Н. П. Алексеева

Лекции по статистическим методам и моделированию для студентов 5 курса восточного факультета

Введение. О роли математических методов в гуманитарных науках. 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 области приложений математической статистики в последние годы наметился большой прогресс, обусловленный распространением использования вычислительной техники и повышением естественно-научного уровня ее пользователей.

Филологи сталкиваются с проблемой анализа данных, например, в лексикографии, например, когда необходимо выяснить, какой фактор вносит более значимый вклад в формирование акцента у школьников – семья, школа или среда обитания (местность).

Этнографы, страноведы, историки – для отсева не значимой информации и выявлении ведущих факторов. Например, насколько значим для продолжительности жизни фактор, связанный с объемами продаж алкогольной промышленности, или как сказывается на специфике питания в Камеруне то, что сельское хозяйство ориентировано на экспорт кофе.

Широкий круг специалистов интересуют особенности поведения людей или животных. Например, можно выяснить, насколько значимо отличается частота и длительность поднятия бровей у наркоманов, алкоголиков и нормальных людей. Ответы на подобные вопросы в состоянии значительно обогатить наши познания в той или иной области науки, или даже обнаружить некоторые скрытые от непосредственного наблюдения факты.

Задача данного курса – познакомить слушателей с некоторыми базовыми понятиями теории вероятностей, математической статистики и моделирования на примерах решения практических задач:

- определение объема выборки, необходимого для того, чтобы наблюдаемые частоты проявления некоторого значения признака отличалась от истинной частоты не более чем на 1%;
- проверка при помощи непараметрического критерия однородности Манна-Уитни того, насколько значимо отличие городов штатов Техас и Калифорния по некоторому признаку (годовое количество осадков, уровень образованности населения, преступности, процент афроамериканцев или людей, для которых родной язык не английский, детская смертность и т.д.);
- измерение значимости корреляции между теми же показателями при помощи рангового критерия Спирмена;
- статистическое подтверждение факта о стабильности показателей преступности по данным уголовной полиции Франции 1826–1831гг.
- решение задачи планирования эксперимента при ограниченности объектов наблюдения (дисперсионный анализ на латинских квадратах).

1. Повторение основных сведений из теории вероятностей

1.1. Классическое определение вероятности. Объектом исследования теории вероятностей является событие. Если при определенном комплексе условий событие обязательно происходит, то оно называется достоверным. Обычно его обозначают через Ω. Если при определенном комплексе условий событие не может произойти, то оно называется невозможным. Его обозначают через знак пустого множества $\varnothing$. Если событие может произойти или не произойти, то оно называется случайным. Случайные события мы обозначаем через $A, B, C, \ldots$. Мера осуществления некоторого случайного события есть вероятность.

Вероятность некоторого события A определяется как отношение числа элементарных исходов, благоприятствующих событию A, к общему числу исходов. Например, вы присутствуете на семинаре, где кроме вас, еще 19 человек, т.е. всего присутствует $n = 20$ студентов. Вероятность того, что преподаватель воспользуется бездушным датчиком случайных и попросит вас ответить на некоторый вопрос, равна $\frac{1}{20} = 0.05$. Если в группе 4 человека не знают ответа на вопрос и 16 знают, то при тех же условиях эксперимента вероятность того, что датчик случайных чисел выберет того, кто ответ знает, равна $P(A) = \frac{16}{20} = \frac{4}{5} = 0.8$.

1.2. Произведение событий, условная вероятность, независимость событий. Пусть в группе 10 юношей и 10 девушек. B – событие, которое заключается в том, что датчик случайных чисел выберет девушку, $P(B) = \frac{10}{20} = 0.5$. Из тех четырех, кто ответа на вопрос не знает, двое юношей и две девушки. Вероятность того, что датчик случайных чисел выберет девушку, которая знает ответа на вопрос, равна $P(AB) = \frac{8}{20} = 0.4$. Событие AB называется *произведением* событий, оно происходит тогда, когда события A и B происходят одновременно.

При изменении условий эксперимента изменится и количество элементарных исходов. *Условной вероятностью* $P(A|B)$ называется вероятность события A при условии, что событие B уже произошло. Эту вероятность удобно вычислять как отношение количества девушек, которые знают ответ на вопрос, к общему количеству девушек: $P(A|B) = \frac{8}{10} = 0.8$. Заметим, что эту вероятность мы могли бы вычислить как

$$P(A|B) = \frac{8}{10} = \frac{8/20}{10/20} = \frac{P(AB)}{P(B)} = 0.8.$$

Если условная вероятность совпадает с безусловной, т.е. $P(A|B) = P(A)$, то события A и B называются *независимыми.* В нашем случае это означает, что знание или незнание ответа на вопрос не зависит от пола, среди юношей и девушек одинаковый процент знатоков.

Нетрудно показать, что вероятность произведения независимых событий равна произведению вероятностей этих событий, т.е. $P(AB) = P(A)P(B)$. Например,

$$0.4 = P(AB) = P(A)P(B) = 0.8 \cdot 0.5.$$

1.3. Сумма событий. Событие $A + B$, которое происходит, когда происходит либо событие A, либо событие B, либо оба вместе, называется *суммой событий.* Например, A – студент знает ответ, B – случайно выбрана девушка. Возможны четыре события: AB, $A\bar{B}$, $\bar{A}B$ и $\bar{A}\bar{B}$, которые осуществляются с вероятностями $\frac{8}{20}, \frac{8}{20}, \frac{2}{20}, \frac{2}{20}$. Благоприятствующие сумме $A+B$ события – AB, $A\bar{B}$, $\bar{A}B$. Вероятность $P(A+B) = \frac{8+8+2}{20} = \frac{18}{20} = 0.9$. Иначе эту вероятность можно сосчитать как $P(A+B) = P(A) + P(B) - P(AB) = 0.8 + 0.5 - 0.4 = 0.9$.

Два события, сумма которых есть достоверное событие, т.е. какое-то из них обязательно происходит, но появление одного исключает появление другого, называются *противоположными.* Например, если вы выходите из аудитории, где сдают экзамен, то события "экзамен сдан" и "экзамен не сдан" являются противоположными. Вероятность противоположного события вычисляется из вероятности основного события как $P(\bar{A}) = 1 - P(A)$.

1.4. Повторные независимые испытания. Вспомним известный исторический анекдот. Шевалье де Мере 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 как какое-то количество очков все равно выпадет, $P(\Omega) = 1$. Событие B можно выразить как сумму $B = A_1 + A_2 + A_3 + A_4$. Вероятность $P(B)$ можно сосчитать двумя способами: либо найти вероятности $P(A_t)$, $t = 1, 2, 3, 4$ 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.

1.5. Биномиальный закон распределения. Обозначим через ξ случайное число выпадения шести очков в $n = 4$ 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 обычно обозначают число сочетаний k из n элементов.

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

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

В общем виде биномиальный закон распределения выражается в виде вероятностей получения k успехов в n независимых испытаниях при вероятности успеха p в единичном испытании:

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

Согласно продолжению анекдота, де Мере решил, что нашел путь к обогащению, но с ним довольно скоро никто не стал играть. Вряд ли этого человека можно обвинить в алчности, поскольку для проверки собственного вычисления вероятности того, что за 24 раза при подбрасывании двух кубиков хотя бы раз выпадут две "шестерки" (у него получилось $\frac{2}{3}$), он экспериментальным образом установил вероятность, близкую к половине. Озадаченный полученным несовпадением, этот "основоположник метода статистического моделирования" стимулировал оживление переписки между Ферма и Паскалем, которые аналитическим образом оба подтвердили его экспериментальный результат.

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.

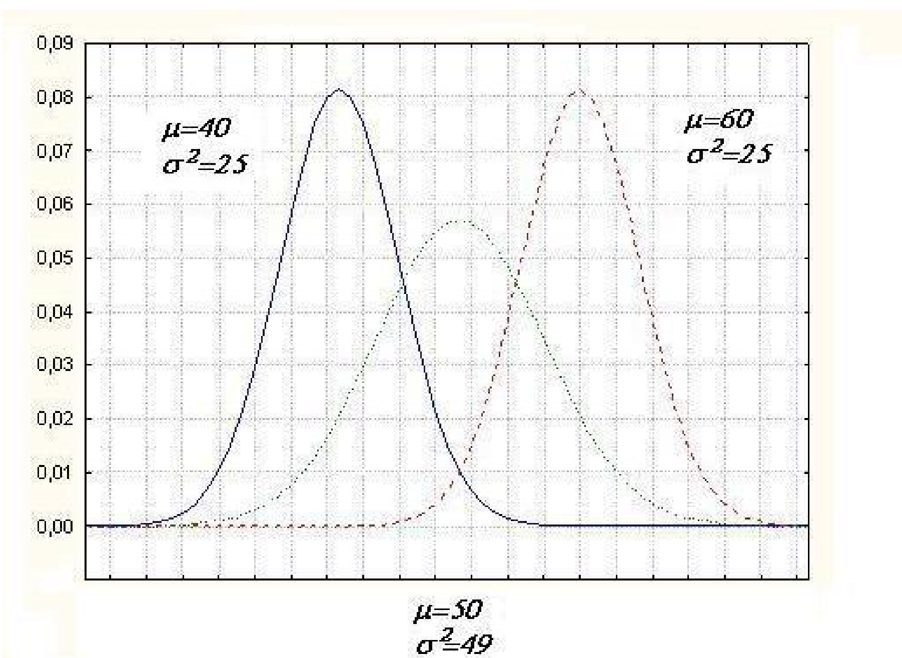

Рис. 1: Вероятности биномиального закона распределения при числе испытаний $n = 100$ и вероятности успеха $p = 0.4$, кривые нормального закона распределения с параметрами $\mu = 40$, $\sigma = 5$; $\mu = 60$, $\sigma = 5$; $\mu = 50$, $\sigma = 7$.

1.6. Нормальный закон распределения. При увеличении числа испытаний биномиальнй закон распределения может быть выражен аналитически в виде функции

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

Это означает, что если число испытаний велико, то для случайного числа успехов ξ из n независимых испытаний вероятность $P\{\xi = k\}$ может быть вычислена двумя способами: либо при помощи формулы (1) биномиального закона распределения, либо при помощи формулы (2) нормального закона распределения, где $x = k$.

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 (1) эта вероятность равна 0.048, по формуле (2) равна 0.046.

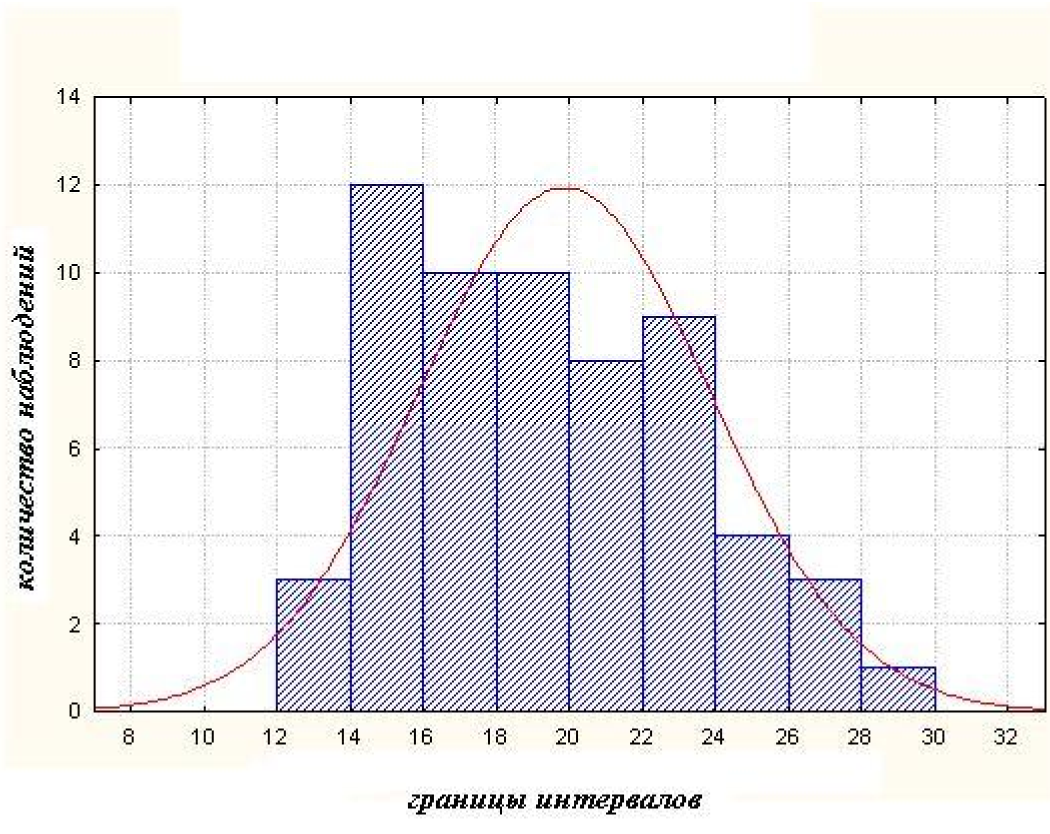

Рис. 2: Распределение числа некурящих студентов по 60 вузам.

В реальных задачах иногда важнее бывает определить не вероятность того, что число некурящих точно равно 15, а вероятность того, некурящих будет не больше 15. При помощи формулы (1) нужно вычислить все вероятности $P\{\xi = k\}$ для $k = 0, 1, \ldots, 15$ и затем их сложить. Должно получиться 0.128. В случае нормального закона распределения эта вероятность равна площади, ограниченной функцией (2) при $\mu = np$, $\sigma = \sqrt{np(1-p)}$ слева до значения $x = k$.

В Excel можно воспользоваться функцией НОРМРАСП(k;np;$\sqrt{np(1-p)}$;1), при помощи которой получается значение 0.106. Графики функции (2) при различных значениях параметров μ и σ изображены на рис. 1.

1.7. Среднее, дисперсия. Параметр μ нормального распределения имеет смысл среднего значения, характеризующего центр положения, параметр σ^2 – дисперсии (средний квадрат отклонения от среднего) . В примере среднее число некурящих равно 20, дисперсия равна 16.

Выражение, равное квадратному корню из дисперсии, называется стандартным отклонением σ. В примере с некурящими стандартное отклонение равно $\sigma = 4$. Известно, что с вероятностью 0.999 случайные наблюдения, распределенные по нормальному закону, сосредоточены в интервале от $\mu - 3\sigma$ до $\mu + 3\sigma$, т.е. вряд ли наблюдаемое число некурящих из 100 будет меньше 8 или больше 32 (рис.2). Можно показать, что 95% наблюдений сосредоточено в интервале от 12 до 28.

Табл.1. *Данные о городах штатов Калифорния и 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	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	15	7
SACRAMEN	CA	15	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	2	17
EL PASO	TX	1.2	245.4	3	13
AUSTIN	TX	3	217.8	6	16
FORT WOR	TX	2	281.1	4	18
ARLINGTO	TX	3.9	93	8	18
CORPUS C	TX	0.9	135	1	15

Как показывает опыт, понятия среднего и дисперсии лучше усваиваются при рассмотрении статистических задач. В табл.1 представлены статстические данные о городах штатов Техас и Калифорния. Используя среднее арифметическое, получаем, что в городах Калифорнии средний процент лиц азиатского происхождения равен 14,4, а в Техасе 2,3, среднеиюльская температура в Калифорнии 72,3, а в Техасе 84,5.

Оценки дисперсии показывают, насколько сильно выражен разброс данных, например, оценка дисперсии температуры в Калифорнии равна 42, а в Техасе 1.4. Т.е. для Техаса среднеиюльская температура не является информативной, поскольку по этому показателю города практически не отличаются.

1.8. Квантиль. Нормальное распределение (2) при $\mu = 0$ и $\sigma = 1$ называется стандартным нормальным распределением. В течении длительного времени, пока была недоступной вычислительная техника, для статистических вычислений приходилось пользоваться таблицами. Для краткости таблицы были составлены только для стандартного распределения. При необходимости вычислений вероятностей для распределений с другими параметрами μ и σ производились манипуляции, соответствующие преобразованиям системы координат.

По этой же причине многие статистики (функции от выборочных значений) преобразуются к виду случайных величин, распределенных по стандартному нормальному закону. Для них характерно, прежде всего то, что с вероятность 0.95 значения этих случайных величин сосредоточены на интервале от -1.96 до 1.96. Число 1.96 для стандартного нормального закона распределения называется 0.975-квантилью. Число 0.975 равно вероятности того, что случайная величина, распределенная по стандартному нормальному закону, меньше числа 1.96.

2. Проверка гипотез однородности зависимых выборок и доверительные интервалы.

2.1. Пример зависимых выборок. Рассмотрим первый реальный пример статистических данных (табл.2). 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 задушением, то скорее всего эта тенденция сохранится и в 1827 году. В инструкциях по статистическому анализу такие данные называются *зависимыми выборками.*

Табл.2. Данные уголовной полиции Франции (Кетле)

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

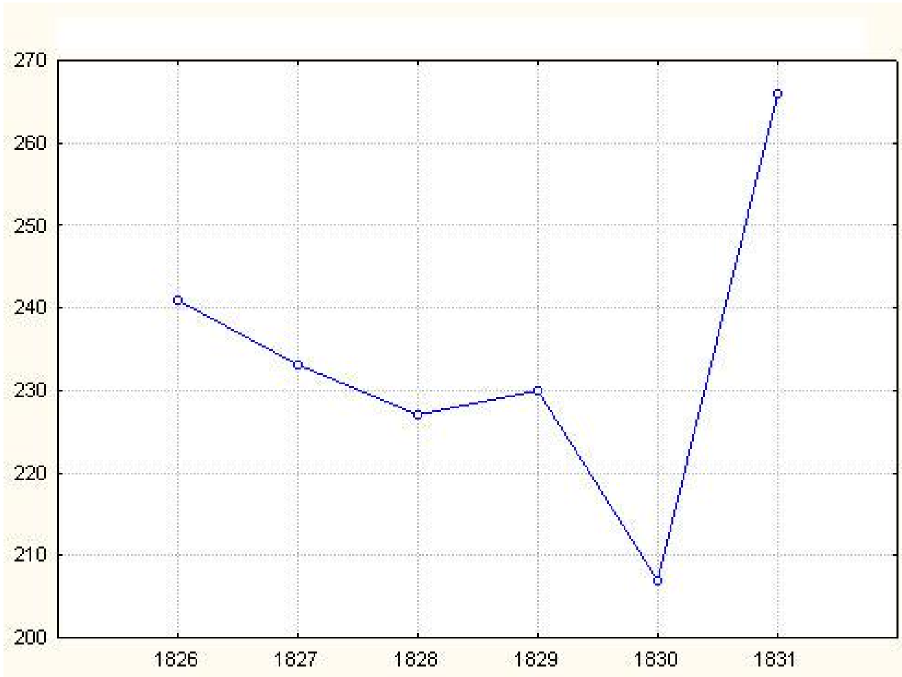

Рис. 3: Динамика числа уголовных преступлений во Франции в 1826-1831гг.

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.3). 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. Если бы во всех пяти случаях наблюдался рост числа убийств или, наоборот, все пять наблюдений соответствовали снижению числа преступлений, то все было бы понятно. А вот насколько значимо сказывается на общем снижении эти повышения, следует обратиться к статистической проверке гипотез.

2.2. Принцип маловероятных событий. Проверка статистических гипотез основана на принципе маловероятных событий, который заключается в том, что событие, вероятность которого мала, считается невозможным. Маловероятным событием в статистике считается событие, вероятность которого меньше 0.05.

2.3. Критерий знаков. Этот критерий используется для проверки гипотезы однородности 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.

Для каждого числа испытаний на основе принципа маловероятных событий вычисляется критическое число "успехов" (табл.3). Оно вычисляется как максимальное число успехов K, такое что $P\{X \leq K\} \leq 0.05$, где X – случайное число успехов.

Табл.3. *Соотношение между числом испытаний и критическим числом "успехов"*

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

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

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

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

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

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

Правило состоит в следующем: если наблюдаемое число "успехов" меньше критического, то проверяемую гипотезу следует отвергнуть, иначе нет оснований ее отвергнуть. При исследовании динамики убийств в 1827-1838гг. $n = 10$, $k = 5 > K = 1$, следовательно, нет оснований отвергнуть гипотезу о случайном изменении числа убийств.

Задача 1 (продолжение). В $n = 5$ испытаниях 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 табл.3, 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. Показать, что прирост численности населения городов с 1982 по 1992гг. изменился не значимо.

город	1982	1992	прирост
CHICAGO	2768483	3005072	+
BOSTON	551675	562994	+
VIRGINIA	417061	262199	-
ALBUQUER	398492	332920	-
OAKLAND	373219	339337	-
WICHITA	311746	279838	-
NEWARK	267849	329248	+
ST PETE	235306	238647	+
BATON RO	224704	220394	-

В четырех города из девяти наблюдался положительный прирост численности населения. По табл.3 критическое значение для 9 испытаний равно 1, следовательно, нет оснований отвергнуть гипотезу о случайном характере прироста населения.

Задача 4. Предположим, что из десяти пациентов восьмерым лекарство помогло, а двоим нет. Можно ли говорить об эффективности лекарства? В табл. 3 для $n = 10$ критическое число равно 1, оно меньше 2, следовательно, нет оснований отвергнуть гипотезу случайном эффекте препарата.

Задача 5. Пусть в группе из 18 человек 4 мальчика. По табл.3. 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.

В случае большего числа испытаний лучше использовать специальные компьютерные программы.

3. Проверка однородности независимых выборок.

3.1. Оценки средних, дисперсий, гистограммы. ПРИМЕР 3.1. Рассмотрим данные из табл.1. о проценте лиц азиатского происхождения. Обозначим через $x_1 = 9.8$ – этот процент в Лос-Анжелесе, $x_2 = 11.8$ – в Сан-Диего, и так далее до Стоктона – последнего представленного города из штата Калифорния – $x_{12} = 22.8$, $n_1 = 12$ – это объем выборки городов этого штата. Через $y_1, \ldots, y_{n_2}$, где $n_2 = 8$, обозначим значения рассматриваемого признака в восьми городах штата Техас. Для получения самого общего представления о том, в каком из штатов этот процент выше, вычислим средние значения:

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{9.8 + 11.8 + \ldots + 22.8}{12} = 14.4,$$
$$\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i = \frac{4.1 + 2.2 + \ldots + 0.9}{8} = 2.3.$$

Оценки дисперсии, показывающие, насколько велик разброс наблюдений от среднего значения, вычисляются по формулам:

$$S_x^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2 = \frac{1}{n_1}\left(\sum_{i=1}^{n_1}x_i^2 - n_1\bar{x}^2\right) =$$
$$= \frac{9.8^2+11.8^2+\ldots+22.8^2-12\cdot 14.4^2}{12-1} = 43.4,$$
$$S_y^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i-\bar{y})^2 = \frac{1}{n_2}\left(\sum_{i=1}^{n_2}y_i^2 - n_1\bar{y}^2\right) =$$
$$= \frac{4.1^2+2.2^2+\ldots+0.9^2-8\cdot 2.3^2}{8-1} = 1.6.$$

В Excel среднее и дисперсию можно оценить по выборке при помощи встроенных функций СРЗНАЧ() и ДИСП() (см. ПРИЛОЖЕНИЕ 1).

Получаем, что в штате Калифорния процент лиц азиатского происхождения выше, чем в штате Техас, и больше средний квадрат отклонения от среднего значения (дисперсия). Для этого построим гистограмму – график, показывающий как распределены города по рассматриваемому признаку (рис.4). На горизонтальной оси откладываются интервалы с шагом 5% – значения процентов лиц азиатского происхождения: $0\%, 5\%, 10\%$ так далее. По вертикальной оси откладывается количество наблюдений или количество 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%. Все это отражено в гистограмме.

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.7):

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

3.2. Критерий Манна-Уитни. Этот критерий используется для проверки гипотезы однородности независимых выборок. Он не зависит от вида выборочного распределения, поэтому относится к непараметрическим критериям.

Пусть имеются две независимые выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$ объемов n_1 и n_2 соответственно. Например, процент лиц азиатского происхождения в городах штатов Калифорния и Техас. Проверяемая гипотеза заключается в том, что по этому признаку штаты распределены одинаково, а наблюдаемое различие объясняется случайным характером наблюдений.

Для того чтобы построить статистику критерия, объединим эти выборки и поставим каждому наблюдению в соответствие ранг – его место в вариационном (упорядоченном) ряду (табл.4). Например, в Сан-Франциско самый высокий процент лиц азиатского происхождения, поэтому ему соответствует наибольший ранг 20, а в Корпусе-Кристи самый низкий процент 0.9%, ему соответствует самый низкий ранг 1. Для компактности в таблице 4 некоторые значения округлены.

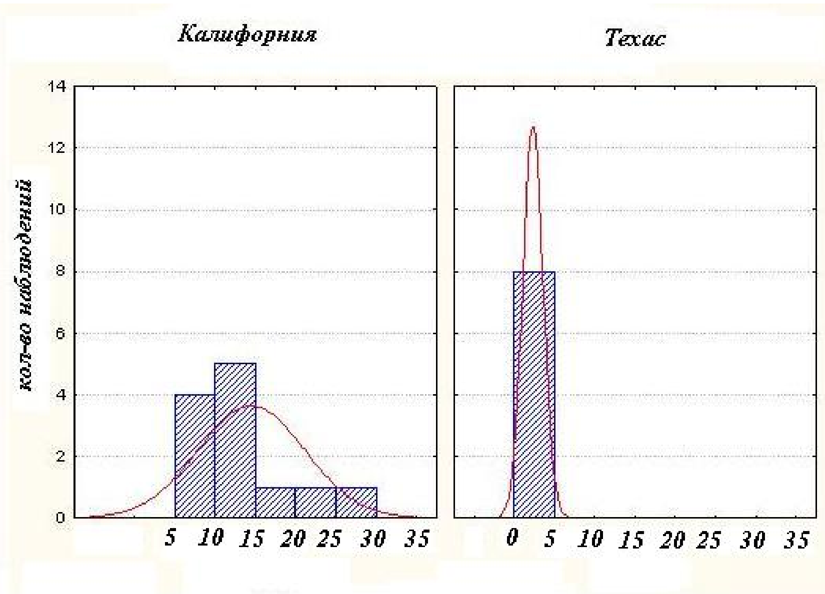

Рис. 4: Категориальные гистограммы признака – процент лиц азиатского происхождения в штатах Калифорния и Техас. Длина интервала группировки равна 5%.

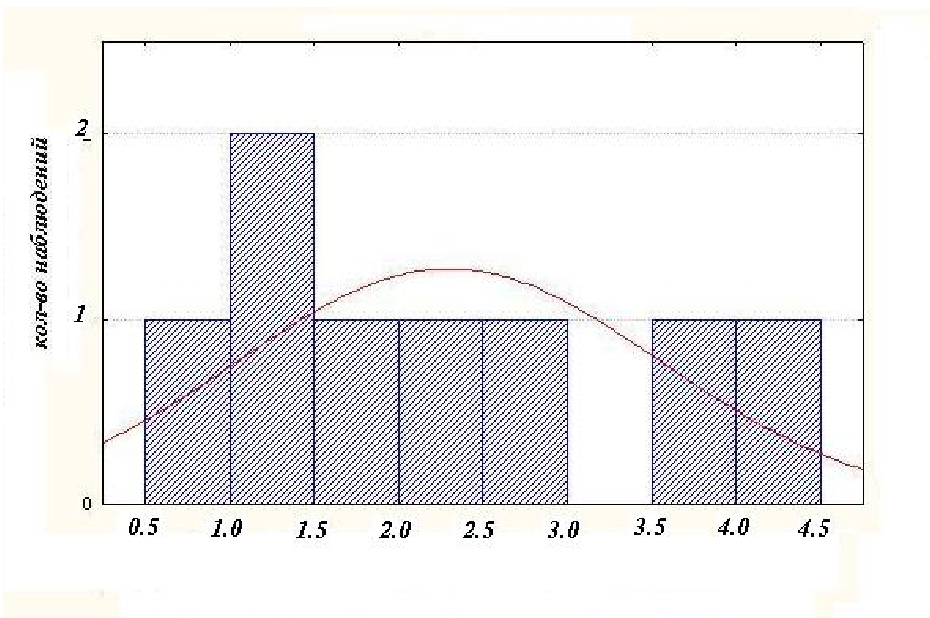

Рис. 5: Гистограмма признака – процент лиц азиатского происхождения в штате Техас. Длина интервала группировки равна 0.5%.

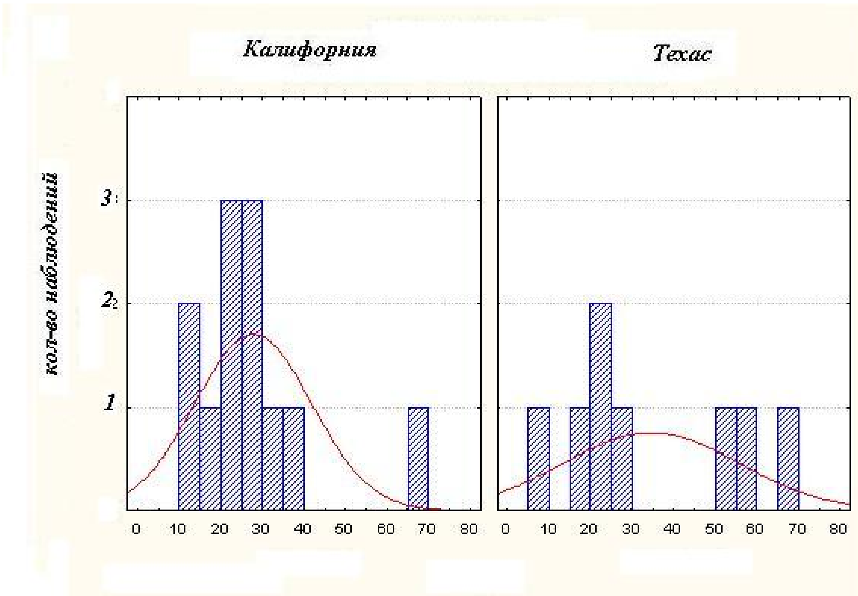

Рис. 6: Гистограммы признака – процент испаноязычного населения в штатах Калифорния и Техас. Длина интервала группировки равна 5%.

Табл.4. *Признак – процент лиц азиатского происхождения – и его ранги.*

	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

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.

Для проверки гипотезы H_0 о равенстве вероятностей $p_{xy} = P\{X < Y\}$ и $p_{yx} = P\{X > Y\}$ при альтернативной гипотезе $p_{xy} > p_{yx}$ 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. Этот критерий носит название критерия Вилкоксона. Однако он не очень удобен для использования, поскольку требует либо большого количества специальных таблиц, либо компьютерных вычислений. Рассмотрим модификацию этого критерия Манна-Уитни.

Сначала заметим, что общее число инверсий можно выразить через ранги.

$$U = (r_1 - 1) + (r_2 - 2) + \ldots + (r_{n_1} - n_1) =$$
$$= \sum r_i - \frac{n_1(n_1+1)}{2}.$$

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:

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

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

Если выборки однородны, то 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

$$Z = \frac{U - \mathbf{E}U}{\sqrt{\mathbf{D}U}} = \frac{R_2 - \frac{n_2(n_2+1)}{2} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}, \tag{3}$$

которая при справедливости нулевой гипотезы однородности выборок имеет стандартное нормальное распределение. Через R_2 обозначена сумма рангов выборки объема n_2. При наличии компьютера вычисляется доверительный уровень вероятности p как $p = P\{|Z| > |Z_*|\}$, а без него наблюдаемое значение статистики Z_* сравнивается с критическим $Z = 1.96$. Если $|Z_*| > 1.96$, то гипотеза однородности отвергается с уровнем значимости $\alpha = 0.05$.

ПРИМЕР 3.1. (продолжение) 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

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

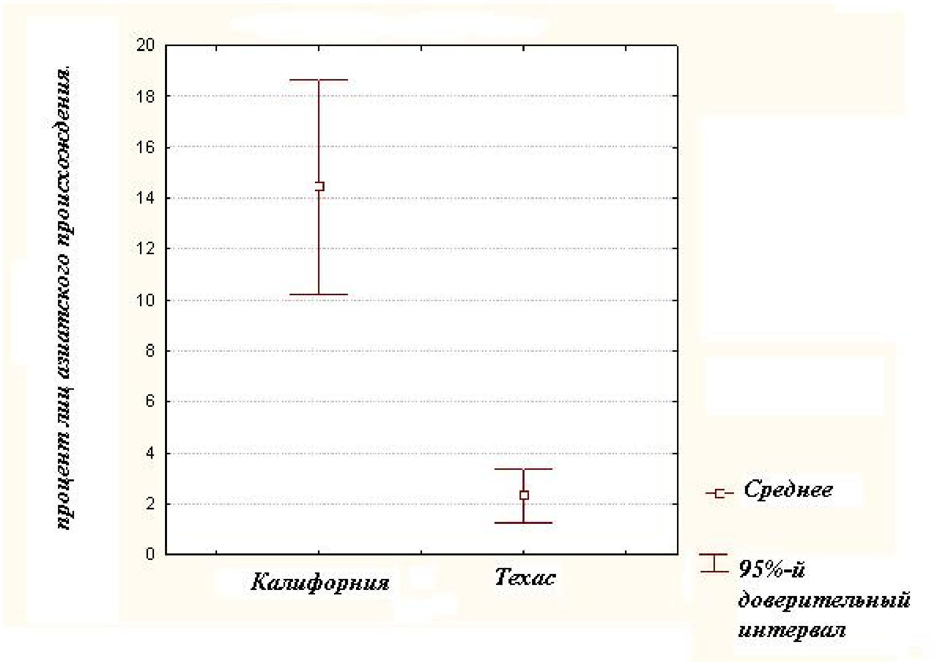

Рис. 7: Средние проценты лиц азиатского происхождения и 95%-е доверительные интервалы в городах штатов Калифорния и Техас.

Это равенство справедливо, так как всего имеется n рангов от 1 до n, и известно, что

$$1 + 2 + \ldots + n = \frac{n(n+1)}{2}.$$

$174 + 36 = 210 = \frac{20 \cdot 21}{2}$. 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 $DU = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = 168$,

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

Если рассматривать обратные инверсии $\bar{U}$, соответствующие $X < Y$, то за основу нужно взять сумму рангов R_1

$$U = R_1 - \frac{n_1(n_1+1)}{2} = 174 - \frac{12 \cdot 13}{2} = 96,$$
$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{DU}} = \frac{96 - 12 \cdot 8/2}{\sqrt{168}} = 3.7.$$

Наблюдаемые значения статистики $Z_* = -3.7$ и $Z_* = 3.7$ 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. То же самое получаем по доверительному уровню вероятности $p = 0.00022$, для вычисления которого в Excel можно воспользоваться функцией

$$f_x = 2 * (1 - \text{НОРМСТРАСП}(ABS(-3,7))).$$

так как $p = 0.00022 < 0.05$, то гипотеза однородности двух независимых выборок отвергается, и различие между городами двух штатов по проценту лиц азиатского происхождения нельзя объяснить случайностью.

ПРИМЕР 3.2. Проверим значимость различия средних процентов испаноязыяного населения городов штатов Калифорния и Техас. В Калифорнии в среднем 27.7% испаноязычного населения, в Техасе 34.4%. Дисперсии равны соответственно 196 и 450 (рис.8).

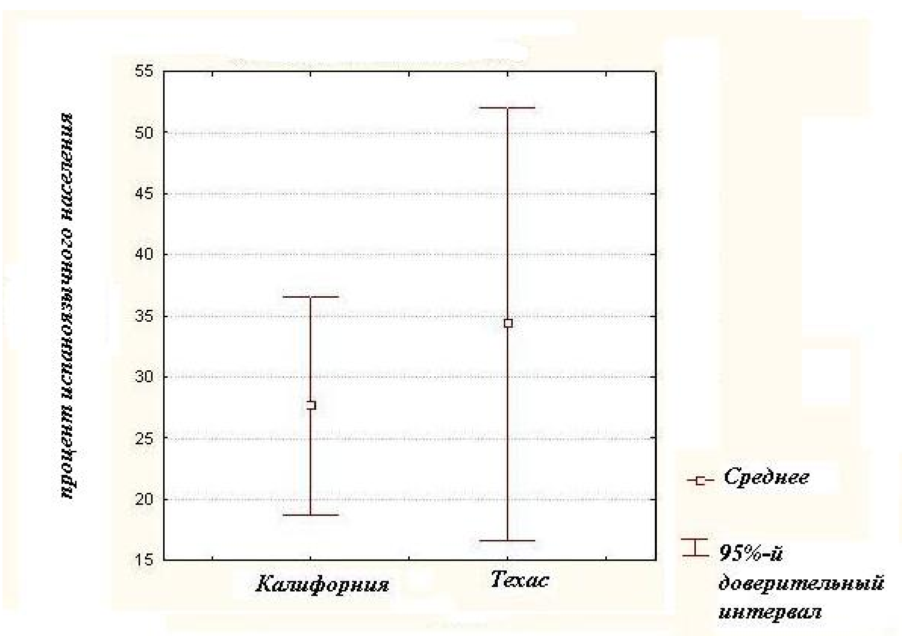

Рис. 8: Средние проценты испаноязычного населения и 95%-е доверительные интервалы в городах штатов Калифорния и Техас.

Объединим выборки процентов испаноязычного населения обоих штатов и вычислим ранги (табл.5). В случае совпадения значений признака ранги 2 и 3 заменяются на 2.5 и 2.5, и ранги 6 и 7 на 6.5 и 6.5. Сумма рангов меньшей выборки равна $R_2 = 88.5$, число инверсий равно $U = 88.5 - 8 \cdot 9/2 = 52.5$, значение статистики Z из формулы (3) равно $Z = \frac{52.5-12\cdot 8/2}{\sqrt{168}} = 0.35 = Z_*$, оно по абсолютной величине не больше 0.975-квантили стандартного нормального распределения 1.96, следовательно, нет оснований отвергнуть гипотезу однородности процентов испаноязычного населения штатов. То же самое подтверждаем при помощи доверительного уровня вероятности $p = 2 * (1 - \text{НОРМСТРАСП}(ABS(0,35))) = 0,73 > 0.05$.

Табл.5. *Признак – процент испаноязычного населения – и его ранги.*

	Калифорния (X)												Техас (Y)							
%	40	21	27	14	24	16	30	14	65	31	26	25	28	21	56	69	23	20	9	50
ранг	16	6.5	12	2.5	9	4	14	2.5	19	15	11	10	13	6.5	18	20	8	5	1	17

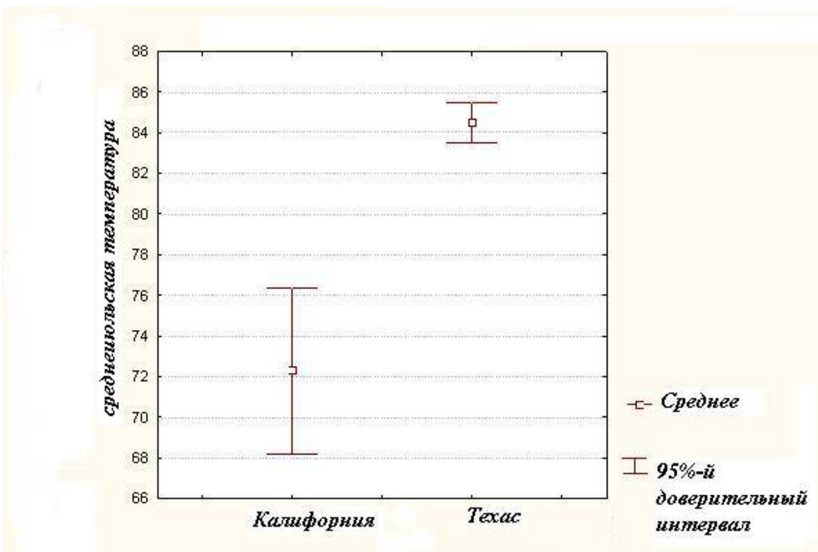

Рис. 9: Среднеиюльская температура и 95%-е доверительные интервалы в городах штатов Калифорния и Техас.

ПРИМЕР 3.3. Проверим, насколько значимо различие среднеиюльских температур в этих штатах. Можно ли считать штат Техас теплее штата Калифорния?

Табл.5. *Признак – среднеиюльская температура – и его ранги.*

	Калифорния (X)												Техас (Y)							
t^o	74	71	70	60	73	76	82	62	73	73	78	79	83	86	85	83	84	85	85	83
ранг	8	4	3	1	6	9	12	2	6	6	10	11	14	20	18	14	16	18	18	14

Сумма рангов в штате Техас $R_2 = 3 \cdot 14 + 16 + 3 \cdot 18 + 20 = 132$, $U = 132 - 8 \cdot 9/2 = 132 - 36 = 96$, $Z = \frac{96 - 8 \cdot 12/2}{\sqrt{168}} = \frac{96-48}{13} = 3.69 > 1.96$, следовательно, различие температур является значимым с $\alpha = 0.05$. То же самое подтверждаем при помощи доверительного уровня вероятности $p = 2 * (1 - \text{НОРМСТРАСП}(ABS(3,69))) = 0.0002 < 0.05$.

3.3. Критерий Краскела-Уоллиса. Этот непараметрический критерий предназначен для проверки гипотезы однородности сразу для нескольких независимых выборок.

ПРИМЕР 3.4. Рассмотрим данные о количестве ошибок, допущенными студентами, обучаемых по одной из трех методик. По этому количеству ошибок 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 величину

№	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

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

При выборе весов c_i была получена статистика Краскела-Уоллиса

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

которая асимптотически сходится к χ^2 распределению с $r-1$ степенью свободы.

$$\frac{12}{15 \cdot 16} \left(\frac{37.5^2}{7} + \frac{43^2}{5} + \frac{39.5^2}{3} \right) - 3 \cdot 16 = 6.84.$$

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.10). Доверительный уровень вероятности в данном случае равен $p = 0.03 < 0.05$, его можно вычислить в Excel при помощи функции f_x = ХИ2СТРАСП(6.84; 2)). Второй параметр означает число степеней свободы распределения статистики H, которое на единицу меньше количества групп.

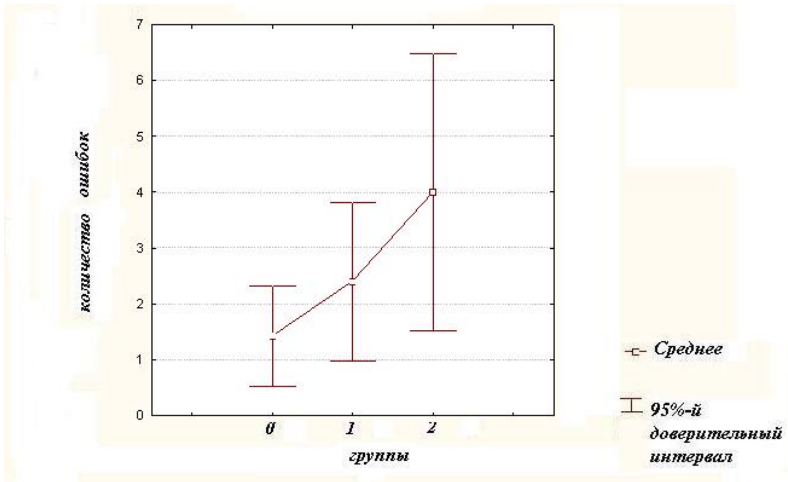

Рис. 10: Зависимость количества ошибок от методики преподавания.

3.4. Дисперсионный анализ на латинских квадратах. Иногда при малом количестве наблюдений удается спланировать эксперимент таким образом, чтобы учесть влияние сразу нескольких факторов.

ПРИМЕР 3.5. Исследуется группа из 16 студентов с различной учебной подготовкой: по 4 выпускника гимназии, лицея, специализированных и обычных школ. В каждой группе один дополнительно не занимался – будем обозначать таких студентов через A, другой B занимался самостоятельно, третий C – систематически с репетитором, четвертый D – с репетитором от случая к случаю. Таким образом, имеется два признака с одинаковым количеством градаций, равным $r = 4$. Это X_1 – фактор образования, X_2 – фактор дополнительного образования. Обозначим через x_{ij} количество ошибок у студента с фактором основного образования i, дополнительного образования j и представим данные в виде квадратной матрицы:

$X_1 \setminus X_2$	A	B	C	D
1	3	8	4	4
2	30	13	6	2
3	2	14	5	25
4	11	2	16	6

Такой план называют неповторяемым двуфакторным. На его основе можно выяснить, насколько значимо влияние факторов основного и дополнительного образования на количество допущенных ошибок. Однако эту задачу мы рассмотрим в структуре более общей задачи, связанной с особенностями тестирования.

Предположим, что проверка осуществляется по четырем тестам, ориентированным на исследование ресурсов памяти α, логического мышления β, любознательности γ и быстроты реакции δ. Вообще, для этого эксперимента требуется $4 \cdot 4 \cdot 4 = 64$ наблюдения. В реальности это оказывается практически невозможно.

При введении разной структуры тестов задача состоит в исследовании влияния на количество допущенных ошибок не только факторов основного и дополнительного образования, но и типа тестирования.

Обозначим через X_3 – фактор специализации проверочной работы, и спланируем эксперимент таким образом, чтобы чтобы у четырех представителей каждой школы были использованы все 4 варианта теста, и у четырех представителей с одинаковой системой дополнительной подготовки также были использованы 4 варианта теста, например, так:

	A	B	C	D
1	α_3	β_8	γ_4	δ_4
2	β_{30}	α_{13}	δ_6	γ_2
3	δ_2	γ_{14}	α_5	β_{25}
4	γ_{11}	δ_2	β_{16}	α_6

Такое размещение r объектов по r строкам и r столбцам, при котором каждый объект встречается один раз в каждой строке и один раз в каждом столбце, называется *латинским квадратом.*

Например, студент, окончивший лицей ($X_1 = 2$), занимающийся самостоятельно своей подготовкой ($X_2 = B$), в тесте α, ориентированном на проверку ресурсов памяти, допустил $x_{22(\alpha)} = 13$ ошибок.

Средний процент ошибок у выпускников гимназий равен $\bar{x}_{1*} = 4.75$, у выпускников лицеев $\bar{x}_{2*} = 12.75$, у выпускников спецшкол $\bar{x}_{3*} = 11.5$ и у выпускников обычных школ $\bar{x}_{4*} = 8.75$. Получаем средний процент ошибок $\bar{x}_{*1} = 11.5$ у студентов A, не использовавших возможности дополнительного образования, $\bar{x}_{*2} = 9.25$ у студентов B, которые занимались самостоятельно, $\bar{x}_{*3} = 7.75$ у студентов C, занимающихся систематически с репетиторами, $\bar{x}_{*4} = 9.25$ у студентов D, которые занимались дополнительно от случая к случаю. Средние проценты ошибок в тестах равны $\bar{x}_{**(\alpha)} = 6.75$, $\bar{x}_{**\beta} = 19.75$, $\bar{x}_{**(\gamma)} = 2.5$, $\bar{x}_{**(\delta)} = 8.75$. Общее среднее равно $\bar{x}_{**} = 9.4375$.

Нам нужно выяснить, насколько значимо от общего среднего $\bar{x}_{**} = 9.4375$ отличаются различные внутригрупповые средние. Представим, что наши случайные наблюдения количества ошибок складываются из нескольких слагаемых:

$$x_{ij(k)} = \mu + a_i + b_j + c_k + \varepsilon_{ij(k)}, \quad i, j = 1, \dots, r, \ k \in \{\alpha, \beta, \gamma, \delta\},$$

где μ генеральное среднее числа ошибок; $\varepsilon_{ij(k)}$ – случайная ошибка; c_k, a_i, b_j – дифференциальные эффекты факторов тестирования, основного и дополнительного образования. Их оценками являются величины

$$\hat{a}_i = \bar{x}_{i*} - \bar{x}_{**}, \ b_j = \bar{x}_{*j} - \bar{x}_{**}, \ c_k = \bar{x}_{**(k)} - \bar{x}_{**}.$$

Статистики для проверки гипотез $H_0 : a_i = 0$, $H_0 : b_j = 0$, $H_0 : c_k = 0$ о незначимости факторов строятся из сопоставления различных источников вариации, которые мы обозначим через Q, Q_1, Q_2, Q_3 Q_4:

- $Q = \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{r} (x_{ij(k)} - \bar{x})^2 = \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{r} x_{ij(k)}^2 - r^2 \bar{x}_{**}^2 = 1055.938$ – общий источник вариации;

- $Q_1 = r\sum_{i=1}^{r}(\bar{x}_{i*} - \bar{x}_{**})^2 = r\left(\sum_{i=1}^{r}\bar{x}_{i*}^2 - r\bar{x}_{**}^2\right) = 150.6871$ – источник вариации, обусловленный различием по виду основного образования;

- $Q_2 = r\sum_{j=1}^{r}(\bar{x}_{*j} - \bar{x}_{**})^2 = r\left(\sum_{j=1}^{r}\bar{x}_{*j}^2 - r\bar{x}_{**}^2\right) = 28.6875$ – источник вариации, обусловленный влиянием различной системой дополнительной подготовки;

- $Q_3 = r\sum_{k=\alpha,\beta,\gamma,\delta}(\bar{x}_{**(k)} - \bar{x}_{**})^2 = r\left(\sum_{k=\alpha,\beta,\gamma,\delta}\bar{x}_{**(k)}^2 - r\bar{x}_{**}^2\right) = 648.6875$ – источник вариации, обусловленный различием форм тестирования;

- $Q_4 = Q - Q_1 - Q_2 - Q_3 = 227.875$ называется ошибкой.

Поскольку $Q_3 > Q_2$, то, скорее всего различия между формами тестирования более выражены, чем различия между типами дополнительного образования. То, насколько

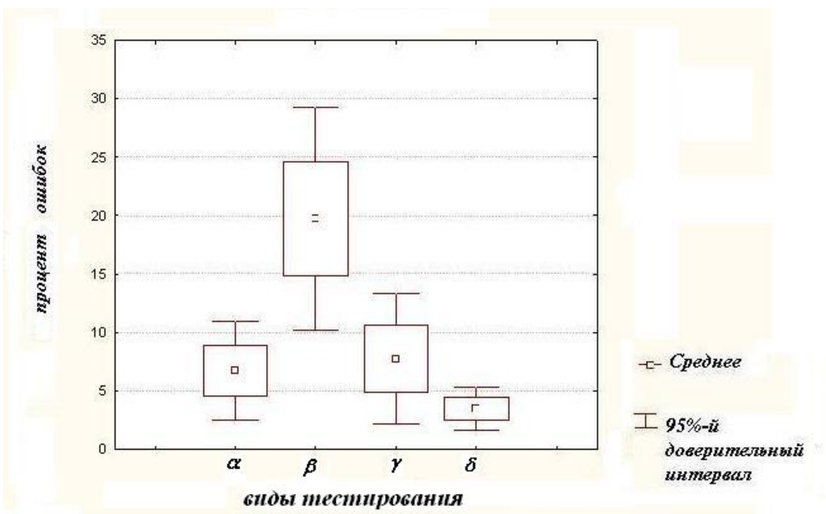

Рис. 11: Зависимость количества ошибок от вида тестирования.

это неравенство значимо, можно выяснить при помощи статистик

$$F_1 = \frac{Q_1/\nu_1}{Q_4/\nu_4} = 1.32, \quad F_2 = \frac{Q_2/\nu_2}{Q_4/\nu_4} = 0.25, \quad F_3 = \frac{Q_3/\nu_3}{Q_4/\nu_4} = 5.69,$$

$\nu_1 = \nu_2 = \nu_3 = (r-1) = 3$, $\nu_4 = (r-1)(r-2) = 6$. Значения этих статистик сравниваются с критическим значением $F_{\text{кр}}$, которое при $r = 4$ равно $F_{\text{кр}} = 4.76$. *Если наблюдаемое значение статистики больше критического, то гипотеза об отсутствии влияния соответствующего фактора отвергается.* Получаем, что значимым фактором оказывается фактор формы тестирования.

4. Критерии для проверки независимости

4.1. Ранговая корреляция Спирмена. Коэффициент корреляции – это численная характеристика, отражающая меру линейной зависимости между признаками. Напомним, что признаки X и Y зависимы линейно, если $Y = kX + b$, где k и b – некоторые числа. Коэффициент корреляции принимает значения от -1 до 1 и вычисляется по выборочным наблюдениям, лучше при помощи компьютера. Если коэффициент корреляции равен 1, то $Y = kX + b$, $k > 0$; если он равен -1, то $k < 0$; если он равен 0, то признаки считаются некоррелированными.

ПРИМЕР 4.1. Рассмотрим данные о смертности от несчастных случаев и других внешних воздействий, об объеме валового внутреннего продукта (ВВП) и естественном приросте населения в 1997 году в 21 стране.

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	-7.0
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

На рис.12 представлена двумерная диаграмма, иллюстрирующая отрицательную зависимость между смертностью от несчастных случаев и ВВП, на рис.13 – двумерная диаграмма, иллюстрирующая положительную зависимость между естественным приростом населения и ВВП, но при условии исключения из рассмотрения выпадающей из общей картины Мексики.

Из этих диаграмм видно, что данные неоднородны. Они подразделяются на три группы: к одной можно отнести такие страны как Россия, Польша, Венгрия, Румыния, Болгария, к другой – остальные страны, за исключением Мексики, и к третьей – саму Мексику.

Неоднородность данных заставляет осторожнее относиться к значимости этого результата и использовать дополнительно еще какую-нибудь методику для анализа такого рода данных. Одним из распространенных критериев, особенно, если количество наблюдений ограничено, является проверка значимости рангового коэффициента корреляции Спирмена.

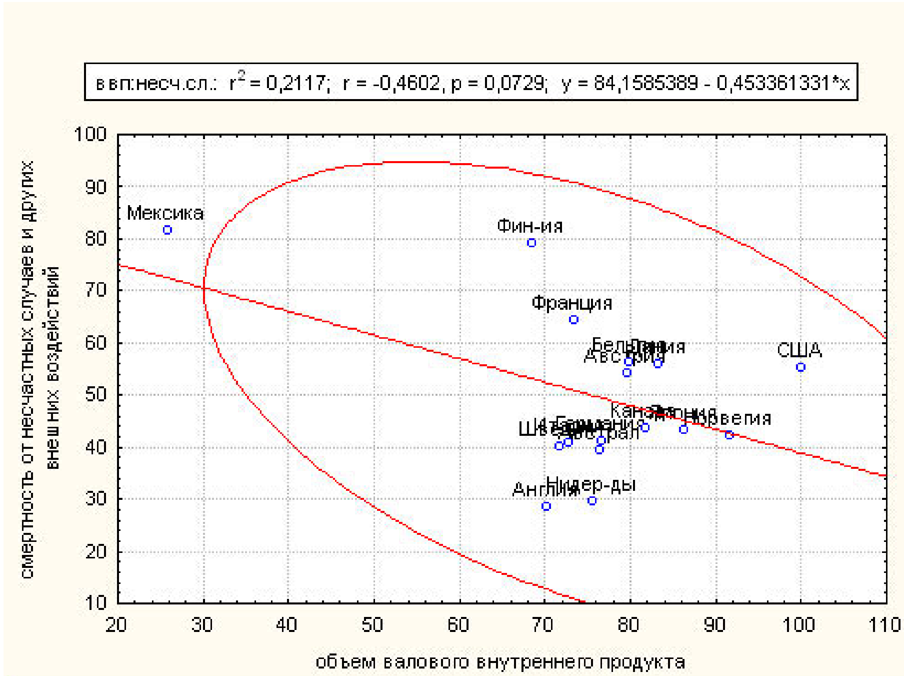

Рис. 12: Наблюдаемая отрицательная зависимость между объемом валового внутреннего продукта и смертностью от несчастных случаев во второй группе стран, $R = -0.46$.

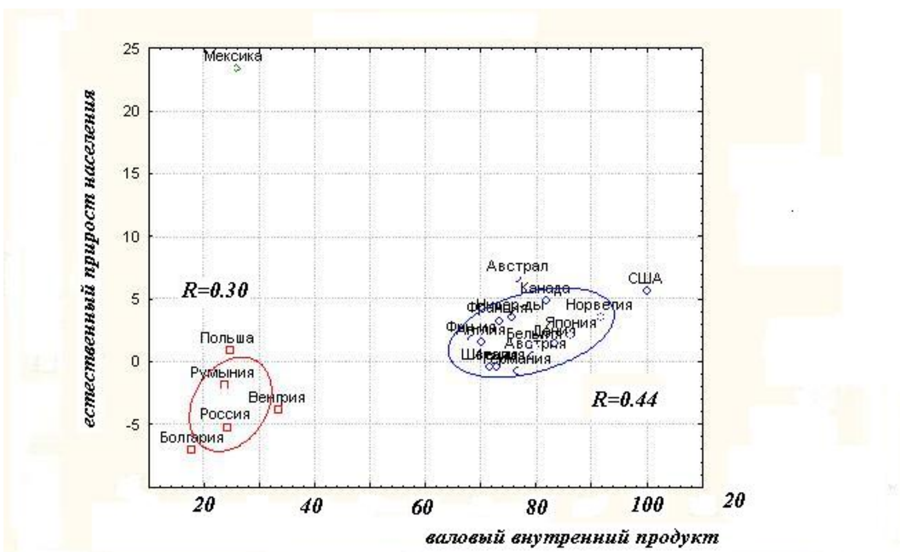

Рис. 13: Наблюдаемая положительная зависимость между объемом валового внутреннего продукта и естественным приростом населения в обеих группах стран, исключая Мексику.

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.

	значения					ранги				
страны	x_i ест.прир.	y_i ввп	x_{i*}	y_{i*}	$x_{i*}y_{i*}$	x'_i	y'_i	$x'_iy'_i$	$(x'_i)^2$	$(y'_i)^2$
Россия	-5.2	24.2	-2	0	0	2	3	6	4	9
Болгария	-7	17.9	-4	-4	16	1	1	1	1	1
Венгрия	-3.8	33.5	0	4	0	3	5	15	9	25
Польша	0.9	24.7	4	2	8	5	4	20	25	16
Румыния	-1.9	23.7	2	-2	-4	4	2	8	16	4
Сумма					20 Q_{xy}	15 R_x	15 R_y	50 R_{xy}	55 R_{xx}	55 R_{yy}

Обозначим через $x_{i*} = k_i - l_i$, где k_i количество индивидов, которых превосходит x_i, l_i– количество индивидов, которые превосходят x_i. Например, $x_1 = -5.2$ естественный прирост населения в России в 1997 году. $k_1 = 1$ – только у одной страны (Болгарии) естественный прирост ниже, чем у России. $l_1 = 3$ – у трех стран прирост выше, чем у России. Следовательно, $x_{1*} = 1 - 3 = -2$. Аналогично построим $y_{i*} = k_i - l_i$. Для России это значение равно $y_1 = 2 - 2 = 0$. Формально эти характеристики вводятся как

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

При $n = 5$ действительно получаем, $x_{1*} = 2(2-3) = -2, \quad y_{1*} = 2(3-3) = 0$.

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

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

принимает наибольшее значение, равное $Q = \sum\limits_{i=1}^{n} x_{i*}^2 = \sum\limits_{i=1}^{n} y_{i*}^2$. В рассматриваемом примере $Q = 0^2 + 2 \cdot 2^2 + 2 \cdot 4^2 = 40$. Если взаимосвязь признаков обратная, то есть 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

$$R_S = \frac{Q_{xy}}{Q}, \tag{4}$$

которая называется коэффициентом корреляции Спирмена. $R_S = \frac{20}{40} = 0.5$. Если между признаками существует прямая зависимость:

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

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

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

Поскольку $\frac{n+1}{2}$ представляет собой среднее арифметическое последовательности $1, 2, \ldots, n$, то коэффициент корреляции Спирмена есть не что иное, как коэффициент корреляции между рангами. Формула для вычисления R_S имеет вид:

$$R_S = \frac{R_{xy} - \frac{R_x R_y}{n}}{\sqrt{R_{xx} - \frac{R_x^2}{n}}\sqrt{R_{yy} - \frac{R_y^2}{n}}} = \frac{50 - \frac{15\cdot 15}{5}}{\sqrt{55 - \frac{15\cdot 15}{5}}\sqrt{55 - \frac{15\cdot 15}{5}}} = \frac{50 - 45}{\sqrt{55 - 45}\sqrt{55 - 45}} = \frac{5}{10}.$$

Значимость коэффициента R_s проверяется при помощи статистики

$$T = \frac{R_S\sqrt{n-2}}{\sqrt{1 - R_S^2}}, \tag{5}$$

которая при $H_0 : \varrho_S = 0$ (некоррелированности рангов X' и Y') имеет распределение Стьюдента с $n - 2$ степенями свободы. При отсутствии вычислительной техники можно воспользоваться таблицей критических значений для статистики Стьюдента. Если наблюдаемое значение статистики T по абсолютной величине превышает критическое значение, то гипотеза о $H_0 : \varrho_S = 0$ отвергается.

Таблица двусторонних критических значений статистики Стьюдента ($\alpha = 0.05$)

объем выборки n	5	6	7	8	9	10	11	12	13	14
$T_{\text{кр}}$	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23	2.20	2.18

Проверим значимость положительной корреляции между рангами естественного прироста и ВВП. $T = \frac{0.5}{\sqrt{1-0.5^2}}\sqrt{5-2} = 1 < 3.18$, следовательно, нет оснований отвергнуть гипотезу о некоррелированности признаков ВВП и естественного прироста, т.е. корреляция не значима. Для вычисления доверительного уровня вероятности $p = P\{|T| > |T_*|\} = 0.20 > 0.05$, согласно которому получается такой же вывод, можно воспользоваться функцией f_x = СТЬЮДРАСП$(1; 5-2; 2)$, параметрами которой являются значение статистики $T_* = 1$, число степеней свободы $n - 2 = 5 - 2$ и последнее число 2 означает, что рассматривался двусторонний критерий с нулевой гипотезой $H_0 : \varrho_S = 0$ и альтернативной гипотезой $H_0 : \varrho_S \neq 0$.

ПРИМЕР 4.2. Покажем, что в штате Техас отрицательная корреляция между процентом лиц азиатского происхождения и процентом испаноязычного населения не значима.

i	город	% лиц аз.пр. x_i	% исп. y_i	x_i'	y_i'	$x_i'y_i'$	$(x_i')^2$	$(y_i')^2$
1	HOUSTON	4.1	27.6	8	5	40	64	25
2	DALLAS	2.2	20.9	5	3	15	25	9
3	SAN ANTO	1.1	55.6	2	7	14	4	49
4	EL PASO	1.2	69	3	8	24	9	64
5	AUSTIN	3	23	6	4	24	36	16
6	FORT WOR	2	19.5	4	2	8	16	4
7	ARLINGTO	3.9	8,9	7	1	7	49	1
8	CORPUS C	0.9	50.4	1	6	6	1	36
	Сумма			36 R_x	36 R_y	138 R_{xy}	204 R_{xx}	204 R_{yy}

$$R_S = \frac{138 - \frac{36 \cdot 36}{8}}{\sqrt{204 - \frac{36 \cdot 36}{8}}\sqrt{204 - \frac{36 \cdot 36}{8}}} = \frac{4}{7} = -0.57, \quad n = 8,$$

$$T_* = \left| \frac{-0.57\sqrt{8-2}}{1 - 0.57^2} \right| = 1.7 < 2.45.$$

Следовательно, гипотеза о равенстве нулю корреляции Спирмена не отвергается, отклонение от нуля корреляции -0.57 для выборки объема 8 можно объяснить случайностью. Для вычисления доверительного уровня вероятности $p = P\{|T| > |T_*|\} = 0.14 > 0.05$, согласно которому получается такой же вывод, можно воспользоваться функцией f_x = СТЬЮДРАСП$(abs(-1, 7); 8 - 2; 2)$, параметрами которой являются значение статистики $T_* = -1.7$, число степеней свободы $n - 2 = 8 - 2$ и последнее число 2 означает, что рассматривался двусторонний критерий с нулевой гипотезой $H_0 : \varrho_S = 0$ и альтернативной гипотезой $H_0 : \varrho_S \neq 0$.

ПРИЛОЖЕНИЕ 1. **Задание по моделированию.**

1. Промоделировать в Excel при помощи датчика случайных чисел "слчис()"$N = 50$ раз случайный выигрыш в игре: из 36 игральных карт наугад вынимается одна карта. Если это туз или шестерка (p=2/9), то выигрыш $a = 35$ руб., иначе проигрыш $b = 10$ руб. Найти по формулам средний случайный суммарный в 50 партиях выигрыш $M = (ap + (-b)(1-p))N$ и дисперсию $D = (a^2 p + (-b)^2(1-p) - M^2)N$.
2. Повторить 30 раз эксперимент (как будто играем целый месяц каждый день). Оценить по выборке n=30 средний выигрыш (функция "срзнач") и дисперсию (функция "дисп"). Сравнить результат с теоретическими характеристиками M и D.
3. При помощи функции "нормрасп"найти вероятность того, что а) проигрыш больше 200, б) выигрыш больше 200, в) от-100 до 100.
4. Найти относительную частоту выигрыша более 200 рублей.

Сначала вычислим размер среднего ожидаемого суммарного выигрыша, который вычисляется по указанной формуле:

$$M = \left(35 \cdot \frac{2}{9} - 10 \cdot \frac{7}{9}\right) 50 = \left(\frac{35 \cdot 2 - 10 \cdot 7}{9}\right) 50 = 0,$$

и дисперсию случайного суммарного выигрыша

$$D = \left(35^2 \cdot \frac{2}{9} + (-10)^2 \cdot \frac{7}{9} - 0^2\right) N = \frac{(1225 \cdot 2 + 100 \cdot 7)\, 50}{9} = 17500.$$

Стандартное отклонение равно $S = \sqrt{D} = 132$. Так как для нормально распределенной случайной величины Z имеет место $P\{M - \sqrt{D} < Z < M + \sqrt{D}\} = 0.68$, то случайный суммарный выигрыш будет находиться в пределах от -132 до 132 в 68% случаев. Итак, обозначим через X случайный суммарный выигрыш, который, будем считать, имеет нормальное распределение со средним $M = 0$ и стандартным отклонением $S = \sqrt{D} = 132$. Для вычисления вероятности того, что проигрыш больше 200, что означает, что $P\{X < -200\}$, воспользуемся функцией

f_x=НОРМРАСП(-200;0;132;1)=0.065.

Для вычисления вероятности того, что выигрыш больше 200, что означает, что $P\{X > 200\} = 1 - P\{X < 200\}$, воспользуемся функцией

f_x=1-НОРМРАСП(200;0;132;1)=0.065.

Для вычисления вероятности того, что выигрыш находится в пределах от -100 до 100, что означает, что $P\{-100 < X < 100\} = P\{X < 100\} - P\{X < -100\}$, воспользуемся функцией

f_x=НОРМРАСП(100;0;132;1)-НОРМРАСП(-100;0;132;1)=0.776-0.224=0.551.

Для моделирования случайного однократного выигрыша воспользуемся встроенными функциями:

f_x=ЕСЛИ(СЛЧИС()<2/9;35;-10) (1)

При нажатии "Enter"случайным образом будет появляться либо число 35, либо -10. Частота их появления связана с указанными параметрами: в 22.2% случаев будет появляться 35, в 77.8% случаев будет появляться -10. Скопируем ячейку с формулой (1) 50 раз по горизонтали и 30 раз по вертикали. Получим таблицу, состоящую из 50 столбцов и 30 строчек со случайными выигрышами. Просуммируем выигрыши в каждой из 30 строк и получим выборку случайных суммарных выигрышей из 30 наблюдений, которая располагается в ячейках $AY1 : AY30$, если начинать с ячейки $A1$. Например, для выборки

85	40	−50	−95	220	40
−50	−5	175	−185	40	40
130	−5	−5	−185	−95	220
−140	85	−5	355	85	−140
130	40	40	265	85	40

найдем оценки среднего $\hat{M}$ и дисперсии $\hat{D}$ при помощи функций

f_x=СРЗНАЧ(AY1:AY30)=38.5= $\hat{M}$ и f_x=ДИСП(AY1:AY30)=16687=$\hat{D}$.

Оценка стандартного отклонения имеет вид: $\hat{S} = \sqrt{16687} = 129.2$. Для вычисления абсолютных частот воспользуемся функциями

f_x=СЧЁТЕСЛИ(AY1:AY30;">200")=4,
f_x=СЧЁТЕСЛИ(AY1:AY30;"<-200")=0,
f_x=СЧЁТЕСЛИ(AY1:AY30;"<100")=23,
f_x=СЧЁТЕСЛИ(AY1:AY30;"<-100")=4,

откуда получаем, что относительная частота суммрного выигрыша более 200 рублей равна $\frac{4}{30} = 0.13$, проигрыша более 200 рублей равна 0, а относительная частота того, что выигрыш находится от -100 до 100 равна $\frac{23-4}{30} = 0.63$. Результаты вычислений соберем в итоговой таблице.

характеристика	теоретическая	оценка
среднее	$M = 0$	$\hat{M} = 38.5$
дисперсия	$D = 17500$	$\hat{D} = 16687$
стандартное отклонение	$S = 132$	$\hat{S} = 129.2$
$P\{X < -200\}$	0.065	0.13
$P\{X > 200\}$	0.065	0
$P\{-100 < X < 100\}$	0.551	0.63

ПРИЛОЖЕНИЕ 2. **Проверка однородности по ранговому критерию Вилкоксона для зависимых выборок**.

По данным уголовной полиции Франции о количестве десяти видов преступлений в 1829 и 1845 гг. составляем таблицу для вычисления статистики Вилкоксона.

Расчет статистики Вилкоксона

(1)	(2)	(3)	(4)	(5)	(6)	(7)
вид прест. i	1829 x_i	1845 y_i	разность $y_i - x_i$	модуль разности $\|y_i - x_i\|$	ранги ранги	ранги с учетом один.знач.
1	2	2	0	0	-	-
2	45	46	1	1	1	1
3	46	44	-2	2	2	2.5
4	0	2	2	2	3	2.5
5	1	4	3	3	4	4
6	61	57	-4	4	5	5
7	7	12	5	5	6	6
8	23	17	-6	6	7	7
9	21	11	-10	10	8	8
10	24	12	-12	12	9	9

Заметим, что при выставлении рангов не рассматриваются неинформативные случаи, когда $x_i = y_i$. Суммируем ранги (7), относящиеся к отрицательным и положительным разностям.

$$S_{-} = 2.5 + 5 + 7 + 8 + 9 = 31.5, \qquad S_{+} = 1 + 2.5 + 4 + 6 = 13.5.$$

В качестве статистики S выбираем меньшее из них, т.е. $S_* = min\{S_-, S_+\} = 13.5$. Математическое ожидание и дисперсия статистики S равны

$$\mathbf{E}S = \frac{n(n+1)}{4} = \frac{9 \cdot 10}{4} = 22.5$$

$$\mathbf{D}S = \frac{n(n+1)(2n+1)}{24} = \frac{9 \cdot 10 \cdot 19}{24} = 71.25$$

Затем вычисляется статистика

$$Z = \frac{S - \mathbf{E}S}{\sqrt{\mathbf{D}S}},$$

которая в случае однородности для зависимых выборок имеет стандартное нормальное распределение. Ее значение равно

$$Z_* = \frac{13.5 - 22.5}{\sqrt{71.25}} = -1.06623.$$

Для проверки гипотезы однородности вычисляется доверительный уровень вероятности $P = P\{|Z| > |Z_*|\} =$ НОРМСТРАСП$(-1.06623) = 0.14316$, который сравнивается с уровнем значимости $\alpha = 0.05$. $P = 0.14316 > 0.05$, следовательно, нет оснований отвергнуть гипотезу однородности, и динамика числа преступлений можно объяснить случайностью.

Если $P < 0.05$, то гипотеза однородности для зависимых выборок отвергается, и изменения в динамике числа преступлений оказываются значимыми в сторону увеличения или уменьшения.