое при $F_* > 1$ равно $2 \cdot P\{F > F_*\}$, а при $F_* < 1$ равно $2 \cdot P\{F > 1/F_*\}$,

$$2 \cdot P(F > 1/F_*) = 2 \cdot F\text{РАСП}(1.8; 36; 45) = 0.06.$$

Так как P-значение больше 0.05, то гипотеза о равенстве дисперсий не отвергается. Это означает, что данные 2004 и 2005 года по вариабельности отличаются незначимо.

При условии одинаковости дисперсий для проверки гипотезы о равенстве средних $H_0 : \mu_1 - \mu_2 = 0$ используется статистика Стьюдента вида

$$T = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{где} \quad S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{21}$$

является оценкой **объединенной дисперсии**. При справедливости нулевой гипотезы $H_0 : \mu_1 - \mu_2 = 0$ статистика T имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Для примера, вычисляем по (21) значение статистики Стьюдента, оно равно $T_* = -3.3$. Число степеней свободы равно $46 + 37 - 2 = 81$. При альтернативной гипотезе $H_1 : \mu_1 - \mu_2 \neq 0$ вычисляем значимость

$$\alpha_* = P(|T| > |T_*|) = \text{СТЬЮДРАСП}(|-3.3|; 81; 2) = 0.0014.$$

Так как $\alpha_* < 0.05$, то в доверительной вероятностью 0.95 утверждаем, что уровни депрессии среди студентов университета в разные годы отличаются значимо.

В ситуации, когда гипотеза о равенстве дисперсий отвергается, рекомендуется использовать видоизмененный критерий Стьюдента для разных дисперсий. Вид его более сложный. В *Excel* его можно извлечь в меню "Сервис" - "Анализ данных" - "Двухвыборочный Т-тест с разными дисперсиями".

2.6 Критерий Манна-Уитни

Критерий Манна-Уитни используется для проверки гипотезы однородности независимых выборок либо небольшого объема, либо в случае выборочного распределения, не согласованного с нормальным. Независимые от выборочного распределения критерии называются **непараметрическими**. Рассмотрим данные из таблицы 1 о проценте лиц азиатского происхождения отдельно по каждому штату. Через $x_1, \ldots, x_{n_1}$, где $n_1 = 12$, и $y_1, \ldots, y_{n_2}$, где $n_2 = 8$, обозначим значения признака в 12-ти городах штата Калифорния и восьми городах штата Техас.

город	штат	% лиц аз.происх.	температура	ранги	
		X	Y	X	Y
LOS ANGE	CA	9.8	469.3	12	8
SAN DIEG	CA	11.8	324.0	13	4
SAN JOSE	CA	19.5	171.3	18	3
SAN FRAN	CA	29.1	46.7	20	1
LONG BEA	CA	13.6	50.0	15	7
SACRAMEN	CA	15.0	96.3	17	9
FRESNO	CA	12.5	99.1	14	12
OAKLAND	CA	14.8	56.1	16	2
SANTA AN	CA	9.7	27.1	11	5
ANAHEIM	CA	9.4	44.3	10	5
RIVERSID	CA	5.2	77.7	9	11
STOCKTON	CA	22.8	52.6	19	10
HOUSTON	TX	4.1	539.9	7	14
DALLAS	TX	2.2	342.4	5	20
SAN ANTO	TX	1.1	333.0	2	17
EL PASO	TX	1.2	245.4	3	13
AUSTIN	TX	3.0	217.8	6	16
FORT WOR	TX	2.0	281.1	4	18
ARLINGTO	TX	3.9	93.0	8	18
CORPUS C	TX	0.9	135.0	1	15

Таблица 1: Данные о городах штатов Калифорния и Техас.

2.6.1 Первичная статистика

Первичная статистика включает в себя вычисление среднего, дисперсии, доверительных интервалов и построение гистограмм. Для получения самого общего представления о том, в каком из штатов этот процент выше, по (16) оценим средние значения:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{9.8 + 11.8 + \ldots + 22.8}{12} = 14.4,$$

$$\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i = \frac{4.1 + 2.2 + \ldots + 0.9}{8} = 2.3.$$

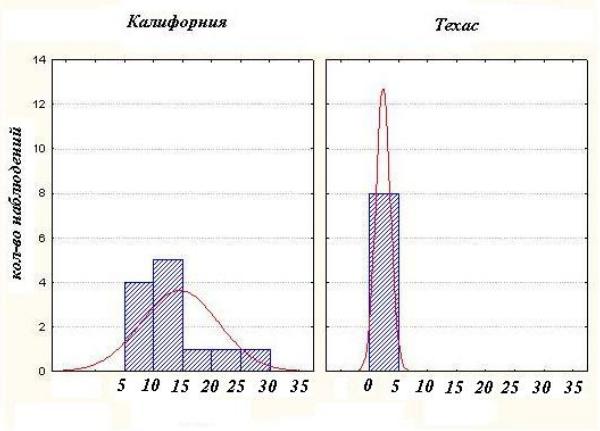

Рис. 11: Категориальные гистограммы признака – процент лиц азиатского происхождения в штатах Калифорния и Техас с длиной интервала группировки 5%.

Оценки дисперсии лиц азиатского происхождения в штатах Калифорния и Техас вычисляем по формуле (17):

$$S_x^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2 = \frac{1}{n_1-1}\left(\sum_{i=1}^{n_1}x_i^2 - n_1\bar{x}^2\right) = \\ = \frac{9.8^2+11.8^2+\ldots+22.8^2-12\cdot 14.4^2}{12-1} = 43.4,$$

$$S_y^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i-\bar{y})^2 = \frac{1}{n_2-1}\left(\sum_{i=1}^{n_2}y_i^2 - n_1\bar{y}^2\right) = \\ = \frac{4.1^2+2.2^2+\ldots+0.9^2-8\cdot 2.3^2}{8-1} = 1.6.$$

В Калифорнии больше средний процент лиц азиатского происхождения и его дисперсия. Построим **гистограмму** – график, показывающий как распределены города по рассматриваемому признаку (рис.11). На горизонтальной оси откладываются интервалы с шагом 5% – значения процентов лиц азиатского происхождения: $0\%, 5\%, 10\%$ так далее. По вертикальной оси откладывается количество наблюдений (городов), в которых процент лиц азиатского происхождения находится в пределах указанных интервалов. Например, в штате Калифорния 4 города имеют процент лиц азиатского происхождения в пределах от 5 до 10%, 5 городов – от 10 до 15%, один город – от 15 до 20%, еще один – от 20 до 25% и последний от 25 до 30%.

В штате Техас все 8 городов имеют процент лиц азиатского происхождения не более 5%. При уменьшении длины интервалов до 0.5% получаем следующее распределение городов (рис.12):

%лиц аз.пр.	$\leq 1\%$	$(1;1.5]$	$(1.5;2]$	$(2;2.5]$	$(2.5;3]$	$(3;3.5]$	$(3.5;4]$	$(4;4.5]$
кол-во гордов	1	2	1	1	1	0	1	1

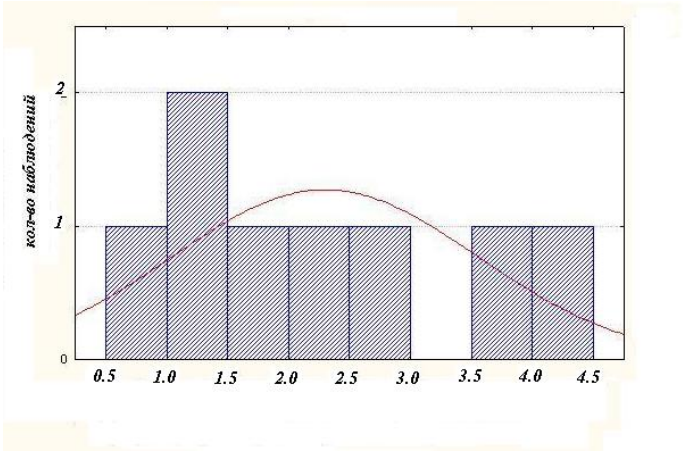

Рис. 12: Гистограмма признака – процент лиц азиатского происхождения в штате Техас с длиной интервала группировки 0.5%.

Доверительные интервалы представлены на рисунке 13.

Итак, имеются две настолько малые выборки, что установить их согласие с нормальным законом распределения не представляется возможным. Критерии однородности Фишера и Стьюдента в данном случае неуместны. Проверяемая гипотеза в **критерии Манна-Уитни** заключается в том, что выборки наблюдений $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$ объемов n_1 и n_2, например, процента лиц азиатского происхождения города штатов Калифорния и Техас, распределены одинаково, а наблюдаемое различие объясняется случайным характером наблюдений. В основе критерия Манна-Уитни лежит **критерий Вилкоксона**.

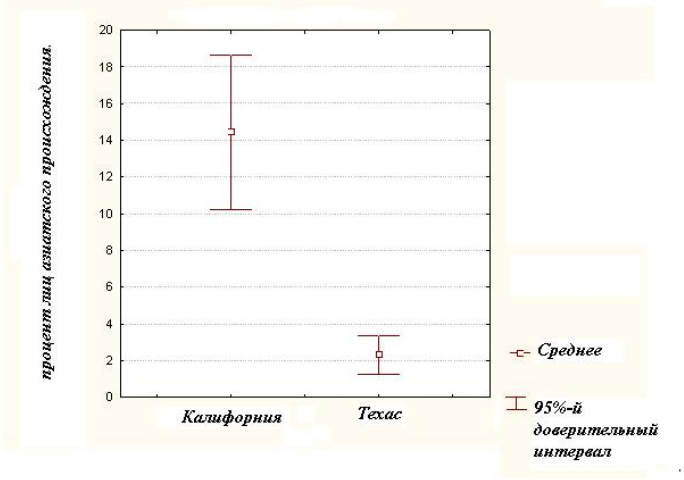

Рис. 13: Средние проценты лиц азиатского происхождения и 95%-е доверительные интервалы в городах штатов Калифорния и Техас.

2.6.2 Критерий Вилкоксона

Если какой-либо элемент из выборки X встречается после какого-то элемента из выборки Y, то говорят, что имеет место одна **инверсия**. Например, получена возрастающая последовательность, в которой элементы выборок X и Y чередуются в следующем порядке: $X\,Y\,Y\,X\,Y\,X\,Y\,Y\,X\,X$. Первый элемент из X меньше всех элементов выборки Y, поэтому его вклад в общее число инверсий U равен нулю. Для второго по порядку элемента из X имеются два меньших по значению элемента из Y, следовательно, его вклад в общее число инверсий U равен 2, и так далее, получаем $U = 0+2+3+5+5 = 15$. Очевидно, что при максимальном числе инверсий $U = 25$, соответствующем последовательности $Y\,Y\,Y\,Y\,Y\,X\,X\,X\,X\,X$, выборки не однородны, и, в целом, значения признака в выборке X больше значений признака Y. Аналогично, минимальное количество инверсий $U = 0$ в случае последовательности $X\,X\,X\,X\,X\,Y\,Y\,Y\,Y\,Y$ свидетельствует о том, что выборки не однородны, и значения признака в выборке X меньше значений Y. Критерий Вилкоксона позволяет установить степень случайности в последовательности $X\,Y\,Y\,X\,Y\,X\,Y\,Y\,X\,X$.

В критерии Вилкоксона для проверки гипотезы $H_0 : p_{xy} = p_{yx}$ о равенстве вероятностей $p_{xy} = P\{X < Y\}$ и $p_{yx} = P\{X > Y\}$ при альтернативной $H_1 : p_{xy} > p_{yx}$ выбирается критическое число инверсий U_α такое, что если наблюдаемое число инверсий больше критического $U > U_\alpha$, то H_0 отвергается в пользу H_1 с уровнем значимости α. Однако алгоритм вычисления критического числа инверсий U_α является непростой задачей, поэтому требуются либо специальные таблицы, либо компьютерные программы. На практике обычно используют более простое его обобщение в виде критерия Манна-Уитни.

2.6.3 Выражение числа инверсий через ранги

Для вычисления число инверсий U через ранги выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$ объединяются в одну и каждому наблюдению ставится в соответствие ранг – его место в вариационном (упорядоченном) ряду.

$$U = (r_1 - 1) + (r_2 - 2) + \ldots + (r_{n_1} - n_1) = \sum r_i - (1 + 2 + \ldots + n) =$$
$$= \sum r_i - \frac{n_1(n_1+1)}{2}, \quad \text{так как} \quad 1 + 2 + \ldots + n = \frac{n(n+1)}{2}.$$

Смесь выборок	66	67	68	70	75	78	79	80	81	92
ранги	1	2	3	4	5	6	7	8	9	10
выборки	y	y	x	x	x	y	y	y	x	x

Число инверсий равно $U = 2 + 2 + 2 + 5 + 5 = 16$. Выразим это число через ранги. Элементы выборки X имеют ранги:

$$r_1 = 3, r_2 = 4, r_3 = 5, r_4 = 9, r_5 = 10,$$
$$U = (3 - 1) + (4 - 2) + (5 - 3) + (9 - 4) + (10 - 5) =$$
$$= 3 + 4 + 5 + 9 + 10 - (1 + 2 + 3 + 4 + 5) = 16.$$

Для проверки вычислений использую тождество

$$R_1 + R_2 = \frac{n(n+1)}{2}, \quad \text{где} \quad n = n_1 + n_2, \tag{22}$$

так как всего имеется n рангов от 1 до n, а сумма элементов натурального ряда равна $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$.

Если s значений в вариационном ряду совпадают, то их ранги $r_1, \ldots, r_s$ заменяются на одинаковые значения, равные среднему арифметическому $\bar{r} = \frac{1}{s}\sum_{k=1}^{s} r_k$.

При справедливости нулевой гипотезы $H_0 : p_{xy} = p_{yx}$ и при $n_1, n_2 \geq 8$ случайное число инверсий U имеет нормальное распределение со средним числом инверсий $\mathbf{E}U = \frac{n_1 n_2}{2}$. Дисперсия U равна

$$\mathbf{D}U = \frac{n_1 n_2}{12}(n_1 + n_2 + 1).$$

Для того чтобы использовать квантили стандартного нормального распределения качестве критических значений, преобразуем статистику U в статистику Манна-Уитни

$$Z = \frac{U - \mathbf{E}U}{\sqrt{\mathbf{D}U}} = \frac{R_2 - \frac{n_2(n_2+1)}{2} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}, \tag{23}$$

которая при справедливости нулевой гипотезы однородности выборок имеет стандартное нормальное распределение. Через R_2 обозначена сумма рангов выборки объема $n_2 < n_1$. При наличии компьютера в *Excel* вычисляется доверительный уровень вероятности p как

$$p = P\{|Z| > |Z_*|\} = 2(1 - \text{НОРМСТРАСП}(abs(Z_*))), \tag{24}$$

при $p < \alpha = 0.05$ нулевая гипотеза отвергается. При отсутствии компьютера наблюдаемое значение статистики Z_* сравнивается с критическим $Z = 1.96$. Если $|Z_*| > 1.96$, то гипотеза однородности отвергается с уровнем значимости $\alpha = 0.05$.

2.6.4 Пример использования критерия Манна-Уитни

Объединяем выборки процентов лиц азиатского происхождения двух штатов в одну и ставим ранги (таблица 2). В Сан-Франциско самый высокий процент лиц азиатского происхождения, поэтому ему соответствует наибольший ранг 20, а в Корпус-Кристи самый низкий процент 0.9%, ему соответствует самый низкий ранг 1. Для компактности в таблице 2 некоторые значения округлены.

Найдем сумму из $n_1 = 12$ рангов, относящихся к штату Калифорния: $R_1 = 174$. Сумма из $n_2 = 8$ рангов, относящихся к штату Техас равна $R_2 = 36$. Для проверки правильности вычислений используем свойство (22). Проверяем:

$$R_1 + R_2 = 174 + 36 = 210 = \frac{20 \cdot 21}{2}.$$

Число инверсий U получаем из суммы рангов

$$U = R_2 - \frac{n_2(n_2+1)}{2} = 36 - \frac{8 \cdot 9}{2} = 0,$$

дисперсия $DU = \frac{n_1 n_2 (n_1+n_2+1)}{12} = 168$,

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{DU}} = \frac{0 - 8 \cdot 12/2}{\sqrt{168}} = -3.7$$

Наблюдаемое значение статистики $Z_* = -3.7$ больше по абсолютной величине 0.975-квантили стандартного нормального распределения 1.96, поэтому гипотеза однородности выборок отвергается с уровнем значимости $\alpha = 0.05$ и можно утверждать, что процент лиц азиатского происхождения значимо меньше в штате Техас, чем в штате Калифорния.

	Калифорния (X)												Техас (Y)							
%	10	12	20	30	14	15	13	15	9.7	9.4	5	23	4	2.2	1.1	1.2	3	2	3.9	1
ранг	12	13	18	20	15	17	14	16	11	10	9	19	7	5	2	3	6	4	8	1

Таблица 2: Признак – процент лиц азиатского происхождения – и его ранги.

Если для вычисления доверительного уровня вероятности p воспользоваться таблицей 7 значений функции стандартного нормального распределения $\Phi(x)$ и выражением $\Phi(-x) = 1-\Phi(x)$, то для максимально представленного в этой таблице значения $\Phi(3.4) = 0.9998$ имеем $\Phi(-3.4) = 1 - 0.9998 = 0.0002$, следовательно доверительный уровень вероятности $p < 0.0002$ меньше $\alpha = 0.05$ и гипотеза однородности штатов по проценту лиц азиатского происхождения отвергается, то есть в Техасе этот процент значимо меньше.

2.7 Критерий Краскела-Уоллиса.

Этот непараметрический критерий предназначен для проверки гипотезы однородности сразу для нескольких независимых выборок.

Для примера рассмотрим данные в таблице 3 о количестве ошибок, допущенными студентами, обучаемых по одной из трех методик. Нужно выяснить, являются ли значимыми различия в методиках.

Для ответа на вопрос объединяем все три группы в одну и упорядочиваем данные. Учитывая одинаковые значения признака числа ошибок, вычисляем ранги. Затем суммируем ранги R_i, $i = 1, \ldots, r$, где через r обозначено количество групп, $r = 3$. $n_1, \ldots, n_r$ – это количество наблюдений в каждой группе, $n_1 = 7$, $n_2 = 5$, $n_3 = 3$, $n_1 + \ldots + n_r = n$. Получаем $R_1 = 37.5, R_2 = 43, R_3 = 39.5$. Если между группами нет систематических различий, то средние ранги $\frac{1}{n_i} R_i$ внутри каждой группы не должны очень сильно отличаться от среднего ранга $\frac{n+1}{2}$, рассчитанного по всей совокупности.

В качестве меры отступления от чистой случайности рассматривают статистику Краскела-Уоллиса

$$H = \sum_{i=1}^{r} c_i \left(\frac{1}{n_i} R_i - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{R_i^2}{n_i} - 3(n+1),$$

полученную при подходящем выборе весов c_i, так что H асимптотически сходится к χ^2 распределению с $r-1$ степенью свободы. $\frac{12}{15 \cdot 16} \left(\frac{37.5^2}{7} + \frac{43^2}{5} + \frac{39.5^2}{3} \right) - 3 \cdot 16 = 6.84$.

№	кол-во ошибок	методика	ранг
1	0	0	1
2	1	0	3.5
3	1	0	3.5
4	1	0	3.5
8	1	1	3.5
5	2	0	7.5
6	2	0	7.5
9	2	1	7.5
10	2	1	7.5
7	3	0	11
11	3	1	11
13	3	2	11
12	4	1	13.5
14	4	2	13.5
15	5	2	15

Таблица 3: Данные о количестве ошибок, допущенных студентами, обучавшимися по разным методикам.

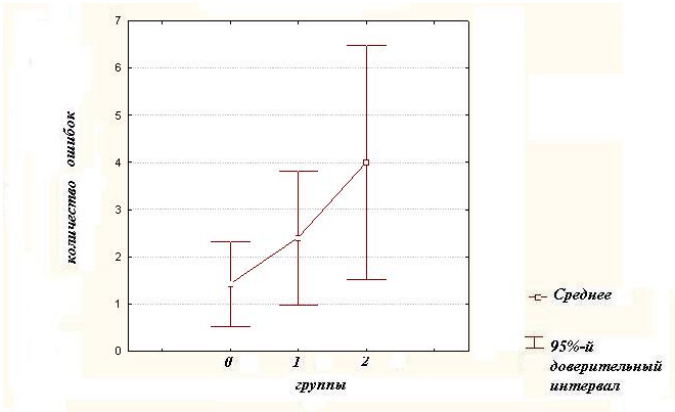

Рис. 14: Зависимость количества ошибок от методики преподавания.

В случаях $r = 2, 3, 4$ используются соответственно критические значения 3.84, 5.99, 7.81. В нашем случае $r = 3$, имеем 6.84>5.99, следовательно, гипотеза об отсутствии систематических различий между группами отвергается (рис.14).

2.8 Проверка гипотез однородности в случае зависимых выборок (критерий знаков)

Рассмотрим пример статистических данных (табл.4). В качестве признаков, характеризующих явление преступности, рассматривается количество убийств, совершенных определенным образом. Это данные типа процессов, так как в качестве повторностей выступают годы наблюдений. Они могут оказаться зависимыми, т.е. если количество убийств при помощи ружья и пистолета в 1826 году было больше, чем количество убийств при помощи задушения, то скорее всего эта тенденция сохранится и в 1827 году. В инструкциях по статистическому анализу такие данные называются **зависимыми выборками**.

Формы убийства	1826	1827	1828	1829	1830	1831
Ружье, пистолет	56	64	60	61	57	88
Сабля, шпага кинжал, стилет	15	7	8	7	12	30
Нож	39	40	34	46	44	34
Палка, трость	23	28	31	24	12	21
Камень	20	20	21	21	11	9
Тупые, острые и колющие орудия	35	40	42	45	46	49
Задушение	2	5	2	2	2	4
Сбрасывание, утопление	6	16	6	1	4	3
Удар ноги, кулака	28	12	21	23	17	26
Неизвестное	17	1	2	0	2	2
Всего	241	233	227	230	207	266

Таблица 4: Данные уголовной полиции Франции (Кетле)

Задача 1. *Обратим внимание на то, что в период 1826–1830гг. наблюдалось в целом общее снижение преступности, за исключением небольшого повышения в 1828–1829гг, а в 1831 ее значительный рост (рис.15). Выясним, нельзя ли считать, что, в основном, преступность снижалась, а ее редкое повышение объясняется случайностью.*

Воспользуемся для этого критерий знаков. В качестве испытаний будем рассматривать $n = 5$ пар лет: 1826–1827, 1827–1828, 1828–1829, 1829–1830, 1830–1831. "Успехом" условимся считать события, которые появляются реже, в данном случае это повышение числа преступлений. Таким образом получаем два "успеха" $k = 2$ в 1828–1829гг. и в 1830–1831гг. и три неудачи. Если бы во всех пяти случаях наблюдался рост числа убийств или, наоборот, все пять наблюдений соответствовали снижению

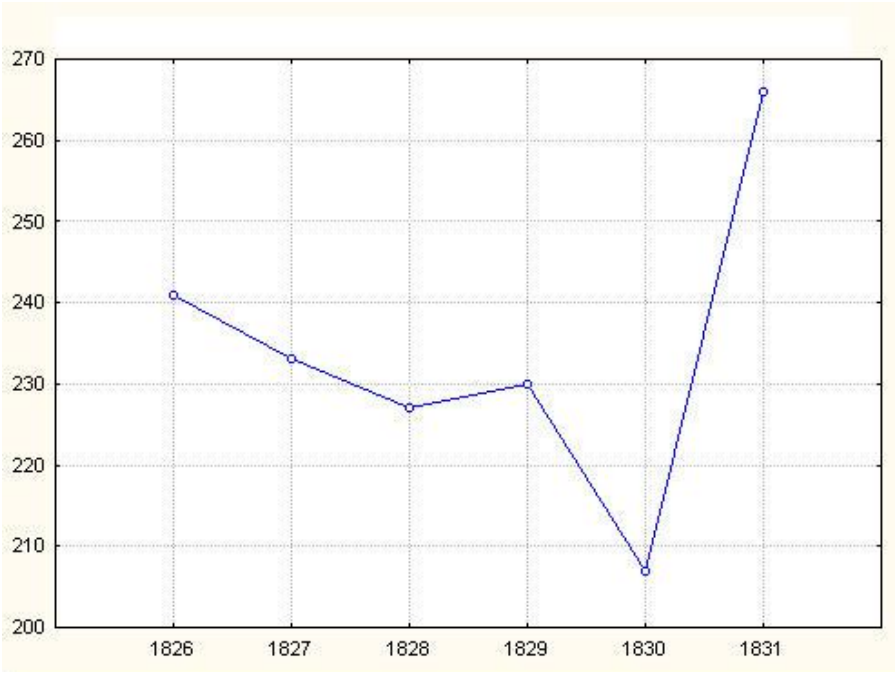

Рис. 15: Динамика числа уголовных преступлений во Франции в 1826-1831гг.

числа преступлений, то все было бы понятно. А вот насколько значимо сказывается на общем снижении эти повышения, следует обратиться к критерию знаков.

Критерий знаков используется для проверки гипотезы однородности двух зависимых выборок $(x_1, y_1), \ldots, (x_n, y_n)$. Например, $x_1 = 64$, $y_1 = 60$ – число убийств при помощи ружья и пистолета в 1827 и 1828 годах, $x_2 = 40, y_2 = 34$ – число убийств при помощи ножа в тех же 1827-1828 годах, и так далее, всего $n = 10$ способов убийства – это число независимых испытаний. Если $x_i = y_i$, то соответствующие наблюдения исключается как мало информативное, и число испытаний уменьшается на единицу.

Рассмотрим знаки разностей $x_i - y_i$, $i = 1, \ldots, n$. При однородных выборках частоты знаков плюс и минус должны быть одинаковы. Пусть знаков плюс больше знаков минус, поэтому условимся считать появление знака плюс "успехом". Проверяемая гипотеза состоит в том, что вероятность "успеха" равна $p = 0.5$. В примере эта гипотеза означает, что изменение количества убийств в сторону уменьшения или увеличения являются случайными.

n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
K	0	0	0	1	1	1	2	2	3	3	3	4	4	5	5	5

Таблица 5: Соотношение между числом испытаний и критическим числом "успехов".

Для каждого числа испытаний на основе принципа маловероятных событий вычисляется критическое число "успехов" (таблица 5). Оно вычисляется как максимальное число успехов K, такое что $P\{X \leq K\} \leq 0.05$, где X – случайное число

успехов. Рассмотрим на примере $n = 10$ способ вычисления критического значения. При вероятности успеха $p = \frac{1}{2}$ справедливы выражения

$$P\{X \le 0\} = \frac{C_{10}^0}{2^{10}} \approx 0.001 < 0.05,$$

$$P\{X \le 1\} = P\{X = 0\} + P\{X = 1\} = \frac{C_{10}^0 + C_{10}^1}{2^{10}} = \frac{1 + 10}{1024} = 0.01 < 0.05,$$

$$P\{X \le 2\} = P\{X = 0\} + P\{X = 1\} + P\{X = 2\} =$$

$$= \frac{C_{10}^0 + C_{10}^1 + C_{10}^2}{2^{10}} = \frac{1 + 10 + 45}{1024} = 0.055 > 0.05,$$

Таким образом, при $n = 10$ максимальное число успехов K, при котором случайное число успехов X не превышает $K = 1$. В случае большего числа испытаний лучше использовать специальные компьютерные программы.

Критерий знаков состоит в следующем: если наблюдаемое число "успехов" меньше критического K, то проверяемую гипотезу $H_0 : p = 0.5$ следует отвергнуть, иначе нет оснований ее отвергнуть.

Задача 1 (продолжение). При исследовании динамики общего числа преступлений дважды наблюдался рост числа преступлений, из $n = 5$ число "успехов" равно 2, оно больше критического $K = 0$, следовательно, нет оснований для того, чтобы отвергнуть гипотезу о случайном характере динамики числа убийств.

Задача 2. *По формам убийства нужно проверить гипотезу о том, что нет ни прогрессирования количества убийств с 1826 по 1827 гг., ни их уменьшения.*

С 1826 по 1827 гг. уменьшилось число убийств по трем формам (сабля, шпага, стилет; удар ноги, кулака; неизвестное), увеличилось число убийств по шести формам, число убийств при помощи камня осталось на том же уровне. Таким образом, из 9 испытаний имеем три успеха. Согласно таблице 5, это число не меньше критического, соответствующего девяти испытаниям, следовательно нет оснований отвергнуть гипотезу о случайности повышения или понижения числа убийств в эти годы.

Задача 3. *Предположим, что из десяти пациентов восьмерым лекарство помогло, а двоим нет. Можно ли говорить об эффективности лекарства?*

В таблице 5 для $n = 10$ критическое число равно 1, оно меньше 2, следовательно, нет оснований отвергнуть гипотезу случайном эффекте препарата.

Задача 4. *Пусть в группе из 18 человек 4 мальчика. Насколько значимо преобладание девочек?*

По таблице 5 критическое значение для $n = 18$ равно 5, оно больше наблюдаемого значения 4, следовательно, гипотеза о равновероятном количестве мальчиков и девочек отвергается, и мальчиков значимо меньше девочек.

2.9 Проверка независимости категориальных признаков на основе критерия хи-квадрат

2.9.1 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 $y_1, \ldots, y_s$. Например, 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"с предложением "У меня бывают периоды плача и близости к слезам", из них 6 девушек и 8 юношей, "скорее не согласны"пятеро (4 девушки и 1 юноша), семь человек "скорее согласны"(1 юноша и 6 девушек) и 10 "совершенно согласны"(один юноша и 9 девушек).

Представим эти данные в виде **таблицы сопряженности**, где строки определяются градациями признака X, столбцы – градациями признака Y, а элементы таблицы равны количеству n_{ij} наблюдений с соответствующими градациями: $X = x_i$, $Y = y_j$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$. Через n_{i*} и n_{*j} обозначены соответствующие суммы: $n_{i*} = \sum\limits_{j=1}^{s} n_{ij}$, $n_{*j} = \sum\limits_{i=1}^{r} n_{ij}$. $\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s} n_{ij} = n$.

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

$$p_{ij} = p_{i*}p_{*j}, \tag{25}$$

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

При выборочном наблюдении над признаками, когда вероятности оцениваются как отношения частот, это равенство может быть нарушено из-за того, что количество наблюдений ограничено. При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{26}$$

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

Для того чтобы определить, не является ли это нарушение случайным, вычисляют статистику – функцию от выборочных наблюдени – в виде взвешенной суммы квадратов отклонений наблюдаемых частот $\widehat{p}_{ij} = \frac{n_{ij}}{n}$ от того, что бы получилось, если бы признаки были независимы, то есть произведения $\widehat{p}_{i*} \cdot \widehat{p}_{*j}$:

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

Коэффициенты c_{ij} выбираются таким образом, чтобы сумма квадратов получилась с чрезвычайно простыми свойствами. Оказывается, если взять в качестве коэффициентов $c_{ij} = \frac{n^3}{n_{i*}n_{*j}}$, тогда статистика χ^2 имеет вид:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right). \tag{27}$$

Для признаков X и Y получаем значение статистики $\chi^2 = 7.79$. Возникает вопрос: насколько велико расхождение между наблюдаемыми частотами и частотами, которые получились бы в случае независимых признаков.

В математической статистике на основе **теоремы Пирсона** показано, что при условии независимости признаков X и Y статистика χ^2 из (27) имеет распределение хи-квадрат с числом степеней свободы, равным $n = (r-1)(s-1)$. Эта статистика используется в качестве статистики критерия для проверки гипотезы о независимости качественных признаков X и Y.

Примем уровень значимости $\alpha = 0.06$. При помощи функции ХИ2ОБР найдем значения (рис.16), определяющие критическую область $[x_0; +\infty)$:

$$x_0 = \text{ХИ2ОБР}(\alpha; n) = \text{ХИ2ОБР}(0.06; 3) = 7.41.$$

При числе степеней свободы, равном 3, $x_0 = 7.41$. Наблюдаемое значение статистики $\chi^2_* = 7.79 > 7.41$ попадает в критическую область $[7.41; +\infty)$, поэтому гипотезу о независимости признаков следует отвергнуть с уровнем значимости $\alpha = 0.06$. Если взять $\alpha = 0.05$, то получаем $x_0 = 7.81$, и наблюдаемое значение статистики $\chi^2_* = 7.79$ не попадает в критическую область $[7.81; +\infty)$, хотя и очень близко около нее находится. По установленному правилу мы не можем отвергнуть гипотезу с уровнем значимости $\alpha = 0.05$. То есть при уровне значимости $\alpha = 0.06$ мы гипотезу отвергаем,

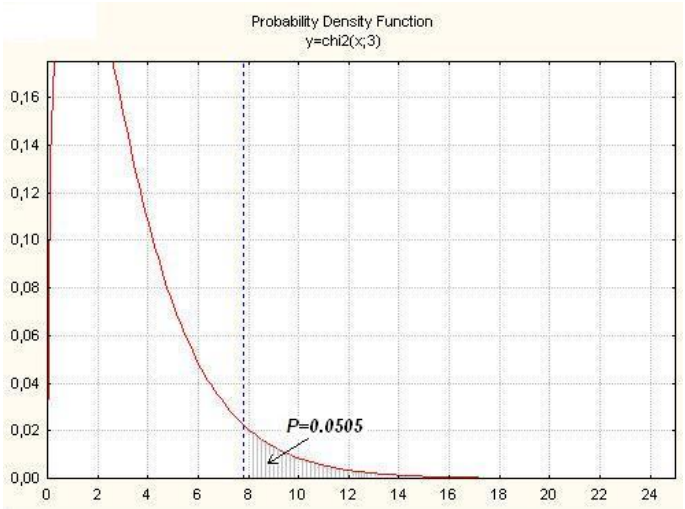

Рис. 16: Плотность распределения хи-квадрат с 3 степенями свободы.

при $\alpha = 0.05$ уже нет. Чтобы при некотором наблюдаемом значении статистики χ^2_* сразу видеть все уровни значимости, при которых гипотеза отвергается или не отвергается, нужно вычислить P-значение или

$$P = P(\chi^2 > \chi^2_*) = \text{ХИ2РАСП}(\chi^2_*; n) = P(\chi^2 > 7.79) = \text{ХИ2РАСП}(7.79; 3) = 0.0505. (28)$$

В случае, когда значение статистики χ^2_* попадает в критическую область, P-значение меньше заранее выбранного уровня значимости. Поэтому *для любого уровня значимости α, большего P-значения, нулевая гипотеза отвергается.*

Вывод: для любого уровня значимости $\alpha > 0.0505$ гипотеза о независимости признаков отвергается. Как мы и получили, при $\alpha = 0.05$ гипотеза не отвергается, а при $\alpha = 0.06$ отвергается. Можно сказать, что различие в ответах по полу оказывается значимым и не может быть объяснено случайностью наблюдений.

2.10 Регрессионный анализ

2.10.1 Проверка значимости коэффициента корреляции

Пусть имеется n наблюдений парной выборки $(x_1, y_1), \ldots, (x_n, y_n)$ двух метрических признаков X и Y.

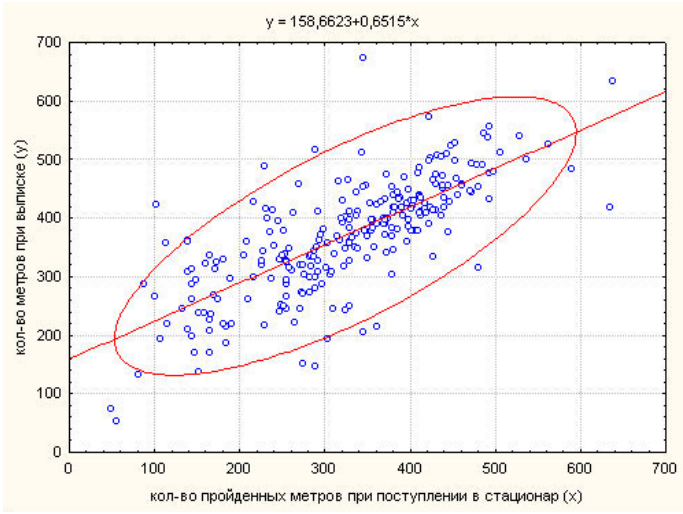

Рис. 17: Линия регрессии, определяющая зависимость между тестами нагрузочной пробы (с шестиминутной ходьбой) при поступлении в стационар и при выписке.

На рис. 17 представлена двумерная диаграмма признаков X – результат теста шестиминутной ходьбы больных с сердечной недостаточностью при поступлении в стационар, Y – результат того же теста при выписке. Как видно из диаграммы, прослеживается тенденция: чем больше больной прошел метров при поступлении, тем больше он прошел их при выписке.

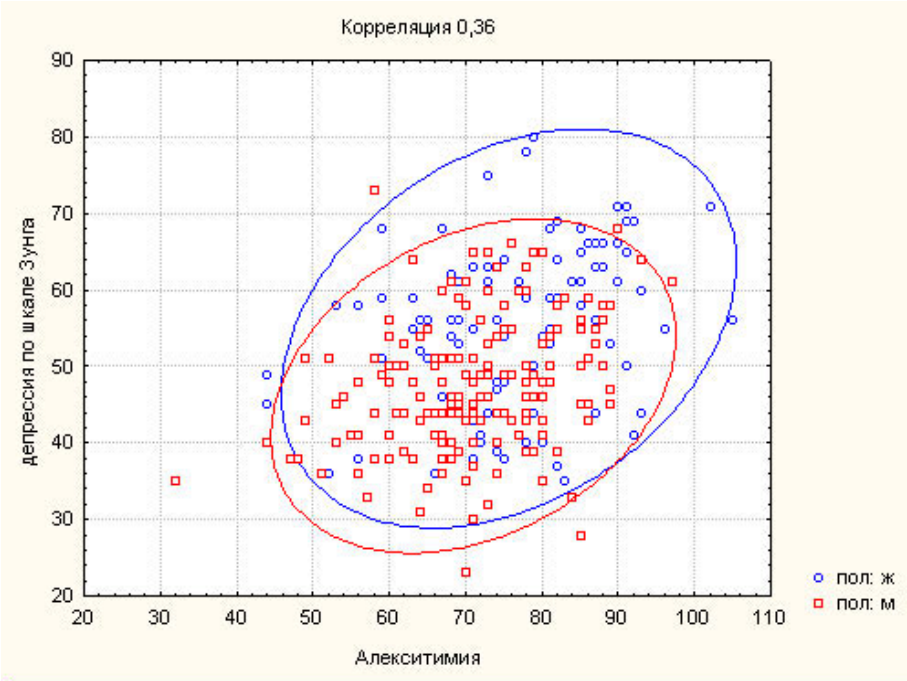

Рис. 18: Исследование зависимости между депрессией и алекситимией.

На рис.18 представлена двумерная диаграмма двух других признаков: X – торонтская алекситимическая шкала и Y – уровень депрессии по шкале Зунга в той же группы больных с сердечной недостаточностью. Разные эллипсы соответствуют отдельно женщинам и мужчинам. Визуально оцениваем, что степень зависимости между этими признаками меньше, чем в случае тестов шестиминутной ходьбы при поступлении и при выписке.

Численно уровень линейной зависимости между признаками оценивается при помощи *коэффициента корреляции*, который можно вычислить по формуле:

$$R = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum\limits_{i=1}^{n} y_i^2 - n\bar{y}^2}} \tag{29}$$

или при помощи встроенной функции КОРРЕЛ в электронных таблицах *Excel*. В качестве аргументов этой функции нужно указать массивы ячеек, в которых находятся значения признаков X и Y. Например, значения признака X находятся в ячейках от $A2$ до $A101$, признака Y в ячейках от $B2$ до $B101$. Функция, вычисляющая коэффициент корреляции, имеет вид: КОРРЕЛ($A2 : A101; B2 : B101$). Заметим, что количество ячеек в этих массивах должно совпадать, иначе функция выдает ошибку.

Если между признаками существует линейная зависимость, т.е. один признак линейным образом выражается через другой: $Y = kX + b$, тогда коэффициент корреляции равен $R = 1$ при положительном k и $R = -1$ при отрицательном k (рис.19).

Если признаки являются независимыми, то коэффициент корреляции равен $R = 0$.

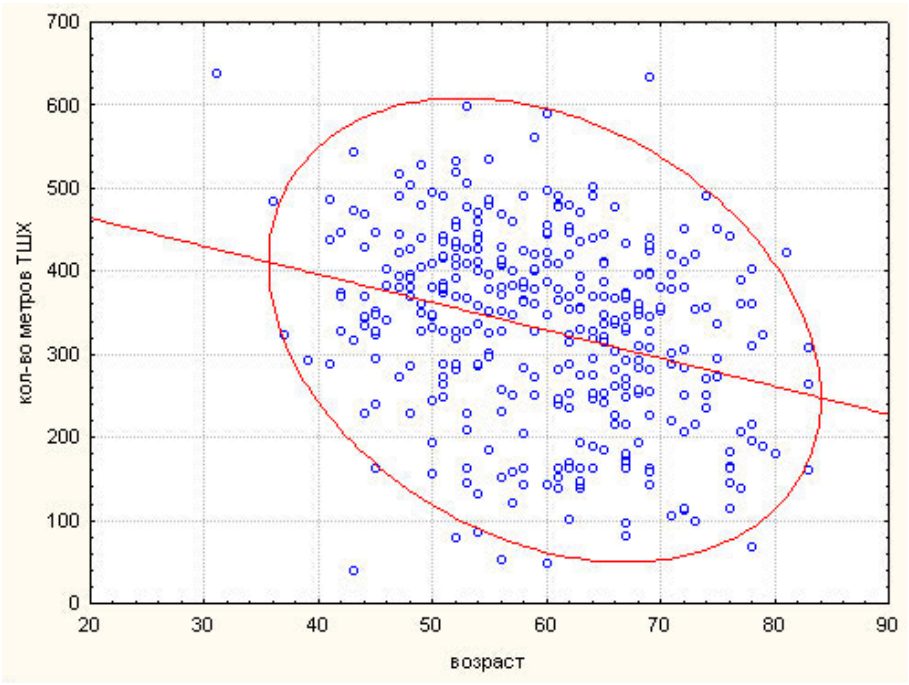

Рис. 19: Отрицательная зависимость между возрастом и результатом теста шестиминутной ходьбы $R = -0.29,\ p < 0.01$.

В примере исследования зависимости между тестами шестиминутной ходьбы при поступлении в стационар и при выписке коэффициент корреляции равен 0.76, в примере исследования взаимосвязи между алекситимической шкалой и депрессией по Зунгу коэффициент корреляции равен $R = 0.36$.

Насколько значимо отклонение коэффициентов корреляции от нуля можно выяснить при помощи критерия значимости корреляции Стьюдента. При независимости признаков X и Y статистика критерия вида:

$$T = T(R, n) = \frac{R}{\sqrt{1 - R^2}}\sqrt{n - 2} \tag{30}$$

имеет распределение Стьюдента с $n - 2$ степенями свободы.

$$T(0.74; 255) = 17.5, \quad P = \text{СТЬЮДРАСП}(17.5; 253; 2) = 2 \cdot 10^{-45},$$
$$T(0.36; 255) = 6.16, \quad P = \text{СТЬЮДРАСП}(17.5; 253; 2) = 3 \cdot 10^{-9}.$$

В обоих случаях P-значение меньше 10^{-8}, отсюда с вероятностью, близкой к 1, можно говорить о том, что между указанными парами признаков имеется какая-то взаимосвязь.

2.10.2 Частные коэффициенты регрессии

Статистический анализ позволяет исследовать влияние на некоторый признак (его называют "зависимой" переменной) ряда других признаков (их называют "независимые" переменные). Если это влияние существенно, то появляется возможность спрогнозировать поведение "зависимой" переменной при известных значениях "независимых"переменных.

Математическая постановка задачи заключается в следующем: пусть имеются s "независимых" переменных $X_1, \ldots, X_s$ и "зависимая"переменная $Y = X_0$. Матрица наблюдений имеет вид $\{x_{ij}\}$, $i = 0, \ldots, s$, $j = 1, \ldots, n$, где x_{ij} – значение i-го признака j-го индивида, n – количество индивидов.

ПРИМЕР 1. В качестве "зависимой" переменной Y будем рассматривать среднюю продолжительность жизни, а в качестве независимых переменных – национальное богатство X_1 и объем X_2 продаваемой водки за год, приходящийся на одного человека (в бутылках), таким образом $s = 2$. В качестве индивидов $n = 7$ будем рассматривать годы: 1970, 1975, 1980, 1985, 1990, 1995 и 1998.

j	Годы	Продолж-ть жизни Y	Нац.богатсво (млрд.дол.) X_1	Водка (бут.в год) X_2
1	1970	68.9	2372*	25.3
2	1975	68.1	2372*	28
3	1980	67.6	2489	30
4	1985	69.2	3379	23.5
5	1990	69.2	4130	18
6	1995	64.6	1171	38.4
7	1998	67	689	29.6

Например, $x_{15} = 4130$ – это объем национального богатства в 1990 году. Звездочками отмечены артефакты или неполные наблюдения. Поскольку данные о национальном богатстве в 1970 и 1975 годах отсутствуют, вместо прочерков указаны средние значения национального богатства.

Регрессией называют функцию, являющуюся условным математическим ожиданием

$$E(Y|X_1, \ldots, X_s)$$

переменной Y в зависимости от значений переменных $X_1, \ldots, X_s$. Модель множественной регрессии представляет собой уравнение

$$Y = E(Y|X_1, \ldots, X_s) + \varepsilon,$$

где ε – случайная не систематическая ошибка с нулевым математическим ожиданием и дисперсией σ^2. Для практических целей удобной оказывается линейная аппроксимация множественной регрессии в виде

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

Коэффициенты $\beta_0, \beta_1, \ldots, \beta_s$ называются **частными коэффициентами регрессии.** Эти коэффициенты вычисляются при помощи программы регрессионного анализа из условия построения наилучшего линейного предсказания

$$\hat{Y}(X_1, \ldots, X_s) = \beta_0 + \beta_1 X_1 + \ldots + \beta_s X_s$$

в смысле минимума расстояния между наблюдениями и прогнозом, то есть частные коэффициенты регрессии ищутся из условия минимальности выражения $\sum_{j=1}^{n}(y_j - \beta_0 - \beta_1 x_{1j} - \ldots - \beta_s x_{sj})^2$. При $s = 1$ уравнение одномерной линейной регрессии имеет вид: $Y = \beta_1 X + \beta_0$, где $X = X_1$. Коэффициенты вычисляются из обычных выборочных характеристик: средних $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ и $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$, а также $\sum_{j=1}^{n} x_j y_j$ и $\sum_{j=1}^{n} x_j^2$:

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}}{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{31}$$

Рис. 20: Влияние национального богатства на продолжительность жизни. Коэффициент корреляции $R = 0.78$, значимость $p = 0.04$.

ПРИМЕР 1 (продолжение). Построим наилучшее предсказание средней продолжительности жизни Y по объему национального богатства X_1. Из (31) получаем $\hat{\beta}_1 = 0.00108, \hat{\beta}_0 = 65.25$. Значимость одномерной линейной регрессии совпадает со значимостью $p = 0.04$

коэффициента корреляции $R = 0.78$, что свидетельствует о наличии линейной зависимости между этими признаками.

Наилучшее предсказание средней продолжительности жизни по двум признакам X_1 (национальное богатство) и X_2 (продажа водки) имеет вид:

$$Y = 75.05 - 0.00009X_1 - 0.2557X_2.$$

Заметим, что по указанному уравнению регрессии влияние национального богатства на среднюю продолжительность жизни отрицательно, хотя когда мы рассматривали отдельную линию регрессии, то влияние было положительным. Нам предстоит объяснить этот парадокс.

Рис. 21: Влияние продажи водки на продолжительность жизни. Коэффициент корреляции $R = -0.93$. Значимость линейной регрессии $p < 10^{-4}$.

2.10.3 Частный коэффициент корреляции

Разность между признаком и его линейным предсказанием $Y - \hat{Y}$ называется *остатком*. Важным свойством остатка является его не коррелированность ни с одной из переменных $X_1, \ldots, X_s$, входящих в предиктор $\hat{Y}(X_1, \ldots, X_s)$. Иллюстрацией этого является двумерная диаграмма остатка и одной из вычитаемых компоненты (рис.22).

Это свойство оказывается чрезвычайно важным. Оно позволяет на определенном этапе устранять влияние той или иной переменной. Рассмотрим наилучшие линейные предсказания $\hat{Y}(X_2, \ldots, X_s)$ и $\hat{X}_1(X_2, \ldots, X_s)$ переменных Y и X_1 по переменным $X_2, \ldots, X_s$. Остатки $Y - \hat{Y}(X_2, \ldots, X_s)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_s)$ не коррелированы ни с одной из переменных $X_2, \ldots, X_s$. Это новые переменные, но по сути дела это те же переменные Y и X_1, только из них устранено влияние переменных $X_2, \ldots, X_s$.

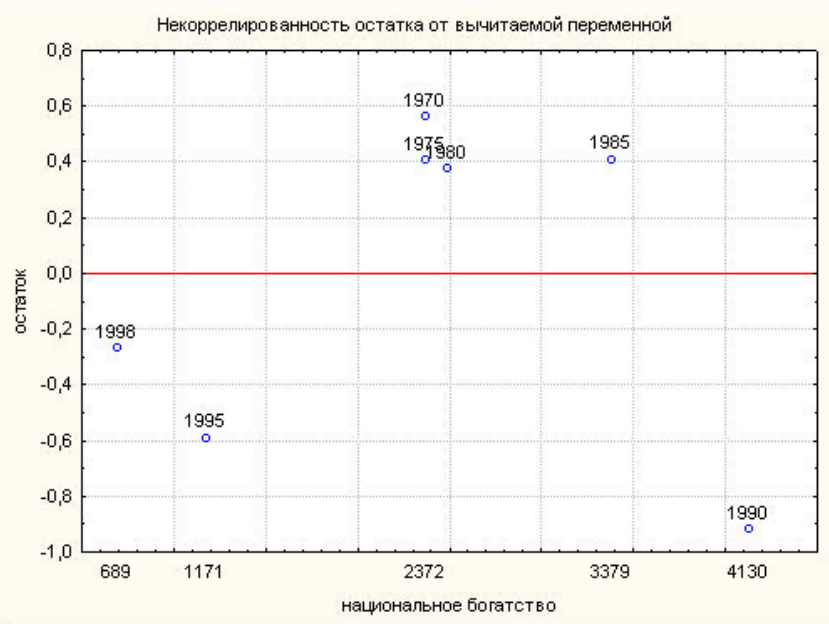

Рис. 22: Некорреллированность остатка $Y - (75.05 - 0.00009X_1 - 0.2557X_2)$ от переменной X_1.

Коэффициент корреляции ρ_{Yx_1} между $Y - \hat{Y}(X_2, \ldots, X_s)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_s)$ называется **частным коэффициентом корреляции**.

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

$$T = \frac{\rho_{Yx_1}}{\sqrt{1 - \rho^2_{Yx_1}}}\sqrt{n - s - 1}, \tag{32}$$

которая в случае равенства нулю генерального коэффициента корреляции имеет распределение Стьюдента с числом степеней свободы, равным $n - s - 1$. Для вычисления p-значения при полученном наблюдаемом значении статистики Стьюдента T_* можно использовать функцию в электронных таблицах *Excel* вида: $P = P(|T| > T_*) = \text{СТЬЮРАСП}(|T_*|; n - s - 1; 2)$.

ПРИМЕР 1 (продолжение). Вычислим частный коэффициент корреляции ρ_{Yx_1} между средней продолжительностью жизни и национальным богатством при условии устранения влияния продажи водки на оба этих признака.

Рассмотрим предикторы $\hat{Y}(X_2)$ и $\hat{X}_1(X_2)$ переменных Y и X_1 по переменной X_2, т.е. линейные предсказания продолжительности жизни и национального богатства в зависимости от продажи водки. Заметим, что на оба этих признака продажа водки действует одинаково отрицательно:

$$\hat{Y}(X_2) = 74.46 - 0.242X_2,$$
$$\hat{X}_1(X_2) = 6429 - 151.159X_2.$$

Составим таблицу, в которой будет отражена динамика продолжительности жизни и национального богатства при условии, если бы продажи водки не были связаны с объемами национального богатства.

Года	Усл.продолж-ть жизни $Y - \hat{Y}(X_2) + 74.46$	Усл.нац.богатсво (млрд.дол.) $X_1 - \hat{X}_1(X_2) + 6429$
1970	75	6196,3*
1975	74,9	6604,5*
1980	74,9	7023,8
1985	74,9	6931,2
1990	73,6	6850,9
1995	73,9	6975,5
1998	74,2	5163,3

Коэффициент корреляции между $Y - \hat{Y}(X_2) + 74.46$ и $X_1 - \hat{X}_1(X_2) + 151.159$ равен -0.11 ($\alpha_* = 0.4072$). Значимость 0.4072 указывает на отсутствие линейной зависимости между продолжительностью жизни и национальным богатством при условии независимости продаж водки от уровня национального богатства.

Оказывается, что между продолжительностью жизни и национальным богатством связи нет. Но при уменьшении национального богатства увеличивается продажа водки, что, в свою очередь, приводит к уменьшению продолжительности жизни. Так объясняется наблюдаемая в действительности положительная корреляции между продолжительностью жизни и национальным богатством.

2.10.4 Значимость прогноза

Основным источником вариации переменной Y является сумма квадратов отклонений от среднего значения $\bar{y}$ вида:

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

В математической статистике показано, что SS_T раскладывается на две суммы квадратов: $SS_T = SS_R + SS_E$, где

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

является источником вариации, обусловленным регрессией, а

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

является источником вариации, обусловленным ошибкой. Через

$$\hat{y}_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_s x_{sj}$$

обозначены значения наилучшего предсказания $\hat{Y}(X_1, \ldots, X_s)$.

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 $\hat{Y}(X_1, \ldots, X_s)$. При $s = 1$ множественный коэффициент корреляции равен обычному коэффициенту корреляции.

В качестве несмещенной оценки дисперсии σ^2 переменной Y используют выражение $\hat{\sigma}^2 = \frac{SS_E}{n-s-1}$. 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. Доверительный уровень вероятности $p = P\{F > F_*\}$ 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.

ПРИМЕР 1 (продолжение). Проверим значимость модели, т.е. проверим гипотезу о равенстве нулю коэффициента детерминации. $SS_T = 16.14$, $SS_R = 14.11$, $SS_E = 2.03$, $R^2 = 14.11/16.14 = 0.874$. Значение статистики Фишера равно

$$F = \frac{14.11/2}{2.03/(7-2-1)} = 13.88.$$

В случае справедливости гипотезы о равенстве $R^2 = 0$ статистика F имеет распределение Фишера с 2 и $7-2-1 = 4$ степенями свободы. Значимость равна $p = P\{F > 13.88\} = 0.0159$. Это означает, что с вероятностью, большей 0.9841 можно доверять прогнозу средней продолжительности жизни по объему национального богатства и объему продаж водки на душу населения.

2.11 Ранговая корреляция Спирмена

Довольно часто, особенно если данные неоднородны, обычный коэффициент корреляции не вполне четко отражает структуру взаимосвязи между признаками.

ПРИМЕР 2. Исследуем зависимость между смертностью от несчастных случаев и других внешних воздействий и объемом валового внутреннего продукта (ВВП) в 1997 году в 21 стране по данным, представленным в таблице 6. Эти страны в силу определенных исторических событий в конце двадцатого века делятся на две группы: Россия, Болгария, Венгрия, Польша, Румыния образуют одну группу, остальные страны другую.

Во второй группе наблюдается отрицательная корреляция между объемом ВВП и смертностью от несчастных случаев. Коэффициент корреляции равен $R = -0.46$, $p = 0.07$. Двумерная диаграмма, изображенная на рисунке 23, свидетельствует о неоднородности данных. В таком случае требуется продублировать статистические вычисления непараметрическими методами. Наиболее доступным в плане понимания является **ранговый коэффициент корреляции Спирмена**, буквально означающий коэффициент корреляции между рангами.

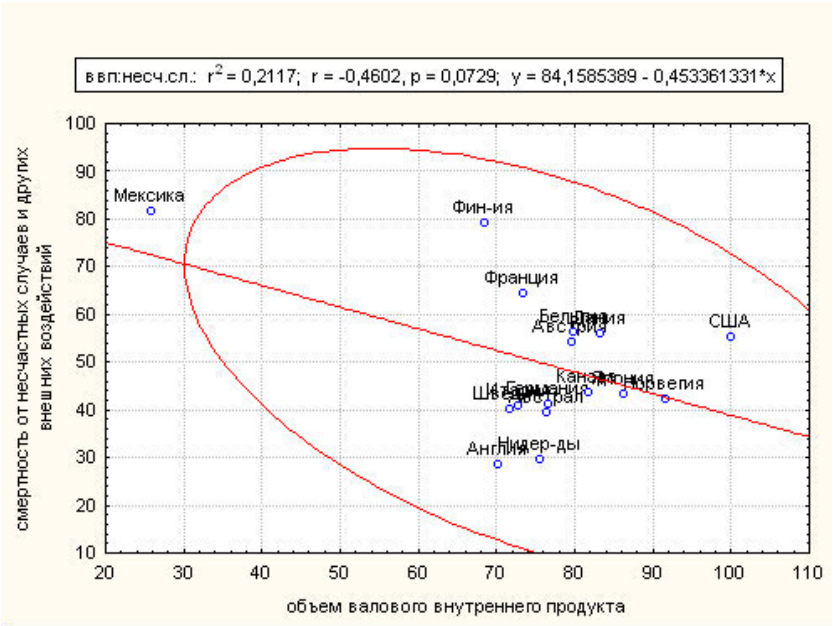

Рис. 23: Наблюдаемая отрицательная зависимость между объемом валового внутреннего продукта и смертностью от несчастных случаев во второй группе стран, $R = -0.46,\ P = 0.07$.

Напомним, что рангом называется порядковый номер элемента в упорядоченной выборке. Пусть $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ – ранги индивидов по признакам X и Y, $x_{i*} = k_i - l_i$, где k_i – количество индивидов, которых превосходит x_i, l_i– количество индивидов, которые превосходят x_i. Аналогично построим y_{i*}. Формально эти характеристики вводятся как

$$x_{i*} = 2\left(x'_i - \frac{n+1}{2}\right),$$
$$y_{i*} = 2\left(y'_i - \frac{n+1}{2}\right).$$

Если признаки взаимосвязаны – большему значению X соответствует большее значение Y, меньшему меньшее и так далее, – то x_{i*} и y_{i*} совпадают, и характеристика

$$Q_{xy} = \sum_{i=1}^{n} x_{i*} y_{i*}$$

принимает наибольшее значение, равное $Q = Q_{xx} = Q_{yy} = \sum_{i=1}^{n} x_{i*}^2 = \sum_{i=1}^{n} y_{i*}^2$. Если взаимосвязь признаков обратная – меньшему значению X соответствует большее значение признака Y и так далее, – то $x_{i*} = -y_{i*}$, и $Q_{xy} = -Q$. Таким образом можно ввести характеристику

$$R_S = \frac{Q_{xy}}{Q}, \tag{33}$$

которая называется коэффициентом корреляции Спирмена. Если между признаками существует прямая зависимость:

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ 1 & 2 & \dots & n-1 & n \end{pmatrix},$$

то коэффиент Спирмена равен единице $R_S = -1$, если обратная:

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix}.$$

Поскольку $\frac{n+1}{2}$ представляет собой среднее арифметическое последовательности $1, 2, \dots, n$, то коэффициент корреляции Спирмена есть не что иное, как коэффицент корреляции между рангами.

Значимость коэффициента Спирмена, как и обычного коэффициента корреляции, проверяется при помощи статистики

$$T = \frac{R_S\sqrt{n-2}}{\sqrt{1-R_S^2}}, \tag{34}$$

которая при некоррелированности рангов X' и Y' имеет распределение Стьюдента с $n-2$ степенями свободы.

ПРИМЕР 2. (продолжение) Ранговый коэффициент корреляции Спирмена между объемом ВВП в 1997 году и смертностью от несчастных случаев во второй группе стран равен $R_S = 0.03$, $P = 0.91$. Поскольку корреляция не значима, то наблюдаемое уменьшение смертности при увеличении объема ВВП является не систематическим, а случайным за счет одной страны (Мексики).

страна	группа	см-ть от несч.сл.	ВВП97	ест.пр.97
Россия	1	187,4	24,2	-5,2
Австрал	2	39,4	76,4	6,7
Австрия	2	54,3	79,5	0,5
Бельгия	2	56,3	79,8	1,2
Болгария	1	66,8	17,9	-7
Англия	2	28,5	70,2	1,6
Венгрия	1	102,4	33,5	-3,8
Германия	2	41,2	76,6	-0,8
Дания	2	55,9	83,2	1,5
Италия	2	40,8	72,8	-0,4
Канада	2	43,6	81,7	4,9
Мексика	2	81,8	25,8	23,4
Нидер-ды	2	29,8	75,6	3,5
Норвегия	2	42,4	91,6	3,5
Польша	1	77	24,7	0,9
Румыния	1	80,8	23,7	-1,9
США	0	55,5	100	5,7
Фин-ия	2	79,1	68,5	1,8
Франция	2	64,5	73,3	3,3
Швеция	2	40,3	71,6	-0,4
Япония	2	43,3	86,2	2,1

Таблица 6: Экономико-демографические данные 1997г.

x	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Таблица 7: Функция стандартного нормального распределения $\Phi(x)$. $\Phi(-x) = 1 - \Phi(x)$.