О.В.Иванов

СТАТИСТИКА

учебный курс для социологов и менеджеров

Часть 2

Доверительные интервалы

Проверка гипотез

Методы и их применение

Москва
2005

Иванов О.В. Статистика / Учебный курс для социологов и менеджеров. Часть 2. Доверительные интервалы. Проверка гипотез. Методы и их применение. – М. 2005. – 220 с.

Учебный курс подготовлен для преподавания студентам-социологам и менеджерам в составе цикла математических дисциплин. Соответствует Государственному образовательному стандарту высшего профессионального образования по специальностям «Социология» и «Менеджмент». Содержит теоретическую часть, примеры, а также задачи для аудиторных и самостоятельных занятий.

Книга может быть полезна преподавателям, студентам, научным сотрудникам, аналитикам, всем, кто занимается прикладным статистическим анализом социальных и экономических данных.

По вопросам, связанным с настоящим изданием, обращаться на кафедру социальной информатики социологического факультета МГУ им.М.В.Ломоносова по электронному адресу: info@socio.msu.ru

ОГЛАВЛЕНИЕ

Предисловие

Этот учебник посвящен второй части общего курса статистики для студентов гуманитарных специальностей. На это указывает, в том числе, нумерация глав. Первая часть курса содержала основы описательной статистики и теории вероятностей. Она завершилась разделом о теоретико-вероятностных основаниях статистического вывода. Во вторую часть включены разделы аналитической статистики: построение доверительных интервалов, проверка статистических гипотез. Особое место отведено непараметрическим методам, значение которых для исследований в гуманитарных областях трудно переоценить.

В учебник включены практические примеры и задачи для решения на семинарских занятиях и самостоятельно. Это позволяет использовать учебник в качестве основного по курсу статистики. Электронная версия учебника и материалы лекций находятся в Интернете по адресу: informatics.socio.msu.ru.

Курс предполагает лекции и семинарские занятия, частично проводимые в компьютерном классе с параллельным изучением статистических пакетов. Базовым пакетом является SPSS.

Автор выражает большую признательность своим коллегам: Соколихину А.А., Самыловскому А.И., Семенову К.Е., Коченкову А.И., Астаховой Н.В. за активную поддержку и помощь в подготовке настоящего учебного курса.

Глава 9. Доверительные интервалы

Эта глава посвящена построению интервальных оценок для параметров генеральной совокупности. На основе анализа выборки мы научимся строить доверительные интервалы, которые с заданной вероятностью содержат значение оцениваемого параметра. В первом параграфе обсуждены точечные и интервальные оценки. В последующих параграфах строятся доверительные интервалы для среднего генеральной совокупности с нормальным законом распределения, для доли признака и для дисперсии.

9-1 Точечные и интервальные оценки

Мы неоднократно отмечали, что важнейшей задачей при проведении исследований является получение информации о генеральной совокупности, ее свойствах и характеристиках. Поскольку генеральные совокупности, как правило, велики, нам приходится ограничиваться детальным изучением выборки, а затем на этой основе делать выводы об изучаемой генеральной совокупности. Эта глава – первая, в которой мы будем делать такие выводы.

Чтобы подойти к этому, нам пришлось познакомиться с методами получения и исследования выборки, с теоретико-вероятностными основаниями для статистических заключений. Теперь мы переходим к изучению аналитической статистики. Напомним, что **аналитическая статистика** включает методы, которые на основе изучения выборочных характеристик позволяют получать выводы о характеристиках генеральной совокупности. Одним из значительных разделов аналитической статистики является оценивание **параметров** генеральной совокупности.

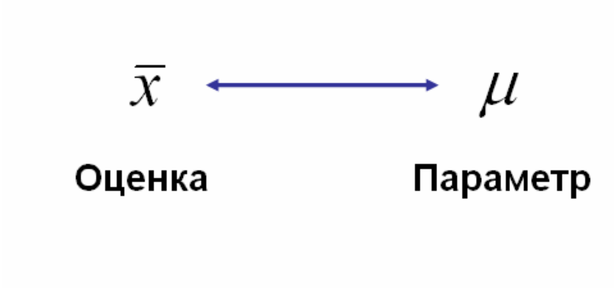


Рисунок 9-1. Статистика есть оценка параметра

Точечные оценки и их критерии

Статистики, вычисляемые по выборке, являются оценками для параметров генеральной совокупности. **Параметр** генеральной совокупности - фиксированное число, которое нам не известно. При его вычислении случайность отсутствует. Тем самым, параметр есть неизвестная и фиксированная величина.

Статистикой мы назвали числовую характеристику выборки. Статистика является случайной величиной, так как в ее основе лежат данные, полученные в результате случайного отбора. Тем самым, статистика является известной и случайной величиной.

Статистики являются **оценочными функциями** параметров генеральной совокупности. Фактическое значение статистики, рассчитанное по данным выборки, мы назвали **оценкой** параметра генеральной совокупности. Оценки бывают точечные и интервальные.

Точечной оценкой (point estimate) называется отдельное число, которое используется в качестве оценки параметра генеральной совокупности.

Например, среднее значение выборки является точечной оценкой среднего значения генеральной совокупности (рисунок 9-1). Доля признака, рассчитанная по выборке, есть оценка для доли признака в генеральной совокупности.

Ошибкой оценки (estimation error) называют разность между оцениваемым параметром генеральной совокупности и оценкой, рассчитанной на основе выборки.

$$\text{Ошибка оценки} = \text{Параметр} - \text{Оценка}$$

Ошибка оценки обычно неизвестна, поскольку неизвестен оцениваемый параметр генеральной совокупности.

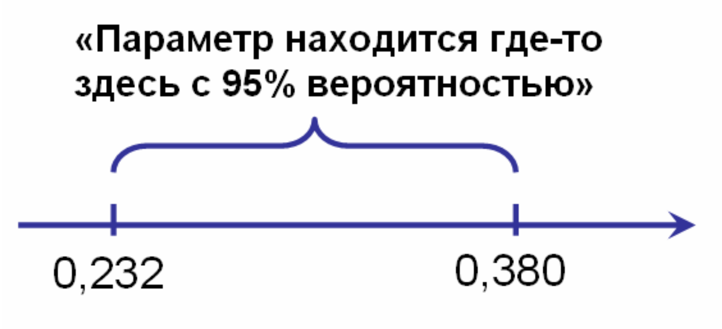


Рисунок 9-2. Интервальная оценка параметра

Поскольку возможны различные оценки для одного и того же параметра, они бывают «хорошими» и «плохими». Считается, что «хорошие» оценки должны удовлетворять следующим критериям.

Критерии точечных оценок:

Несмещенность оценки (unbiased estimator) означает, что математическое ожидание точечной оценки равно значению оцениваемого параметра генеральной совокупности.

Эффективность оценки (relatively efficient estimator) означает, что статистика, используемая в качестве точечной оценки параметра генеральной совокупности, имеет минимальную стандартную ошибку.

Состоятельность оценки (consistent estimator) означает, что по мере увеличения объема выборки значение точечной оценки приближается к значению оцениваемого параметра генеральной совокупности.

Выборочное среднее удовлетворяет всем трем названным критериям и поэтому является наилучшей оценкой для среднего генеральной совокупности.

Интервальные оценки

Наряду с точечными оценками мы будем рассматривать интервальные оценки. Важнейшим для нас является понятие доверительного интервала.

Доверительный интервал (confidence interval) – вычисленный на основе выборки интервал значений признака, который с известной вероятностью содержит оцениваемый параметр генеральной совокупности.

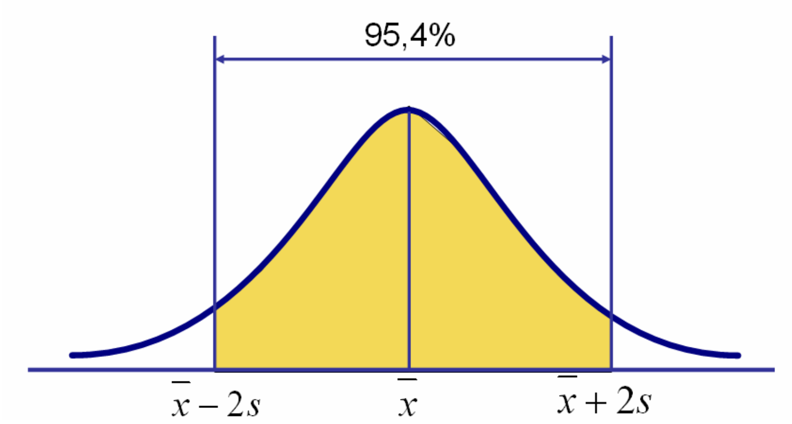


Рисунок 9-3. Интервал в 95,4% для нормального закона

Говорят, «мы на 95% уверены, что доля людей, которым известна наша торговая марка находится где-то между 23,2% и 38,0%». Мы проиллюстрировали это высказывание рисунком 9-2. Это и есть интервальная оценка. Наша задача состоит в том, чтобы научиться строить доверительные интервалы, основываясь на надежных статистических методах.

Доверительная вероятность (или уровень доверия, confidence level) – это вероятность того, что доверительный интервал содержит значение параметра.

Доверительную вероятность принято устанавливать на уровнях 90%, 95% и 99%. Чем выше доверительная вероятность, тем более широкий и менее полезный интервал мы получим. Если доверительная вероятность не задана, мы будем считать, что она равна 0,95 или 95%. Для нормального закона мы отмечали, что в пределах плюс минус двух стандартных отклонений относительно среднего находится 95,4% значений случайной величины.

Существует несколько различных форм записи доверительных интервалов, каждой из которых мы воспользуемся, понимая, что они эквивалентны и применяются в зависимости от удобства и контекста.

Вариант 1. Запись текстом. «Мы на 95% уверены, что среднее значение роста студентов находится где-то между 165 и 175 см».

Вариант 2. Математическая формулировка. Среднее значение μ генеральной совокупности находится в интервале от 165 до 175 с доверительной вероятностью 0,95.

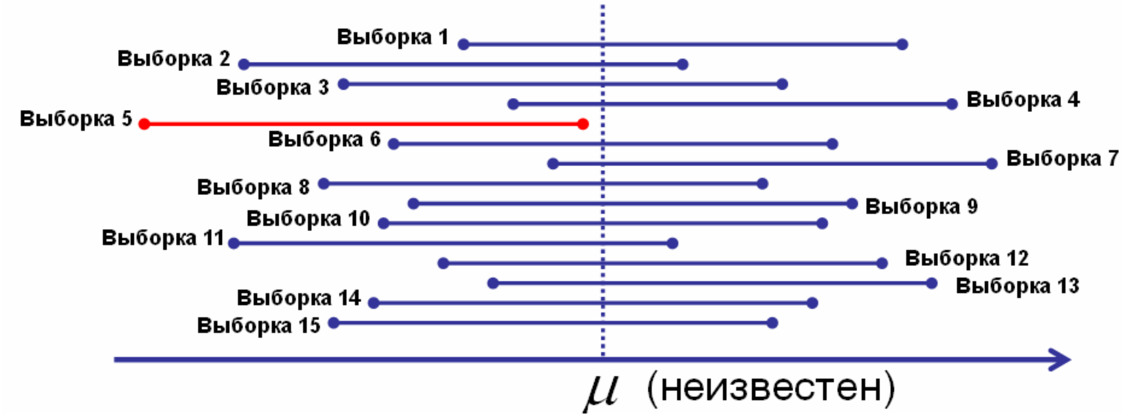


Рисунок 9-4. Доверительный интервал, построенный по пятой выборке, не захватил оцениваемый параметр

Вариант 3. Запись формулой. При помощи формулы доверительный интервал запишется следующим образом:

$$P(165 < \mu < 175) = 0{,}95$$

Доверительный интервал зависит от выборки. Поскольку выборка случайна, для каждой выборки мы будем получать, вообще говоря, свой доверительный интервал. Для доверительной вероятности 95% доверительный интервал будет покрывать неизвестный параметр в 95 случаях из 100.

На рисунке 9-4 показаны доверительные интервалы, построенные для 15 различных выборок. Лишь для пятой выборки оцениваемый параметр не находится внутри построенного доверительного интервала. Это означает, что в приведенном примере только в одном случае из пятнадцати доверительный интервал не содержит неизвестного параметра.

9-2 Доверительный интервал для среднего

Построим доверительный интервал для среднего генеральной совокупности, имеющей нормальный закон распределения.

Предположим, у нас имеется простая случайная выборка из этой генеральной совокупности. Объем выборки равен n. Требуется построить доверительный интервал, который с доверительной вероятностью будет содержать среднее генеральной совокупности:

$$\bar{x} - E < \mu < \bar{x} + E$$

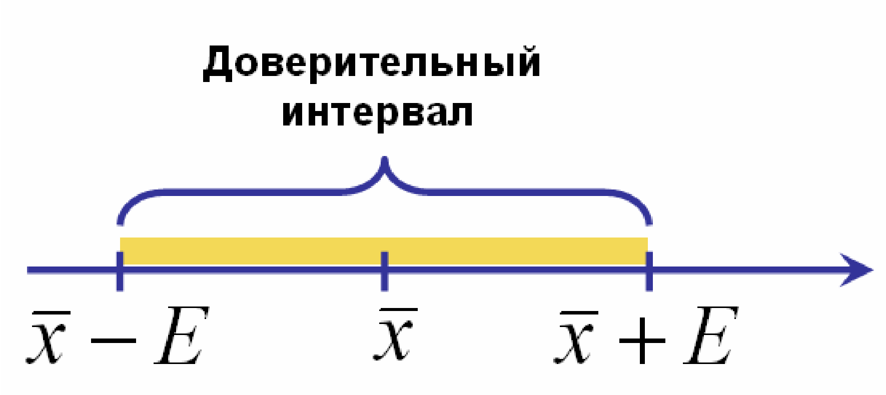


Рисунок 9-5. Доверительный интервал для среднего

По выборке мы можем вычислить выборочное среднее. Это есть точечная оценка для среднего генеральной совокупности. Наша задача сводится к вычислению **точности интервальной оценки** E, что позволит вычислить границы доверительного интервала. При построении доверительного интервала мы будем основываться на известных нам свойствах нормального закона распределения.

Существует два различных случая, которые мы последовательно рассмотрим.

Первый случай: σ известно или n ≥ 30

Предположим, что стандартное отклонение σ генеральной совокупности нам **известно** или **объем выборки n ≥ 30**.

Тогда среднее генеральной совокупности, имеющей нормальный закон распределения, с доверительной вероятностью 1-α находится в доверительном интервале:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Точность интервальной оценки находится по формуле:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Последовательность действий для нахождения доверительного интервала следующая.

ШАГ 1. По выборке вычислить выборочное среднее.

ШАГ 2. По таблице нормального закона найти z-значение для доверительной вероятности 1 - α.

ШАГ 3. Вычислить точность интервальной оценки по формуле:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Если значение σ неизвестно в случае $n \geq 30$, тогда вместо σ в формулу подставляется ее выборочная оценка s.

ШАГ 4. Подставить полученные значения в формулу для доверительного интервала:

$$\bar{x} - E < \mu < \bar{x} + E$$

ШАГ 5. Написать ответ.

Применение таблиц для нахождения z-значений нормального распределения мы уже изучали в главе 7. Приведем z-значения для часто используемых доверительных вероятностей:

Z-значение	Площадь	Доверительная вероятность
1,645	0,9500	0,90 или 90%
1,96	0,9750	0,95 или 95%
2,575	0,9950	0,99 или 99%

Формулы для доверительных интервалов в первом, втором и третьем случаях после подстановки z-значений запишутся в следующем виде:

$$P\left(\bar{x} - 1{,}65 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}65 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9$$

$$P\left(\bar{x} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(\bar{x} - 2{,}58 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2{,}58 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}99$$

Пример. Средний возраст студентов. Ректор университета хочет узнать, каков средний возраст обучающихся студентов. Из предыдущих исследований известно, что стандартное отклонение равно 2 года. Сделана выборка из 50 студентов и вычислено выборочное среднее - 20,3 года. Требуется построить 95%-ый доверительный интервал для генерального среднего.

ШАГ 1. По выборке вычислено выборочное среднее 20,3.

ШАГ 2. Доверительная вероятность 95% соответствует z-значению 1,96.

ШАГ 3. Вычислим точность интервальной оценки:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{2}{\sqrt{50}} = 0{,}55$$

ШАГ 4. Подставим полученные значения в формулу для доверительного интервала:

$$20{,}3 - 0{,}55 < \mu < 20{,}3 + 0{,}55$$

ШАГ 5. Запишем ответ. Средний возраст студентов университета с вероятностью 0,95 находится в интервале между 19,75 и 20,85:

$$19{,}75 < \mu < 20{,}85$$

Можно заметить, что при одном и том же объеме выборки при увеличении доверительной вероятности уменьшается точность интервальной оценки и наоборот. Кроме этого, при постоянной вероятности увеличение объема выборки n приводит к увеличению точности. Имеющиеся формулы позволяют определить минимальный объем выборки, требуемый для получения интервальной оценки с заданной доверительной вероятностью и попадающей в интервал заданного размера. **Объем выборки** определяется по формуле:

$$n = \left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2$$

Пример. Опять про средний возраст. Декан просит преподавателя по статистике оценить средний возраст студентов факультета. Какой размер выборки необходим в этом случае? Преподаватель статистики считает, что оценка должна быть сделана с точностью до 1 года и с вероятностью 99%. Из ранее проведенного исследования известно, что стандартное отклонение возраста – 2 года.

Решение. Для $\alpha = 1 - 0{,}99 = 0{,}01$ z-значение равно 2,58. $E = 1$, $\sigma = 2$. Подставим в формулу:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{2{,}58 \cdot 2}{1}\right)^2 = 26{,}63 \approx 27$$

Тем самым, чтобы быть на 99% уверенным, что полученная оценка отличается от точного значения среднего возраста не больше чем на 1 год, преподавателю нужна выборка как минимум в 27 человек.

Пример. IQ профессоров статистики. Предположим, нам захотелось оценить средний уровень IQ профессоров статистики. Сколько профессоров нам следует случайным образом протестировать, если мы хотим с надежностью 0,95 получить оценку среднего значения IQ, ошибившись при этом не более, чем на 2 балла.

Решение. Общеизвестно, что распределение IQ является нормальным и его стандартное отклонение обычно $\sigma = 15$. Для доверительной вероятности 0,95 z-значение равно 1,96. Точность оценки при данном условии равна:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1{,}96 \cdot 15}{2}\right)^2 = 216{,}09 \approx 217$$

Получили, что объем выборки для получения требуемой точности должен составить не менее 217 профессоров.

Мы не сможем воспользоваться приведенными выше формулами в случае, если стандартное отклонение генеральной совокупности неизвестно, или объем выборки мал ($n \le 30$). Обе ситуации вполне типичны и мы рассмотрим второй случай.

Второй случай: σ неизвестно и n≤30

Строим доверительный интервал для среднего генеральной совокупности, имеющей нормальный закон распределения в следующем виде:

$$\bar{x} - E < \mu < \bar{x} + E$$

У нас имеется простая случайная выборка объема n из этой генеральной совокупности. Дисперсия не известна и объем выборки небольшой ($n \leq 30$), что не позволяет воспользоваться приведенными выше формулами.

При построении доверительного интервала вместо нормального распределения теперь будем использовать t-распределение. Оно было введено в 1908 году В.С.Госсетом, ирландским служащим пивоваренного завода, который участвовал в разработке новых технологий производства пива. Поскольку самостоятельно публиковать результаты исследований работникам завода не разрешалось, Госсет напечатал свои материалы под псевдонимом Стьюдент, поэтому t-распределение часто называют распределением Стьюдента.

Распределение Стьюдента похоже на стандартное нормальное распределение, поскольку имеет колоколообразную форму, симметрично относительно среднего, кривая не соприкасается с осью X. Отличается от стандартного нормального распределения тем, что дисперсия t-распределения больше 1, распределение представляет собой семейство кривых, различающихся числом степеней свободы. С увеличением объема выборки распределение приближается к нормальному. Для нахождения t-значений будем использовать таблицы A-3.

Число степеней свободы (degrees of freedom) – это количество значений, которые могут свободно изменяться после того, как по выборке было вычислено значение статистики.

Например, если среднее для выборки из пяти значений равно 10, тогда четыре из пяти значений могут изменяться. Выберем четыре значения, тогда пятое будет точно определено, поскольку сумма пяти есть 50. Число степеней свободы: $df = 5 - 1 = 4$. Число степеней свободы t-распределения при построении доверительного интервала для среднего равно: $df = n - 1$.

Среднее генеральной совокупности, имеющей нормальный закон распределения с доверительной вероятностью 1-α находится в доверительном интервале:

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Точность интервальной оценки находится по формуле:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Последовательность действий по построению доверительного интервала следующая.

ШАГ 1. По выборке вычислить выборочное среднее и стандартное отклонение.

ШАГ 2. По таблице А-3 найти t-значение для доверительной вероятности 1 - α и числа степеней свободы df = n - 1.

ШАГ 3. Вычислить точность интервальной оценки по формуле:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

ШАГ 4. Подставить полученные значения в формулу для доверительного интервала:

$$\bar{x} - E < \mu < \bar{x} + E$$

ШАГ 5. Написать ответ.

Пример. Как бьется сердце на экзамене. У 20 студентов, сдававших государственный экзамен, сердце билось в среднем со скоростью 96 ударов в минуту. Стандартное отклонение выборки было равно 5 ударам в минуту. Найти 95%-ый доверительный интервал для генерального среднего.

Таблица А-3. Распределение Стьюдента

Df	Односторонняя область				
	0,005	0,01	0,025	0,05	0,1
	Двусторонняя область				
	0,01	0,02	0,05	0,1	0,2
1	63,656	31,821	12,706	6,314	3,078
2	9,925	6,965	4,303	2,920	1,886
3	5,841	4,541	3,182	2,353	1,638
4	4,604	3,747	2,776	2,132	1,533
5	4,032	3,365	2,571	2,015	1,476
6	3,707	3,143	2,447	1,943	1,440
7	3,499	2,998	2,365	1,895	1,415
8	3,355	2,896	2,306	1,860	1,397
9	3,250	2,821	2,262	1,833	1,383
10	3,169	2,764	2,228	1,812	1,372
11	3,106	2,718	2,201	1,796	1,363
12	3,055	2,681	2,179	1,782	1,356
13	3,012	2,650	2,160	1,771	1,350
14	2,977	2,624	2,145	1,761	1,345
15	2,947	2,602	2,131	1,753	1,341
16	2,921	2,583	2,120	1,746	1,337
17	2,898	2,567	2,110	1,740	1,333
18	2,878	2,552	2,101	1,734	1,330
19	2,861	2,539	2,093	1,729	1,328
20	2,845	2,528	2,086	1,725	1,325
21	2,831	2,518	2,080	1,721	1,323

Рисунок 9-6. Использование таблиц t-распределения

ШАГ 1. По выборке вычислено выборочное среднее 96 и стандартное отклонение 5.

ШАГ 2. Доверительная вероятность 95% и количество степеней свободы df = 20 – 1 = 19 соответствуют t-значению 2,093. В заголовке таблицы А-3 пользуемся значениями для двусторонней области (рисунок 9-6).

ШАГ 3. Вычисляем точность интервальной оценки:

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2{,}093\frac{5}{\sqrt{20}} = 2{,}34$$

ШАГ 4. Подставим полученные значения в формулу для доверительного интервала:

$$96 - 2{,}34 < \mu < 96 + 2{,}34$$

ШАГ 5. Запишем ответ. Среднее число ударов сердца у студентов, сдававших государственный экзамен, с доверительной вероятностью 95% находится в пределах (ударов в минуту):

$$93{,}66 < \mu < 98{,}34$$

Резюме

Мы рассмотрели построение доверительных интервалов для среднего генеральной совокупности, имеющей нормальный закон распределения. Его точечной оценкой является выборочное среднее. Интервальную оценку строим, находя значение E, называемое точностью интервальной оценки. В случае, если выборка достаточно велика или нам известно стандартное отклонение генеральной совокупности, мы основываемся на свойствах нормального закона и вычисляем точность, используя таблицу А-2. Если объем выборки небольшой $n \leq 30$ и стандартное отклонение генеральной совокупности нам не известно, построение доверительного интервала происходит при помощи t-распределения, и в этом случае мы пользуемся таблицей А-3.

Формулы позволяют нам также оценивать минимальный объем выборки, требуемый для получения интервальной оценки с заданной точностью и заданной доверительной вероятностью.

9-3 Доверительный интервал для доли

В этом параграфе мы построим доверительный интервал для доли признака генеральной совокупности. Точечной оценкой является доля признака в выборке.

Построение доверительного интервала

Если у нас имеется случайная выборка объема n из генеральной совокупности, то найдя выборочную долю, будем искать интервальную оценку в следующем виде:

$$\hat{p} - E < p < \hat{p} + E$$

Задача состоит в нахождении точности E и вычислении границ доверительного интервала. Из свойств биномиального распределения следует, что доверительный интервал для доли признака вычисляется по следующей формуле:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

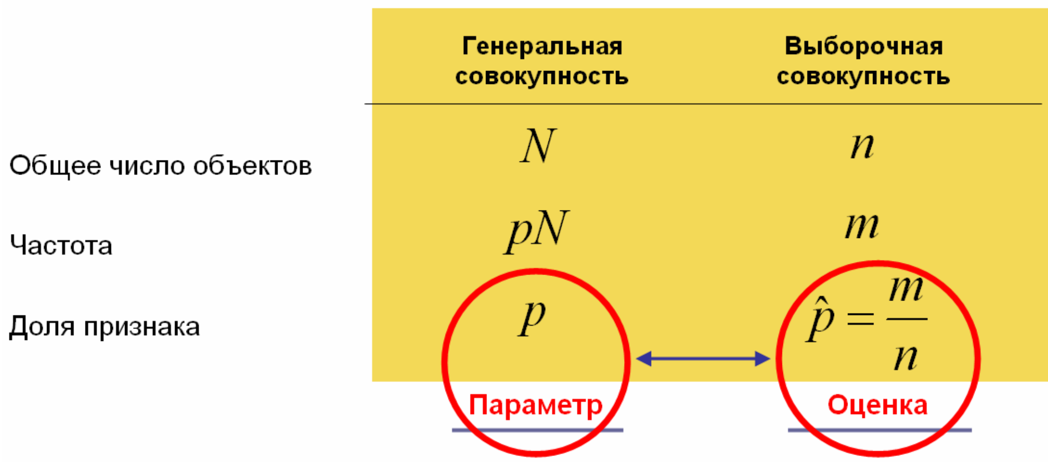


Рисунок 9-7. Оценка доли признака генеральной совокупности

Потребуем соблюдение необходимых условий: $np \geq 5$ и $nq \geq 5$. Последовательность действий по построению доверительного интервала включает пять шагов.

ШАГ 1. По выборке вычислить долю признака.

ШАГ 2. По таблице А-2 найти z-значение для доверительной вероятности $1 - \alpha$.

ШАГ 3. Вычислить точность интервальной оценки по формуле:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

ШАГ 4. Подставить полученные значения в формулу для доверительного интервала:

$$\hat{p} - E < p < \hat{p} + E$$

ШАГ 5. Написать ответ.

Пример. Поддержка для мэра. В ходе проведенного опроса 829 жителей города выяснилось, что 417 опрошенных (51,5%) предполагают поддержать на предстоящих выборах кандидатуру действующего мэра. Можно ли на этом основании утверждать, что более половины жителей города поддерживают перевыборы действующего мэра на следующий срок?

Решение. Построим интервальную оценку для доли признака в генеральной совокупности и затем ответим на поставленный вопрос.

ШАГ 1. По условию доля признака в выборке составила 0,515.

ШАГ 2. Для доверительной вероятности 95% z-значение равно 1,96.

ШАГ 3. Вычислим точность интервальной оценки:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1{,}96\sqrt{\frac{0{,}515\cdot 0{,}485}{829}} = 0{,}034$$

ШАГ 4. Подставим полученные значения в формулу для доверительного интервала:

$$0{,}515 - 0{,}034 < p < 0{,}515 + 0{,}034$$

ШАГ 5. Напишем ответ:

$$48{,}1\% < p < 54{,}9\%$$

Полученный результат можно интерпретировать следующим образом. Доля признака генеральной совокупности с вероятностью 95% находится в пределах между 48,1% и 54,9% голосов. Это означает, что несправедливо по результатам опроса утверждать, что более половины избирателей будут голосовать за выборы действующего мера на следующий срок.

Нахождение объема выборки

Длина доверительного интервала для доли зависит от объема выборки и доверительной вероятности. На эту связь указывает формула точности интервальной оценки E. Мы можем ответить на вопрос о том, какой минимальный объем должна иметь выборка, чтобы с заданной доверительной вероятностью оцениваемый параметр оказался в доверительном интервале заданной длины. **Объем выборки** находится по формуле:

$$n = \hat{p}\hat{q}\cdot\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Значение объема необходимо округлить вверх, чтобы получить целое число в качестве минимального объема выборки. Формула предполагает, что у нас уже имеется оценка для доли признака, полученная из предшествующих исследований. Если все же никакой оценки мы еще не имеем, то минимальный объем выборки вычисляется по формуле:

$$n = 0{,}25 \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2$$

Пример. Домашний компьютер. Исследователь хочет с 95%-ой вероятностью оценить количество студентов, у которых дома имеется персональный компьютер. По данным предыдущего исследования их оказалось 40%. Исследователь не хочет ошибиться больше, чем на 2%. Требуется определить минимальный объем выборки для проведения исследования.

Решение. Для доверительной вероятности 95% z-значение равно 1,96. По условию задана точность интервальной оценки E = 0,02.

Подставляем в формулу:

$$n = \hat{p}\hat{q} \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2 = 0{,}40 \cdot 0{,}60 \cdot \left(\frac{1{,}96}{0{,}02} \right)^2 = 2304{,}96$$

Получили, что требуемый минимальный объем выборки 2 305 студентов. Объем получился большим, поскольку исследователь захотел оценить долю довольно точно – в пределах ±2%.

Резюме

В этом параграфе мы построили доверительный интервал для доли признака в генеральной совокупности. Использовали при этом свойства биномиального распределения и его нормальное приближение. Связь между доверительной вероятностью, точностью оценки и объемом выборки позволяет находить минимальный объем.

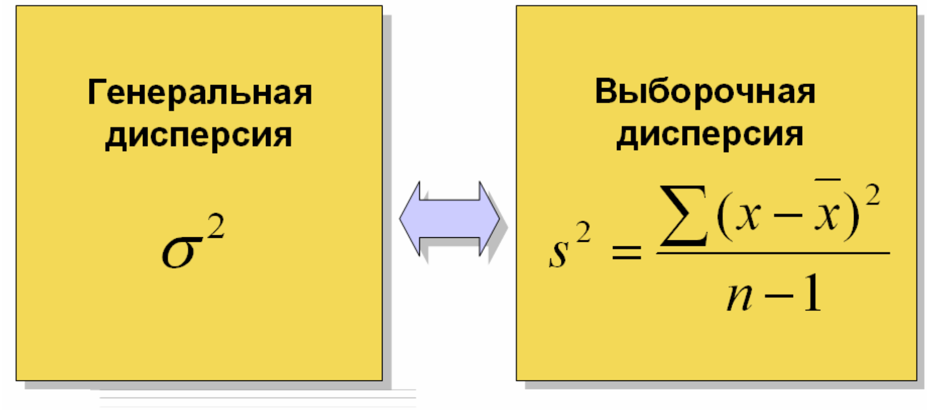


Рисунок 9-8. Оценка дисперсии генеральной совокупности

9-4 Доверительный интервал для дисперсии

Построим доверительный интервал для неизвестной дисперсии нормально распределенной генеральной совокупности. Оценкой для генеральной дисперсии является выборочная дисперсия. Доверительный интервал находится по следующей формуле:

$$\frac{(n-1)\cdot s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi_L^2}$$

Значения χ^2_L и χ^2_R находятся по таблицам хи-квадрат распределения, исходя из следующих условий:

$$P(\chi^2 > \chi_L) = 1 - \frac{\alpha}{2}$$

$$P(\chi^2 > \chi_R) = \frac{\alpha}{2}$$

Доверительный интервал для стандартного отклонения имеет вид:

$$\frac{\sqrt{n-1}\cdot s}{\chi_R} < \sigma < \frac{\sqrt{n-1}\cdot s}{\chi_L}$$

Последовательность действий для построения доверительного интервала для дисперсии следующая.

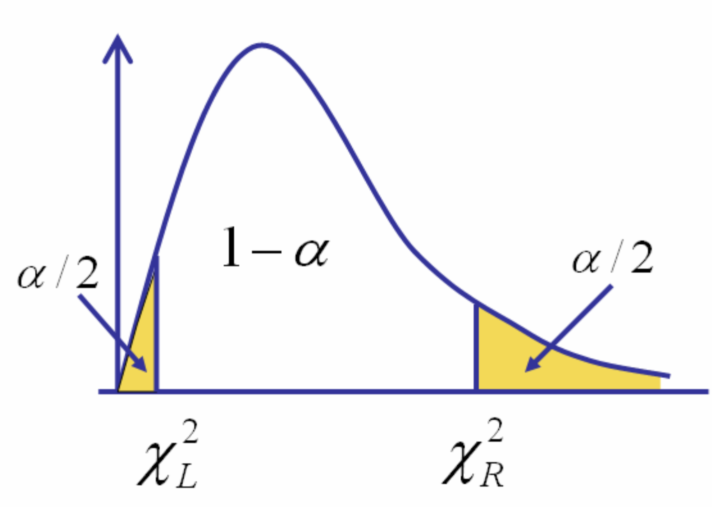


Рисунок 9-9. Нахождение левого и правого хи-квадрат значений

ШАГ 1. По выборке вычислить дисперсию.

ШАГ 2. По таблице А-4 найти два хи-квадрат значения для доверительной вероятности 1 - α с числом степеней свободы df = n - 1:

$$P(\chi^2 > \chi_L) = 1 - \frac{\alpha}{2}$$

$$P(\chi^2 > \chi_R) = \frac{\alpha}{2}$$

ШАГ 3. Подставить полученные значения в формулу:

$$\frac{(n-1)\cdot s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi_L^2}$$

ШАГ 4. Написать ответ.

Индексы R и L возле χ^2-значений являются первыми буквами слов «Left» и «Right» - левое и правое значение, соответственно. Эти значения являются в некотором смысле симметричными, поскольку «отрезают» от распределения χ^2 одинаковые по площади части (рисунок 9-9).

Пример. Интервальная оценка дисперсии. Требуется оценить дисперсию для нормально распределенной генеральной совокупности, оценка которой по выборке объема 10 оказалась равна 28,2. Следуем предложенной последовательности действий. Доверительную вероятность выберем на уровне 90%.

ШАГ 1. По выборке объема 10 вычислена дисперсия, которая равна по условию 28,2.

ШАГ 2. Поскольку доверительная вероятность равна 90%, тогда $\alpha = 1 - 0{,}9 = 0{,}1$. Число степеней свободы df = n – 1 = 9. По таблице А-4 находим два χ^2-значения для 0,95 и 0,05. Эти значения равны $\chi^2_L = 3{,}325$ и $\chi^2_R = 16{,}919$ соответственно.

$$P(\chi^2 > \chi_L) = 1 - \frac{\alpha}{2} = 0{,}95$$

$$P(\chi^2 > \chi_R) = \frac{\alpha}{2} = 0{,}05$$

ШАГ 3. Подставим полученные значения в формулу:

$$\frac{(n-1)\cdot s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi_L^2}$$

$$\frac{(10-1)\cdot 28{,}2}{16{,}919} < \sigma^2 < \frac{(10-1)\cdot 28{,}2}{3{,}325}$$

ШАГ 4. Напишем ответ:

$$15{,}0 < \sigma^2 < 76{,}3$$

Получили, что дисперсия генеральной совокупности находится в интервале от 15,0 до 76,3. Если нас интересует доверительный интервал для стандартного отклонения, то нужно взять корень от обеих частей неравенства:

$$\sqrt{15{,}0} < \sigma < \sqrt{76{,}3}$$

Это означает, что стандартное отклонение генеральной совокупности с доверительной вероятностью 90% находится в следующих границах:

$$3{,}87 < \sigma < 8{,}73$$

Используем компьютер

При построении доверительных интервалов важнейшим является вычисление характеристик выборки: среднего, доли, дисперсии, а также нахождение z-значения, t-значения и χ^2-значений, соответственно. При использовании электронных таблиц указанные значения вычисляются при помощи функций: НОРМСТОБР( ), СТЬЮДРАСПОБР( ), ХИ2ОБР( ). Обработка выборок в SPSS дает возможность получить 95% доверительный интервал для среднего. Желательно сравнить ручные вычисления с результатами, полученными при помощи компьютера. При наличии расхождений – объяснить или найти ошибку.

Что означают термины

Точечная оценка	**Несмещенность**	**Доверительный интервал**
Интервальная оценка	**Эффективность**	**Доверительная вероятность**
Ошибка оценки	**Состоятельность**	**Точность интервальной оценки**

Символы и формулы

$\bar{x} - E < \mu < \bar{x} + E$	Доверительный интервал для среднего (общий вид)
E	Точность доверительного интервала
$\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	Доверительный интервал для среднего (дисперсия известна или объем выборки большой)

$n = \left(\dfrac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$	Объем выборки для нахождения интервальной оценки среднего
$\bar{x} - t_{\alpha/2} \dfrac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \dfrac{s}{\sqrt{n}}$	Доверительный интервал для среднего (дисперсия неизвестна и объем выборки небольшой)
$\hat{p} - z_{\alpha/2} \sqrt{\dfrac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$	Доверительный интервал для доли признака
$n = \hat{p}\hat{q} \cdot \left(\dfrac{z_{\alpha/2}}{E} \right)^2$	Объем выборки для нахождения интервальной оценки доли
$\dfrac{(n-1) \cdot s^2}{\chi_R^2} < \sigma^2 < \dfrac{(n-1) \cdot s^2}{\chi_L^2}$	Доверительный интервал для дисперсии
$P(\chi^2 > \chi_L) = 1 - \dfrac{\alpha}{2}$, $P(\chi^2 > \chi_R) = \dfrac{\alpha}{2}$	Условия для нахождения $\chi 2$-значений
$\dfrac{\sqrt{n-1} \cdot s}{\chi_R} < \sigma < \dfrac{\sqrt{n-1} \cdot s}{\chi_L}$	Доверительный интервал для стандартного отклонения

Задачи и упражнения

9-1. Доход студентов. Из предыдущих исследований известно, что месячный доход студентов университета имеет нормальное распределение со стандартным отклонением 60. Опрошено случайным образом 225 человек. Их средний доход составил 310. Найти 95%-ый доверительный интервал для среднего месячного дохода всех студентов университета.

9-2. Пробег колес. Для случайно отобранных 100 шин фирмы ABC средний пробег составил 40 000 км при стандартном отклонении 8 000 км. Найти 99%-ый доверительный интервал для генерального среднего.

9-3. Возраст студентов. Для случайно отобранных 16 студентов университета средний возраст составил 23 года. Найти 95%-ый доверительный интервал для генерального среднего, если известно, что:

а) распределение возрастов всех студентов нормальное со стандартным отклонением 0,6.

б) распределение возрастов всех студентов нормальное, но стандартное отклонение генеральной совокупности неизвестно, а выборочное стандартное отклонение равно 0,6.

9-4. Быть ли статистике? Выборочный опрос 75 студентов-первокурсников показал, что 15 из них высказываются за исключение курса статистики из учебной программы. Найти 90%-ый доверительный интервал для фактической доли студентов-первокурсников, поддерживающих исключение статистики из программы.

9-5. Возраст медсестер. Средний возраст 12-ти медсестер в крупной городской больнице оказался равен 26,8 года. Стандартное отклонение выборки 4,8 года. Найдите 95%-ый доверительный интервал для среднего возраста генеральной совокупности, состоящей из всех медсестер этой больницы.

9-6. Рождение детей. В больнице в течение 10 отобранных случайным образом недель проводилось исследование. Было установлено, что в среднем за неделю рождается 12 детей. Выборочное стандартное отклонение равно 2. Найдите 99%-ый доверительный интервал для фактического среднего.

9-7. Собаки и почтальоны. В одной городской местности для исследования было случайно отобрано 5 месяцев. Оказалось, что в среднем в каждый из них собаки кусают 28 почтальонов. Стандартное отклонение по выборке равно 3. Найдите 90%-ый доверительный интервал для среднего числа почтальонов, ежемесячно страдающего от укусов собак.

9-8. Почтовые расходы. Исследователь хочет определить с точностью до 25$ среднюю сумму почтовых расходов компании. Каков должен быть объем выборки, если хочется иметь 90%-ую гарантию правильности результатов. Стандартное отклонение равно 80$.

9-9. Студенческий городок. Недавнее исследование, проведенное среди 150 студентов, выявило, что 86 из них проживают за пределами студенческого городка. Найдите 95%-ый доверительный интервал для фактической доли студентов, которые живут не в студенческом городке.

9-10. Статистика происшествий. Исследование 200 несчастных случаев, при которых требовалась срочная медицинская помощь, выявило, что 40% из них произошло с людьми у них дома. Найдите 90%-ый доверительный интервал для действительной доли несчастных случаев, которые случаются дома.

9-11. Ограничение власти правительства. Политический аналитик выявил, что 60% из 300 человек, голосующих за демократов, считают, что у федерального правительства слишком много власти. Найдите 95%-ый доверительный интервал генеральной доли тех, кто голосует за демократов и придерживается этого мнения.

9-12. Исследование диетолога. Диетолог хочет определить с максимальной ошибкой в 2% долю людей, которые едят перед сном. Каков должен быть размер выборки, если он хочет быть на 95% уверен в том, что его оценка содержит значение генеральной доли? Предыдущее исследование выявило, что 18% из 100 опрошенных сказали, что они едят перед сном.

9-13. Теннис и футбол. Исследование показало, что из 200 опрошенных 15% процентов регулярно играют в теннис или футбол. Каков должен быть объем выборки, если исследователь хочет найти 99%-ый доверительный интервал для действительной доли взрослых, которые играют в теннис или футбол, и при этом не отклониться от генеральной доли более чем на 1%?

Глава 10. Проверка статистических гипотез

В этой главе мы рассмотрим общие принципы, которые важны для проверки статистических гипотез, а затем научимся проверять три типа гипотез – о среднем, о доли признака и о дисперсии генеральной совокупности. Все эти гипотезы связаны с оценкой параметров одной генеральной совокупности. В следующей главе мы будем рассматривать и сравнивать две генеральные совокупности и их числовые характеристики.

10-1 Общие принципы проверки гипотез

Перед тем, как рассматривать и проверять конкретные гипотезы, нам необходимо привести основные понятия, а также обсудить общие принципы, которые применяются при проверке статистических гипотез.

Статистической гипотезой (statistical hypothesis) мы называем любое предположение о свойствах и характеристиках исследуемых генеральных совокупностей, которое может быть проверено на основе анализа выборок.

Например, если мы анализируем возраст, пол и семейное положение респондентов, то гипотезой может стать предположение о среднем возрасте неженатых мужчин. Эта гипотеза подлежит проверке, в ходе которой нам предстоит сделать один из следующих выводов:

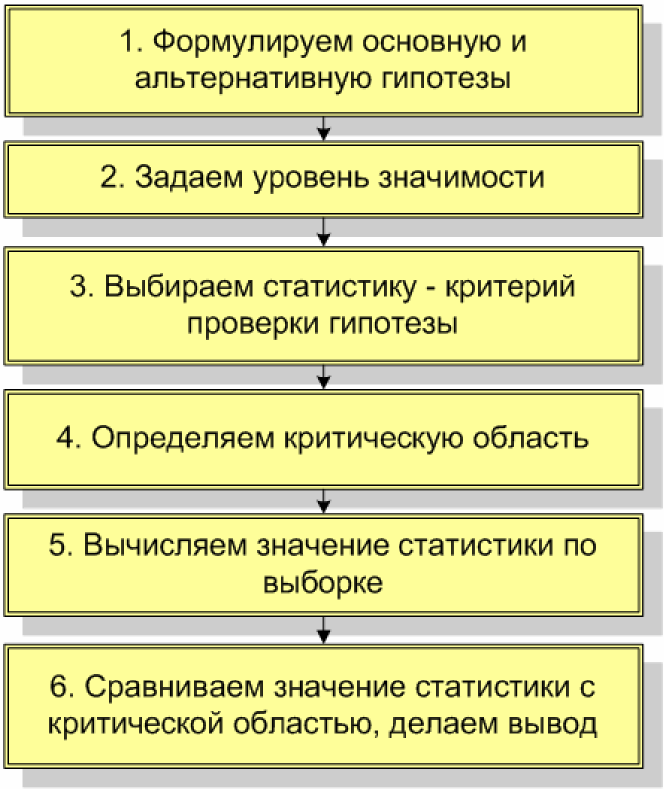


Рисунок 10-1.Этапы проверки статистических гипотез

1) рассматриваемая нами гипотеза не верна и нам следует ее отклонить;

2) мы можем принять рассматриваемую гипотезу, поскольку у нас нет причин ее отклонить.

Гипотезы называют **параметрическими**, если в них делаются предположения относительно значений параметров исследуемого распределения. Если, напротив, исследователь не делает предположений о виде распределения и значениях его параметров, его гипотеза является **непараметрической**. Примером может быть проверка гипотезы о том, что две выборки взяты из одной и той же генеральной совокупности без предположений о конкретном виде распределения.

В этой и следующей главе мы будем рассматривать параметрические гипотезы. Непараметрическим критериям мы посвятим главы 15 – 17 в заключительной части курса.

Основная и альтернативная гипотезы

Проверяемая гипотеза в статистике называется **основной** или **нулевой гипотезой.** Она обычно рассматривается вместе с другой гипотезой, которая называется **альтернативной**. Основная гипотеза обозначается H_0, альтернативная – H_1. В результате проверки имеется две возможности:

1. Принять основную гипотезу H_0, отклонив при этом альтернативную гипотезу H_1
2. Отклонить основную гипотезу H_0 и принять альтернативную гипотезу H_1

В качестве иллюстрации приведем три разных примера исследований.

Ситуация А. Новая методика преподавания. Исследователь хочет проверить, повлияет ли новая методика преподавания на уровень успеваемости студентов. Повысится или понизится успеваемость у студентов, прослушавших курс по новой методике? Исследователю известно, что средняя успеваемость без нововведений составляет 4,23 балла. Гипотезы в этом случае будут сформулированы следующим образом:

$$H_0: \mu = 4{,}23$$
$$H_1: \mu \neq 4{,}23$$

Нулевая гипотеза говорит о том, что среднее останется тем же самым, а альтернативная – о том, что оно изменится. Что будет означать принятие основной гипотезы? Оно будет означать, что в результате проверки (проведения выборочного эксперимента) у нас не появилось оснований думать, что успеваемость существенно отличается от прежней. В случае принятия альтернативной гипотезы мы скажем, что успеваемость отличается и это отличие значимо.

Рассматриваемая гипотеза является двусторонней, поскольку нам не известно заранее, в какую сторону может измениться успеваемость студентов после внедрения методики. Мы понимаем, что она может как повыситься, так и снизиться, что будет в любом случае означать изменение успеваемости.

Ситуация Б. Аккумуляторные батареи для ноутбуков. Производители аккумуляторных батарей для ноутбуков утверждают, что разработали принципиально новый тип батареи, которая существенно дольше может работать без подзарядки. Из предыдущих исследований известно, что среднее время работы существующих аккумуляторов составляет 2,5 часа, после чего их требуется заряжать. Гипотезы принимают следующий вид:

$$H_0: \mu \leq 2{,}5$$
$$H_1: \mu > 2{,}5$$

В данном случае мы заинтересованы только в увеличении продолжительности работы батареи. Принятие альтернативной гипотезы по результатам эксперимента будет означать, что время службы батареи без подзарядки действительно увеличилось и значимо превышает 2,5 часа.

Нулевая гипотеза при этом представляет собой утверждение, что время работы батареи осталось на прежнем уровне или даже уменьшилось. Этот критерий называется односторонним, поскольку в наши интересы входит только увеличение среднего.

Ситуация В. Расходы на канцелярию. Менеджер бюро переводов хочет снизить расходы компании на канцелярские принадлежности. В среднем эти расходы составляют 5 300 рублей в неделю. После принятия определенных мер по экономии бумаги и скрепок менеджер хотел бы проверить, снизились ли расходы или остались на прежнем уровне. Гипотезы о величине расходов выглядят следующим образом:

$$H_0: \mu \geq 5300$$
$$H_1: \mu < 5300$$

Это односторонний критерий, так как менеджер заинтересован исключительно в понижении расходов на канцелярию. Итак, имеется три вида критериев:

Двусторонний	Левосторонний	Правосторонний
H0: =	H0: $\geq$	H0: $\leq$
H1: $\neq$	H1: $<$	H1: $>$

Рассмотренные ситуации показывают, что выбор гипотез зависит от задачи, которую решает исследователь. В реальных исследованиях чтобы получить результаты, которые окажутся важными и полезными, приходится не один раз формулировать и проверять различные гипотезы. Это выглядит как сложный путь проб и ошибок. Исследователь напоминает в этот момент скульптора, который отсекает от камня лишнее, оставляя только безукоризненные формы своего будущего творения.

Чтобы быть корректными, статистики в случае принятия основной гипотезы не утверждают, что она верна, а говорят, что в результате эксперимента не появилось оснований ее отклонить. При принятии альтернативной гипотезы статистики с уверенностью отклоняют основную гипотезу раз и навсегда, окончательно и бесповоротно. В этом случае статистическими методами нельзя доказать какое-либо утверждение, можно лишь отвергнуть ошибочные предположения.

Ошибки первого и второго рода

Статистические гипотезы проверяются статистическими методами, на основании выборки, полученной из генеральной совокупности. Из-за случайности выборки в результате проверки могут возникать ошибки и приниматься неправильные решения.

Назовем ошибкой первого рода ситуацию, в которой мы отвергаем верную гипотезу H_0. При ошибке второго рода принимается гипотеза H_0 в то время, как она неверна.

Ошибка первого рода (type I error) происходит в случае, когда мы отвергаем нулевую гипотезу, если она верна. **Ошибка второго рода (type II error)** происходит, если мы принимаем нулевую гипотезу, когда она неверна.

В ситуации А исследователь по результатам эксперимента решит, что успеваемость изменилась, отклонит основную гипотезу и совершит, тем самым, ошибку первого рода, если на самом деле успеваемость осталась на прежнем уровне. Ошибку второго рода он совершит, если примет основную гипотезу, решив, что успеваемость не изменилась, в то время, как она в действительности изменилась существенно. В первом случае методика будет внедрена, но будет работать без пользы для студентов. Во втором случае новая методика оказалась бы полезной, но не будет принята в результате ошибочных выводов.

	Основная гипотеза верна	Основная гипотеза неверна
Мы приняли основную гипотезу	Верное решение	Ошибка II рода
Мы отклонили основную гипотезу	Ошибка I рода	Верное решение

Рисунок 10-2. Ошибки I и II рода при проверке гипотез

Уровень значимости

Поскольку существуют ненулевые вероятности совершить ошибки первого или второго рода, следует устанавливать разумные значения этих вероятностей. До начала эксперимента принято задавать значение вероятности ошибки первого рода, называемое уровнем значимости.

Уровнем значимости гипотезы (level of significance) называют вероятность совершить ошибку первого рода, то есть принять гипотезу H_0 в то время, как она неверна.

Будем обозначать уровень значимости символом α (альфа). Как правило, уровень значимости выбирают 0,05 или 0,01. Это означает, что если мы отклоняем нулевую гипотезу, то вероятность совершения ошибки первого рода равна 5% или 1%, а вероятность сделать правильное заключение соответственно 95% или 99%. Другими словами, когда $\alpha = 0{,}05$, существует вероятность 5% отвергнуть истинную нулевую гипотезу, а если $\alpha = 0{,}01$, то существует вероятность 1% соответственно.

Желательно, чтобы вероятность совершения ошибок была минимальна. Тем не менее, уровень значимости не может быть выбран очень малым, поскольку уменьшение вероятности совершить ошибку первого рода вызывает увеличение вероятности совершить ошибку второго рода.

Статистика (критерий)

Для поверки гипотез нам потребуется специальная функция, называемая статистикой или критерием. Эта функция зависит от элементов выборки и является случайной.

Статистика (критерий, statistical test) есть специальная функция от элементов выборки, по значениям которой принимают решение о принятии или отклонении основной гипотезы.

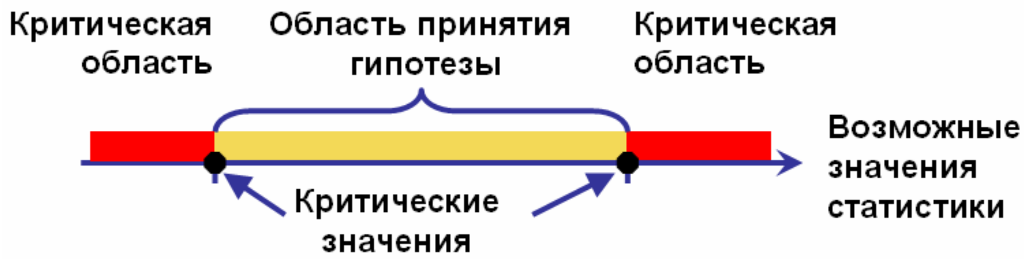


Рисунок 10-3. Критические значения отделяют область принятия гипотезы от критической области

Множество значений статистики включает два непересекающихся подмножества: *область принятия гипотезы*, то есть множество тех значений статистики, при которых гипотеза H_0 принимается, и *критическую область*, то есть множество тех значений статистики, при которых гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1.

Область принятия гипотезы (nonrejection region) есть множество значений статистики, при которых основную гипотезу следует принять.

Критическая область (critical region) есть множество значений статистики, при которых основную гипотезу следует отклонить.

Критические значения (critical value(s)) отделяют критическую область от области принятия гипотезы.

Критическая область

Различают одностороннюю и двустороннюю критическую область. В свою очередь, односторонняя критическая область может быть правосторонней или левосторонней. Вид критической области зависит от вида альтернативной гипотезы. В частности при альтернативной гипотезе H_1: $\mu < \mu_0$ критическая область будет левосторонней, при альтернативной гипотезе H_1: $\mu > \mu_0$ критическая область будет правосторонней. Альтернативная гипотеза H_1: $\mu \neq \mu_0$ соответствует двусторонней критической области.

Критическая область строится, исходя из имеющихся знаний о законе распределения статистики. Критические точки находятся по таблицам. Необходимо при этом учитывать уровень значимости гипотезы, а также количество степеней свободы, зависящее от объема выборки.

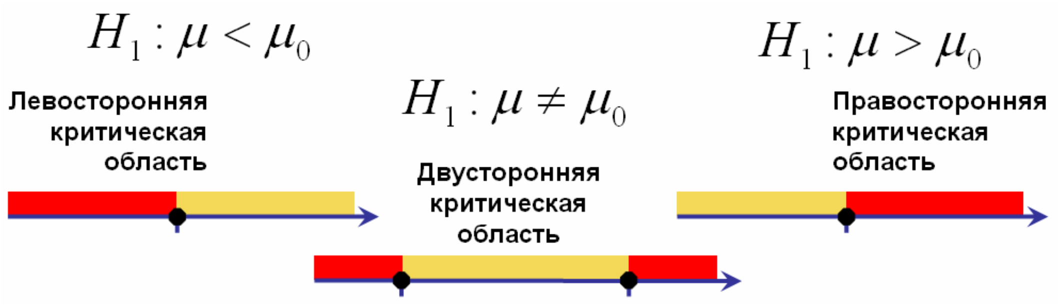


Рисунок 10-4. Критическая область зависит от вида альтернативной гипотезы

После построения критической области значение статистики по выборке сравнивают с критической областью. Если значение статистики попало в область принятия гипотезы, то основная гипотеза принимается. Если значение статистики попало в критическую область, то основная гипотеза отклоняется и принимается альтернативная гипотеза.

Этапы проверки гипотезы

Мы обсудили ключевые понятия и описали последовательность действий по проверке гипотез, теперь запишем основные этапы.

ШАГ 1. Сформулировать основную и альтернативную гипотезы.

ШАГ 2. Задать уровень значимости α.

ШАГ 3. По таблице найти критические значения и построить критическую область.

ШАГ 4. По выборке сосчитать значение статистики.

ШАГ 5. Сравнить полученное значение с критической областью. Если значение попало в критическую область – отклонить основную гипотезу, не попало – принять.

ШАГ 6. Написать ответ.

Как мы увидим, эта последовательность шагов применима ко всем критериям, включая непараметрические. Следует придерживаться этой последовательности даже в очевидных случаях, чтобы не допускать ошибок.

I	II	III
Нулевая гипотеза: $H_0: \mu \geq \mu_0$	Нулевая гипотеза: $H_0: \mu \leq \mu_0$	Нулевая гипотеза: $H_0: \mu = \mu_0$
Альтернативная гипотеза: $H_1: \mu < \mu_0$	Альтернативная гипотеза: $H_1: \mu > \mu_0$	Альтернативная гипотеза: $H_1: \mu \neq \mu_0$

Рисунок 10-5. Три варианта гипотез о среднем

10-2 Гипотеза о среднем

Критерии, которые мы сейчас рассмотрим, позволяют проверять гипотезу о среднем значении генеральной совокупности, имеющей нормальный закон распределения.

Первый случай: σ известно или n ≥ 30

В качестве статистики используется следующая случайная функция:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

где $\bar{x}$ – среднее значение выборки,
μ_0 – гипотетическое генеральное среднее,
σ – стандартное отклонение генеральной совокупности,
n – объем выборки.

Используемая статистика получена путем деления разности между наблюдаемым выборочным средним и ожидаемым генеральным средним на величину стандартной ошибки:

$$\frac{\textit{Наблюдаемое среднее} - \textit{Ожидаемое среднее}}{\textit{Стандартная ошибка}}$$

Эта случайная величина имеет стандартное нормальное распределение. Для нахождения критических значений мы будем использовать таблицы А-2.

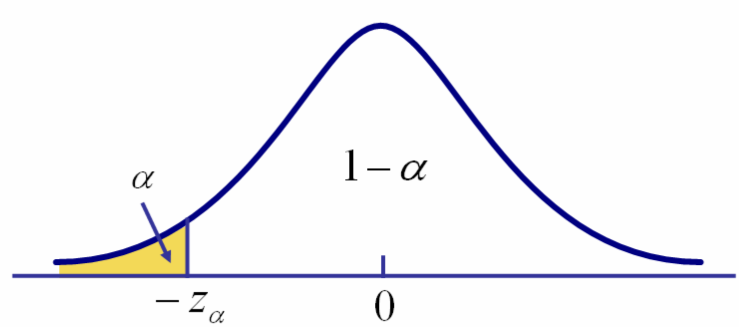


Рисунок 10-6. Левосторонняя критическая область

Построим критическую область для каждого из трех вариантов гипотез.

Вариант I. Левосторонняя критическая область. Гипотезы:

$$H_0: \mu \geq \mu_0$$
$$H_1: \mu < \mu_0$$

Критическая область является левосторонней и определяется уравнением:

$$P(z < -z_\alpha) = \alpha$$

Для нахождения границы пользуемся таблицей А-2. Поскольку таблица составлена только для положительных чисел, воспользуемся свойством функции распределения стандартного нормального закона и получим:

$$\Phi(z_\alpha) = 1 - \alpha$$

Это означает, что внутри таблицы следует отыскать значение вероятности 1 – α, затем по краям таблицы определить z-значение, которому эта вероятность соответствует. Меняем знак z-значения и получаем границу критической области.

Для уровня значимости α = 0,01 находим в таблице 1 – α = 0,99. Этому значению вероятности соответствует z-значение 2,33. Следовательно, критическая область запишется уравнением:

$$P(z < -2{,}33) = 0{,}01$$

Если значение статистики, вычисленное по выборке, окажется меньше значения минус 2,33, основную гипотезу следует отвергнуть на уровне значимости 1%.

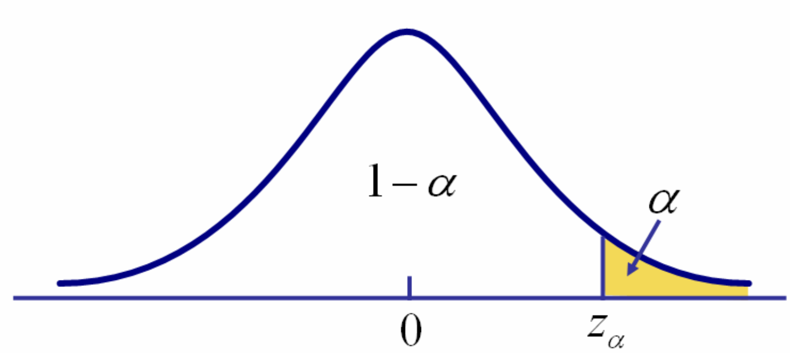


Рисунок 10-7. Правосторонняя критическая область

Вариант II. Правосторонняя критическая область. Гипотезы:

$$H_0: \mu \geq \mu_0$$
$$H_1: \mu < \mu_0$$

Для этого набора критическая область является правосторонней и определяется уравнением:

$$P(z > z_\alpha) = \alpha$$

Для нахождения границы пользуемся таблицей А-2. Будем искать в таблице z-значение, которое соответствует следующей вероятности:

$$\Phi(z_\alpha) = 1 - \alpha$$

Внутри таблицы находим значение вероятности 1 - α, и по краям таблицы определяем z-значение, которому эта вероятность соответствует. Получили границу критической области.

Для уровня значимости α = 0,05 находим в таблице 1 - α = 0,95. Этому значению вероятности соответствует z-значение 1,96. Следовательно, критическая область запишется уравнением:

$$P(z > 1{,}96) = 0{,}05$$

Если значение статистики, вычисленное по выборке, окажется больше значения 1,96, основную гипотезу следует отвергнуть на уровне значимости 5%.

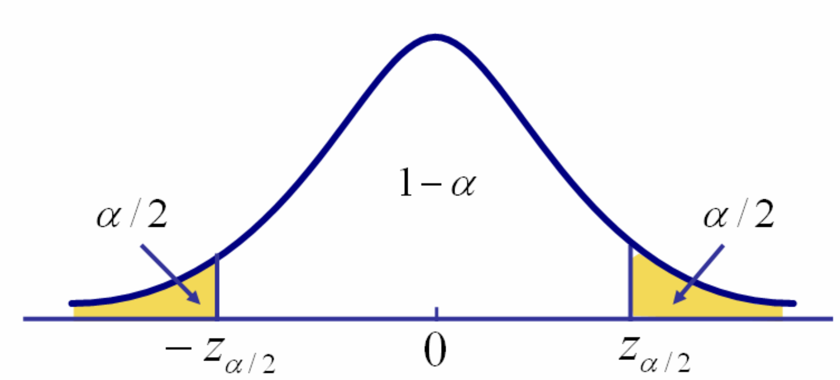


Рисунок 10-8. **Двусторонняя критическая область**

Вариант III. Двусторонняя критическая область. Гипотезы:

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Для этого набора критическая область является двусторонней и определяется двумя уравнениями:

$$P(z > z_{\alpha/2}) = \alpha/2$$
$$P(z < -z_{\alpha/2}) = \alpha/2$$

Для нахождения двух границ пользуемся таблицей А-2. Будем искать в таблице z-значение, которое соответствует вероятности:

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2$$

Это означает, что нам нужно внутри таблицы отыскать значение вероятности 1 - $\alpha/2$, и по краям таблицы определить z-значение, которому эта вероятность соответствует. Полученное z-значение со знаком плюс и со знаком минус дает нам две границы двусторонней критической области.

Для уровня значимости α = 0,05 находим в таблице 1 - $\alpha/2$ = 0,975. Этому значению вероятности соответствует z-значение 1,96. Следовательно, критическая область запишется при помощи уравнений: $P(z > 1{,}96) = 0{,}025$ и $P(z < -1{,}96) = 0{,}025$.

Если значение статистики, вычисленное по выборке, окажется меньше значения минус 1,96 или больше плюс 1,96, основную гипотезу следует отвергнуть на уровне значимости 5%.

Рисунок 10-9. Значение -6,0 попадает в критическую область

Пример. Чем занимаются старшеклассники. В одном из журналов утверждается, что старшеклассники в среднем смотрят телевизор меньше других. Известно, что люди проводят перед телевизором в среднем 29,4 часа в неделю со стандартным отклонением 2 часа. Случайная выборка из 25 старшеклассников имеет среднее 27 часов. Необходимо проверить утверждение на уровне значимости α = 0,01.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

H_0: $\mu \geq 29{,}4$
H_1: $\mu < 29{,}4$

ШАГ 2. Задан уровень значимости α =0,01.

ШАГ 3. Найдем критические значения и построим критическую область. Так как α = 0,01 и критическая область левосторонняя, то критическое значение равно –2,33.

ШАГ 4. По выборке вычисляем значение статистики:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{27 - 29{,}4}{2/\sqrt{25}} = -6{,}0$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики –6,0 попадает в критическую область, мы отвергаем нулевую гипотезу.

ШАГ 6. Формулируем ответ. Поскольку мы отвергли основную гипотезу и приняли альтернативную, делаем вывод, что старшеклассники действительно значимо меньше смотрят телевизор, чем все остальные.

Если дисперсия генеральной совокупности неизвестна и выборка имеет объем n≥30, то в формулу для статистики мы подставляем вместо стандартного отклонения σ его оценку s:

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Во всем остальном проверка гипотезы не отличается от изложенной. Мы привели решение примера только для левосторонней области. Для двух других видов критических областей задачи для решения приведены в конце главы.

Второй случай: σ неизвестно и n<30

В первом случае проверка гипотезы о среднем проводилась при условии, что нам известно стандартное отклонение генеральной совокупности. Теперь рассмотрим проверку гипотезы, если стандартное отклонение неизвестно. Вместо стандартного отклонения мы будем использовать стандартное отклонение выборки. Если объем выборки достаточно велик (n≥30), мы использовали z-статистику. В случае, когда объем выборки не велик, статистика имеет распределение Стьюдента.

В качестве статистики используется случайная функция:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

где $\bar{x}$ – среднее значение выборки,
μ_0 – гипотетическое генеральное среднее,
s – стандартное отклонение выборки,
n – объем выборки.

Эта статистика имеет t-распределение с числом степеней свободы df = n – 1. Критические значения находят по таблицам t-распределения (таблица А-3). Построим критическую область для каждого из трех вариантов гипотез.

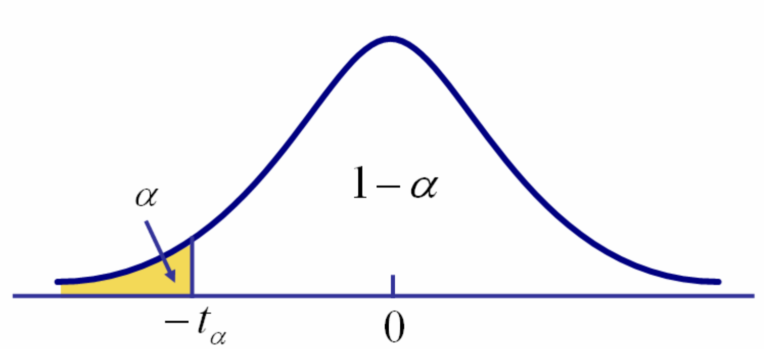


Рисунок 10-9. Левосторонняя критическая область

Вариант I. Левосторонняя критическая область. Гипотезы:

H_0: $\mu \geq \mu_0$
H_1: $\mu < \mu_0$

Для этого набора критическая область является левосторонней и определяется уравнением:

$$P(t < -t_{\alpha}) = \alpha$$

Для нахождения границы пользуемся таблицей А-3. Находим в таблице значение α там, где приведены значения для односторонней области. Поскольку таблица составлена для положительных значений, воспользуемся свойством симметрии. Найдем положительное t-значение, а затем поменяем знак:

$$P(t > t_{\alpha}) = \alpha$$

Отрицательное t-значение есть граница левосторонней критической области. Сравниваем с ней значение статистики, вычисленное по выборке, и делаем вывод о принятии или отклонении основной гипотезы.

Например, для уровня значимости α = 0,01 и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице t-значение 2,539. Критическая область запишется уравнением: $P(t < -2{,}539) = 0{,}01$.

Если значение статистики, вычисленное по выборке, окажется меньше значения минус 2,539, основную гипотезу следует отвергнуть на уровне значимости 1%.

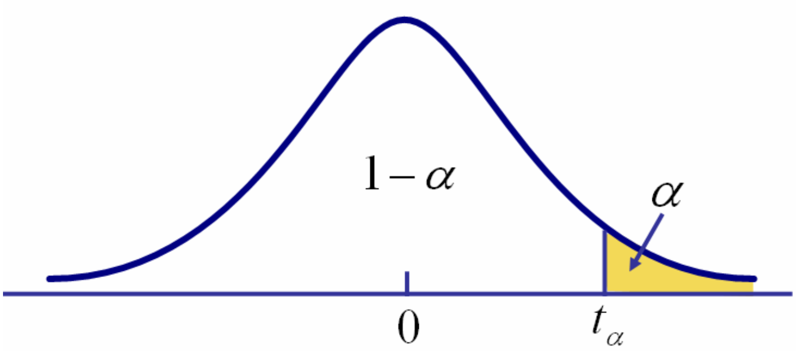


Рисунок 10-10. Правосторонняя критическая область

Вариант II. Правосторонняя критическая область. Гипотезы:

H_0: $\mu \le \mu_0$
H_1: $\mu > \mu_0$

Для этого набора критическая область является правосторонней и определяется уравнением:

$$P(t > t_\alpha) = \alpha$$

Для нахождения границы пользуемся таблицей А-3. Находим в таблице α там, где приведены значения для односторонней области. Полученное из таблицы t-значение есть граница правосторонней критической области. Вычисляем значение статистики по выборке, сравниваем с критической областью и делаем вывод о принятии или отклонении основной гипотезы.

Например, для уровня значимости α = 0,05 и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице t-значение 1,729. Следовательно, критическая область запишется уравнением:

$$P(t > 1{,}729) = 0{,}05$$

Если значение статистики, вычисленное по выборке, окажется больше значения 1,729, основную гипотезу следует отвергнуть на уровне значимости 5%.

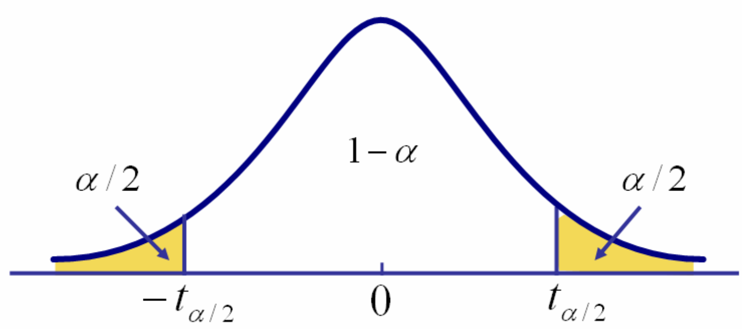


Рисунок 10-11. Двусторонняя критическая область

Вариант III. Двусторонняя критическая область. Гипотезы:

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Для этого набора критическая область является двусторонней и определяется уравнениями:

$$P(t < -t_{\alpha/2}) = \alpha / 2$$
$$P(t > t_{\alpha/2}) = \alpha / 2$$

В таблице А-3 находим как положительное, так и отрицательное t-значение. Вероятность α находим в той части заголовка таблицы, которая соответствует двусторонней критической области. Критическая область показана на рисунке 10-11 и состоит из двух частей, каждая из которых имеет границей соответствующее отрицательное или положительное t-значение.

Например, для уровня значимости α = 0,05 и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице t-значение 2,093. Следовательно, критическая область запишется уравнениями:

$$P(t < -2{,}093) = 0{,}025$$
$$P(t > 2{,}093) = 0{,}025$$

Если значение статистики, вычисленное по выборке, окажется меньше значения минус 2,093 или больше плюс 2,093, основную гипотезу следует отвергнуть на уровне значимости 5%.

Пример. Уровень преступности. За последние 20 лет средний уровень преступности в городе N составляет 399,40 преступлений на 100 тысяч жителей. Руководство города заявило в печати, что преступность находится на среднем региональном уровне. Если известно, что средний уровень преступности в регионе составляет 394,82 со стандартным отклонением 8,93, требуется проверить справедливость утверждения на уровне значимости 5%.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

H_0: $\mu \leq 394{,}82$
H_1: $\mu > 394{,}82$

ШАГ 2. Задан уровень значимости $\alpha = 0{,}05$.

ШАГ 3. Найдем критическое значение и построим критическую область. Так как $\alpha = 0{,}05$ и критическая область правосторонняя, то критическое значение равно 2,093.

ШАГ 4. По выборке вычисляем значение статистики:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{399{,}82 - 394{,}82}{8{,}93/\sqrt{20-1}} = 2{,}234$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики 2,234 попадает в критическую область, мы отвергаем нулевую гипотезу.

ШАГ 6. Формулируем ответ. Поскольку мы отвергли основную гипотезу, мы не можем согласиться с утверждением руководства города о том, что уровень преступности находится на среднем региональном уровне. Отличие является значимым.

I	II	III
Нулевая гипотеза: $H_0: p \geq p_0$	Нулевая гипотеза: $H_0: p \leq p_0$	Нулевая гипотеза: $H_0: p = p_0$
Альтернативная гипотеза: $H_1: p < p_0$	Альтернативная гипотеза: $H_1: p > p_0$	Альтернативная гипотеза: $H_1: p \neq p_0$

Рисунок 10-12. Три варианта гипотез о доле признака

10-3 Гипотеза о доле признака

Для проверки гипотезы о доле признака в генеральной совокупности применяется следующая статистика:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

где $\hat{p}$ - доля признака в выборке,
p - гипотетическая доля признака в генеральной совокупности,
q - вероятность противоположного события, q = 1 – p,
n - объем выборки.

Эквивалентной статистикой является также случайная функция:

$$z = \frac{m - np}{\sqrt{npq}}$$

Поскольку генеральная доля связана с биномиальным распределением, после соблюдения условий $np \geq 5$ и $nq \geq 5$ мы имеем дело со стандартным нормальным распределением и для нахождения границ критической области пользуемся таблицами А-2. В зависимости от вида альтернативной гипотезы мы, как и прежде, имеем три различных варианта критической области. Нахождение критических значений полностью идентично z-критерию для проверки гипотезы о среднем, поэтому мы не будем его излагать повторно.

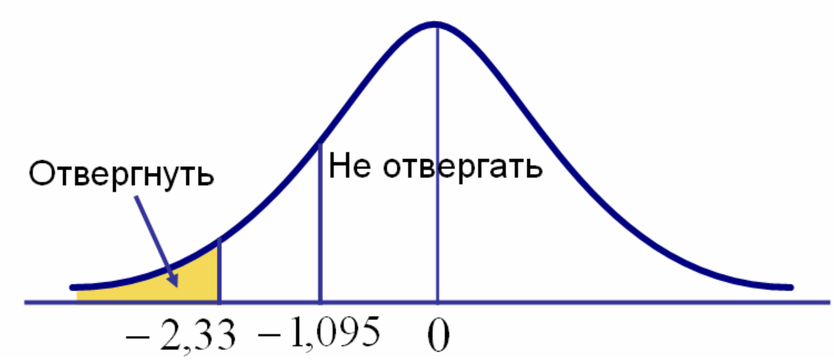


Рисунок 10-13. Значение -1,095 не попало в критическую область

Пример. Хорошие студенты. Профессор статистики утверждает, что в прошлом году более половины студентов второго курса сдали экзамен на пятерки и четверки. Усомнившись, несколько студентов решили провести исследование, в ходе которого из 30 опрошенных студентов лишь 12 сдали экзамен по статистике на пятерки и четверки. Есть ли основания думать, что профессор «слукавил»? Проверить гипотезу на уровне значимости 1%.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

$$H_0: p \geq 0{,}5$$
$$H_1: p < 0{,}5$$

ШАГ 2. Задан уровень значимости α =0,01.

ШАГ 3. Найдем критическое значение и построим критическую область. Так как α = 0,01 и критическая область левосторонняя, то критическое значение равно - 2,33.

ШАГ 4. По выборке вычисляем значение статистики:

$$z = \frac{m - np}{\sqrt{npq}} = \frac{12 - 30 \cdot 0{,}5}{\sqrt{30 \cdot 0{,}5 \cdot 0{,}5}} = \frac{-3}{2{,}74} = -1{,}095$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики - 1,095 не попадает в критическую область.

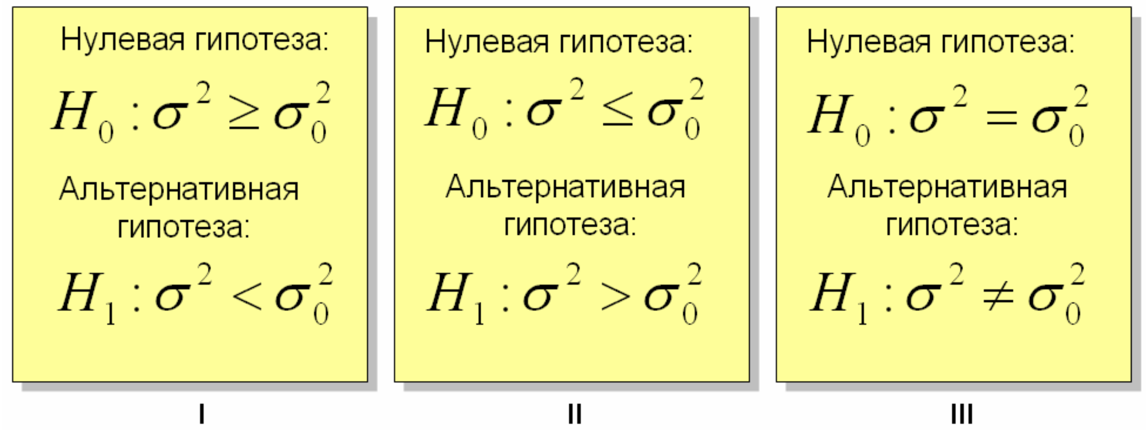


Рисунок 10-14. Три варианта гипотез о дисперсии

ШАГ 6. Формулируем ответ. Поскольку в результате эксперимента мы не получили оснований отклонить нулевую гипотезу, у нас нет оснований сомневаться в истинности утверждения профессора о результатах экзамена.

10-4 Гипотеза о дисперсии

Гипотеза о дисперсии есть предположение относительно значения дисперсии генеральной совокупности, имеющей нормальный закон распределения. Есть три варианта основной и альтернативной гипотез (рисунок 10-14).

Для проверки гипотезы используется следующая статистика:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

где σ_0 – гипотетическое генеральное стандартное отклонение,
s – стандартное отклонение выборки,
n – объем выборки.

Эта статистика имеет χ^2-распределение с числом степеней свободы df = n – 1. Для каждого из трех вариантов построим соответствующие критические области.

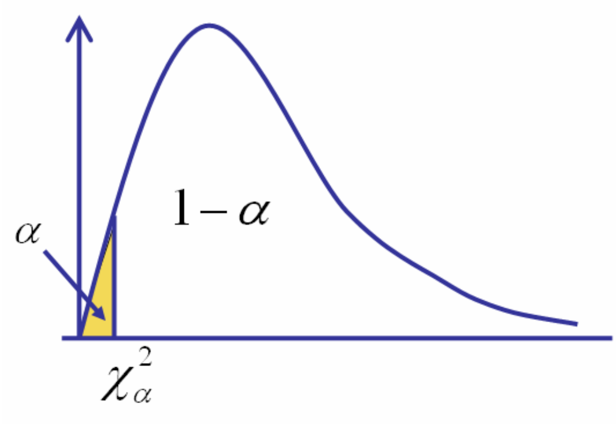


Рисунок 10-15. Левосторонняя критическая область для проверки гипотезы о дисперсии

Вариант I. Левосторонняя критическая область. Гипотезы:

$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

Для этого набора критическая область является левосторонней и определяется уравнением:

$$P(\chi^2 < \chi_\alpha^2) = \alpha$$

Для нахождения границы пользуемся таблицей А-4. Следует заметить, что в таблице находятся значения, отвечающие вероятности так называемого «хвоста» - заштрихованной площади на рисунке вверху таблицы. Находим χ^2-значение при помощи условия:

$$P(\chi^2 > \chi_\alpha^2) = 1-\alpha$$

Сравниваем значение статистики по выборке с критической областью, и делаем вывод о принятии или отклонении основной гипотезы.

Например, для уровня значимости α = 0,01 и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице χ^2-значение 7,633. Следовательно, критическая область запишется условием: $P(\chi^2 < 7{,}633) = 0{,}01$. Если значение статистики, вычисленное по выборке, окажется меньше значения 7,633, основную гипотезу следует отвергнуть на уровне значимости 1%.

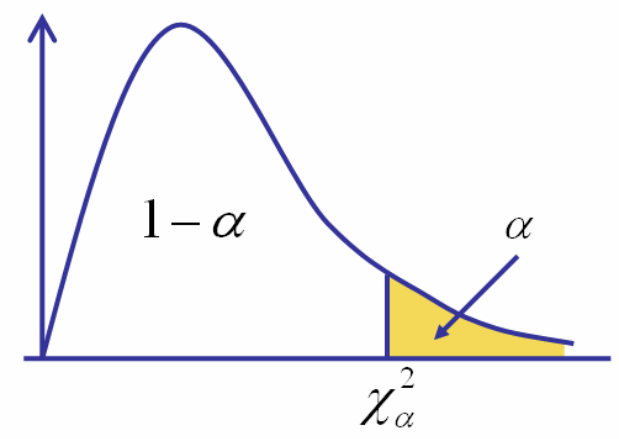


Рисунок 10-16. Правосторонняя критическая область для проверки гипотезы о дисперсии

Вариант II. Правосторонняя критическая область. Гипотезы:

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

Для этого набора критическая область является правосторонней и определяется уравнением:

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

По таблице А-4 находим χ^2-значение, которое является границей критической области. Сравниваем значение статистики по выборке с критической областью, и делаем вывод о принятии или отклонении основной гипотезы.

Например, для уровня значимости $\alpha = 0{,}05$ и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице χ^2-значение 30,144. Следовательно, критическая область запишется условием:

$$P(\chi^2 > 30{,}144) = 0{,}05$$

Если значение статистики, вычисленное по выборке, окажется больше значения 30,144, основную гипотезу следует отвергнуть на уровне значимости 5%.

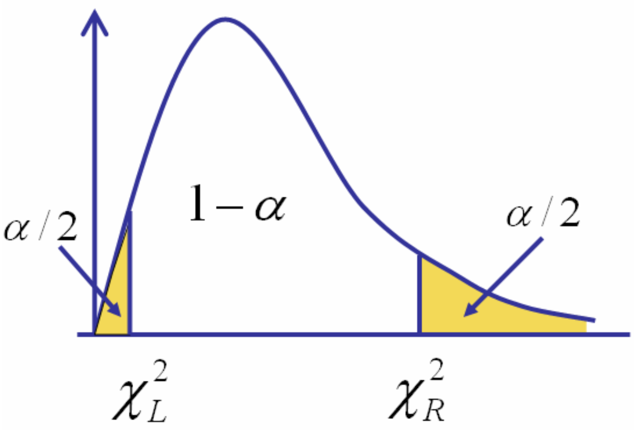


Рисунок 10-17. Двусторонняя критическая область для проверки гипотезы о дисперсии

Вариант III. Двусторонняя критическая область. Гипотезы:

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_1: \sigma^2 \neq \sigma_0^2$$

Для этого набора критическая область является двусторонней и определяется уравнениями:

$$P(\chi^2 < \chi_L^2) = \alpha/2$$
$$P(\chi^2 > \chi_R^2) = \alpha/2$$

По таблице А-4 находим χ^2-значения, которые являются границей критической области. Сравниваем значение статистики по выборке с критической областью, и делаем вывод о принятии или отклонении основной гипотезы.

Например, для уровня значимости α = 0,10 и объема выборки 20, число степеней свободы df = (20 – 1) = 19. Находим в таблице χ^2-значения 10,117 и 30,144. Следовательно, критическая область запишется уравнениями:

$$P(\chi^2 < 10{,}117) = 0{,}05$$
$$P(\chi^2 > 30{,}144) = 0{,}05$$

Если значение статистики, вычисленное по выборке, окажется меньше 10,117 или больше 30,144, основную гипотезу следует отвергнуть на уровне значимости 10%.

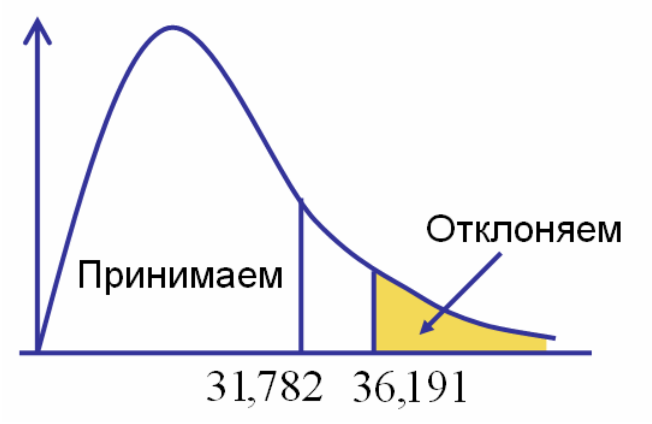


Рисунок 10-18. Значение 31,782 не попало в критическую область

Пример. Стандартное отклонение для теста IQ. Считается, что среднее значение IQ теста равно 100 со стандартным отклонением 15. Если нам захотелось проверить гипотезу о дисперсии, мы создадим случайную выборку, скажем из 20 человек и вычислим стандартное отклонение этой выборки. Оно оказалось равно 19,4. Есть ли у нас основания считать, что стандартное отклонение генеральной совокупности отличается от 15? Требуется проверить на уровне значимости 1%.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

$$H_0 : \sigma^2 \le 15^2$$
$$H_1 : \sigma^2 > 15^2$$

ШАГ 2. Задан уровень значимости $\alpha = 0{,}01$.

ШАГ 3. Найдем критическое значение и построим критическую область. Так как $\alpha = 0{,}01$ и критическая область правосторонняя, то критическое значение равно 36,191.

ШАГ 4. По выборке вычисляем значение статистики:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)\cdot 19{,}4^2}{15^2} = 31{,}782$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики 31,782 не попадает в критическую область.

ШАГ 6. Формулируем ответ. У нас нет оснований считать, что стандартное отклонение больше предполагаемого значения 15.

Используем компьютер

В качестве компьютерного упражнения можно провести проверку гипотезы о среднем при помощи SPSS. Условие задачи должно содержать первичные данные, а не результаты обработки. Раздел для проверки гипотезы находится на пути Analyze → Compare Means → One-Sample T Test. Кроме данных выборки требуется ввести также Test Value – гипотетическое среднее генеральной совокупности. Важным является сопоставление ручного счета и результата вычислений в компьютере. В данном случае не для поиска ошибок, а для перевода и изучения формата отчетов в SPSS. Отчет будет содержать две таблицы. Вычисленные результаты следует изучить и проинтерпретировать. Первая встреча с отчетом в SPSS принесет много неожиданностей.

Что означают термины

Статистическая гипотеза	**Ошибки I и II рода**	**Область принятия гипотезы**
Нулевая гипотеза	**Уровень значимости**	**Критическая область**
Альтернативная гипотеза	**Статистика, критерий**	**Критические значения**

Символы и формулы

H_0	Основная (нулевая) гипотеза
H_1	Альтернативная гипотеза
$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	z-критерий для проверки гипотезы о среднем

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$ t-критерий для проверки гипотезы о среднем

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$ z-критерий для проверки гипотезы о доле

$$z = \frac{m - np}{\sqrt{npq}}$$ Другая формула z-критерия для проверки гипотезы о доле

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$ χ2-критерий для проверки гипотезы о дисперсии

Задачи и упражнения

10-1. Преступность среди несовершеннолетних. Эксперты утверждают, что 29% всех ограблений совершаются людьми, не достигшими 18-ти лет. Проверьте это утверждение на уровне значимости $\alpha = 0{,}05$, если из 83-х ограблений, попавших в выборку, 17 были совершены теми, кому не было еще 18 лет.

10-2. Избыточный вес восьмиклассников. В одном исследовании предполагалось, что не меньше 15% всех восьмиклассников страдают от избыточного веса. В выборке из 80-ти учащихся избыточный вес оказался у 9 человек. Проверьте предположение исследования при $\alpha = 0{,}05$.

10-3. Два телефонных аппарата. Телефонная компания хочет сказать в рекламном объявлении, что более 30% ее абонентов имеют, по крайней мере, два телефонных аппарата. Чтобы подтвердить эту информацию, компания делает выборку из 200 своих абонентов и обнаруживает, что у 72-х из них есть два или более телефонных аппаратов. Подтверждают ли эти данные рекламную информацию? Возьмите $\alpha = 0{,}05$.

10-4. Кредиты клиентам банка. Менеджер банка утверждает, что размер кредита, выдаваемого клиентам, составляет в среднем 4 800$ со стандартным отклонением 800$. В выборке из 25 клиентов, бравших кредиты, средний размер кредита оказался равен 4 235$. При $\alpha = 0{,}10$, есть ли достаточные основания опровергать утверждение менеджера?

10-5. Сколько служат лампочки? Производитель утверждает, что в среднем его лампочки служат три года, или 36 месяцев, со стандартным отклонением 8 месяцев. Выбрали и протестировали 50 лампочек, их средний срок службы оказался равен 32 месяцам. Следует ли признать утверждение производителя ложным на уровне значимости $\alpha = 0{,}01$?

10-6. Превышение скорости. Опытный водитель утверждает, что инспекторы дорожного движения выписывают в среднем 60 штрафов за превышение скорости в день. Приведенные ниже данные показывают, сколько штрафов было выписано в каждый из дней одного месяца. Пусть $\sigma = 13{,}42$. Проверьте утверждение водителя при $\alpha = 0{,}05$.

72	45	36	68	69	71	57	60
83	26	60	72	58	87	48	59
60	56	64	68	42	57	57	
58	63	49	73	75	42	63	

10-7. Больничные листы. Менеджер утверждает, что на заводе среднее количество дней, пропущенных работниками по болезни меньше, чем в среднем по стране, где оно равно 10. Следующие данные показывают, сколько дней пропустили по болезни 40 работников этого завода в прошлом году. Имеются ли достаточные основания, чтобы считать утверждение менеджера истинным, при $\alpha = 0{,}05$? Используйте s для того, чтобы оценить величину σ.

0	6	12	3	3	5	4	1
3	9	6	0	7	6	3	4
7	4	7	1	0	8	12	3
2	5	10	5	15	3	2	5
3	11	8	2	2	4	1	9

10-8. Средний бал за экзамен. Основываясь на своем прошлом опыте, преподаватель полагает, что средний балл за экзамен, который оценивался по 100-бальной шкале, равен 75. Проверьте гипотезу о том, что средний балл студентов в этом году всё ещё равен 75. Возьмите $\alpha = 0{,}01$. Выборка из результатов экзаменов 20 студентов выглядит следующим образом:

80, 68, 72, 73, 76, 81, 71, 71, 65, 50,

63, 71, 70, 70, 76, 75, 69, 70, 72, 74

10-9. Обучение в компьютерном классе. Инженер компьютерного класса прочитал в отчете, что компьютерным классом пользуются в среднем 16 студентов в час. Чтобы проверить эту информацию, он случайным образом выбрал день и подсчитал количество студентов, пользовавшихся компьютерным классом в течение 8 часов. Были получены следующие результаты:

20, 24, 18, 16, 16, 19, 21, 23

При $\alpha = 0{,}05$ может ли инженер сделать вывод, что среднее действительно равно 16?

10-10. Здоровье сотрудников клиники. Крупная клиника ввела программу физической подготовки своих сотрудников, чтобы уменьшить количество пропусков работы по причине болезни. Главный врач сообщил, что служащие пропускают по причине болезни в среднем 48 часов в год. По прошествии года выборка из 18 служащих показала, что они пропустили в среднем 41 час рабочего времени; стандартное отклонение выборки равно 5. Уменьшила ли программа количество пропусков? Возьмите $\alpha = 0{,}10$.

Глава 11. Сравнение двух выборок

В этой главе мы вновь будем строить доверительные интервалы и проверять гипотезы, но теперь это будет связано со сравнением двух выборок. Для начала мы обсудим отличия независимых и парных выборок, поскольку от этого зависят последующие методы. Затем научимся сравнивать средние, доли и дисперсии для двух выборок.

11-1 Независимые и парные выборки

Существует много ситуаций, в которых у исследователя может появиться потребность сравнить две выборки и сделать выводы о генеральных совокупностях, из которых они получены. Нам предстоит рассмотреть несколько вариантов гипотез и построить доверительные интервалы в случае, когда у нас имеются две выборки. Метод зависит от того, каким образом получены эти выборки, являются ли они независимыми или зависимыми.

Ситуация 1. Две генеральные совокупности, выборки независимые. На рисунке 11-1 показана ситуация, в которой исследователь изучает и сравнивает две генеральные совокупности. Из каждой случайным отбором получают независимые выборки. Они являются независимыми в том смысле, что выбор объектов для одной не оказывает влияния или воздействия на выбор объектов для второй. Сравнивая эти выборки, можно получить выводы относительно двух генеральных совокупностей – проверить гипотезу о равенстве средних, о равенстве дисперсий или о равенстве долей, построить соответствующие доверительные интервалы.

Рисунок 11-1. Независимые выборки из двух генеральных совокупностей

Исследователя может интересовать, например, лояльность покупателей к известной торговой марке в двух городах – Москве и Санкт-Петербурге. В каждом городе независимо будет получена выборка. Затем на основе анализа этих выборок будут сделаны выводы о том, совпадает ли уровень лояльности в двух разных городах или отличается, насколько существенно это различие.

Ситуация 2. Одна генеральная совокупность, выборки независимые. Другая исследовательская ситуация возникает, если из одной генеральной совокупности получить одну большую выборку, а затем методом случайного отбора разделить ее на две (рисунок 11-2). В этом случае выборки также являются независимыми.

Рисунок 11-2. Независимые выборки из одной генеральной совокупности

Рисунок 11-3.Парные (зависимые) выборки из двух генеральных совокупностей

Эта процедура может потребоваться исследователю, чтобы проверить гипотезу об отсутствии эффекта воздействия. Влияет ли утренняя пробежка на деловую активность? Ускорит ли новый курс лечения выздоровление пациентов? Оказывает ли влияние на успеваемость студентов новая форма обучения? Такие вопросы требуют наличия двух выборок: первая рассматривается как экспериментальная, а вторая – как контрольная. На экспериментальную группу оказывается воздействие, а на контрольную – нет. Затем сравнивают результаты – если они близки, то эффект воздействия отсутствует.

Ситуация 3. Две генеральные совокупности, выборки парные. Следующая исследовательская ситуация возможна в случае парных сравнений. Две выборки получают из двух генеральных совокупностей (рисунок 11-3), но отбор производится парный, то есть зависимый.

Например, если важно сравнить мнение экспертов о двух различных сортах вина, каждый эксперт выставляет две оценки – первому и второму сорту соответственно. Получим первую выборку – мнения об одном сорте, вторую – о другом. Если эти выборки сравнить, можно сделать вывод о схожести или различии двух сортов вин.

Ситуация 4. Одна генеральная совокупность, выборки парные. В некоторых случаях исследователь имеет две парные выборки и его целью является ответ на вопрос, получены ли эти выборки из одной генеральной совокупности или из разных.

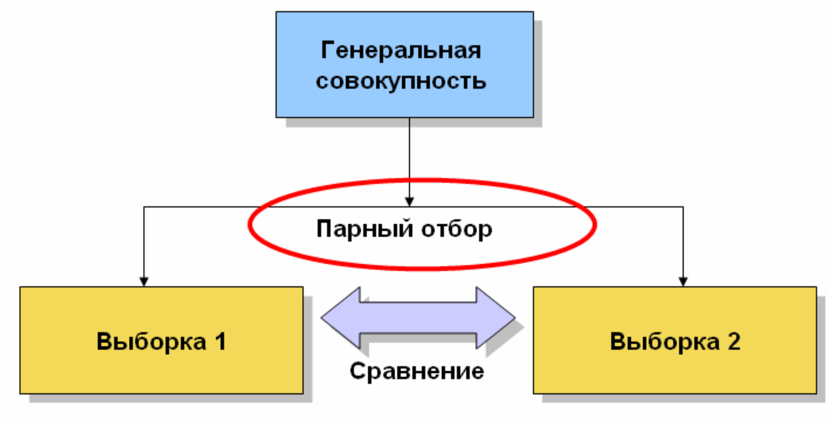


Рисунок 11-4.Парные (зависимые) выборки из одной генеральной совокупности

Другими словами, совпадают ли распределения генеральных совокупностей, из которых получены парные выборки. Эта исследовательская ситуация показана на рисунке 11-4.

Ситуация 5. Одна генеральная совокупность, одна выборка. Проверка эффекта воздействия может проводиться по следующей исследовательской схеме. Из генеральной совокупности делается выборка, которая является экспериментальной группой (рисунок 11-5). В ней дважды проводятся измерения исследуемого признака – до воздействия и после. Затем результаты сравниваются. Исследователь, к примеру, может проверить, повышаются ли знания студентов о современных проблемах в обществе, после изучения курса «Введение в специальность». Оказывает ли этот курс воздействие на уровень знаний студентов в указанной области? Измерения «до» и «после» дают возможность сравнить результаты, и, если они различия окажутся значимыми, сделать вывод о наличии эффекта воздействия.

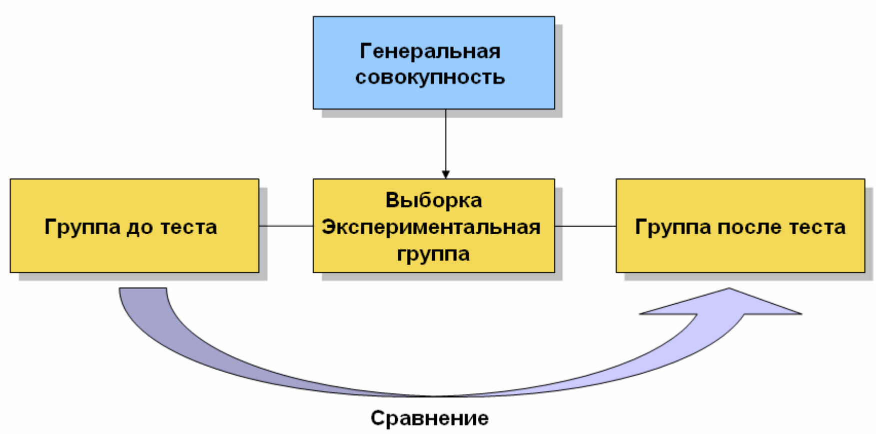


Рисунок 11-5.Выборка из одной генеральной совокупности до и после теста

Резюме

Мы рассмотрели пять исследовательских ситуаций, которые можно назвать типичными, но они не исчерпывают всех возможных схем проведения эксперимента. Далее мы будем различать, какая исследовательская схема используется, поскольку от этого зависит метод. Для парных выборок статистика будет основываться на разностях между значениями в паре. Для независимых выборок будет сравниваться их характеристики.

11-2 Сравнение средних. Независимые выборки

В этом параграфе мы познакомимся с проверкой гипотезы о равенстве средних для независимых выборок, а также построим доверительный интервал для разности средних.

Гипотеза о равенстве средних. Независимые выборки

Из двух генеральных совокупностей получены две простые случайные выборки. Выборки являются **независимыми**, что означает отсутствие связи между объектами каждой из них, и имеют объем $n \geq 30$ или обе взяты из нормально распределенных генеральных совокупностей.

Проверим в этих условиях гипотезу о равенстве средних двух генеральных совокупностей:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Нулевая гипотеза может быть записана иначе:

$$H_0 : \mu_1 - \mu_2 = 0$$

Такая гипотеза называется гипотезой о разности средних и может быть проверена для любого значения разности, в том числе не нулевого. Мы ограничим себя исключительно гипотезой о равенстве средних.

По виду альтернативной гипотезы можно заключить, что критическая область будет двусторонней. Для сравнения средних существуют также односторонние критерии, правосторонний:

$$H_0 : \mu_1 \le \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

и левосторонний:

$$H_0 : \mu_1 \ge \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Статистика, которая будет использоваться для проверки гипотезы, определяется по следующей общей формуле:

$$\frac{\text{Наблюдаемое} \quad \text{значение} - \text{Ожидаемое} \quad \text{значение}}{\text{Стандартная} \quad \text{ошибка}}$$

Наблюдаемым значением является разность двух выборочных средних, ожидаемым значением – разность двух генеральных средних. В нашем случае, при проверке гипотезы о равенстве средних, ожидаемое значение равно нулю. Стандартная ошибка вычисляется по-разному, в зависимости от того, известным нам дисперсии двух генеральных совокупностей или не известны.

Мы рассмотрим три различные ситуации и соответствующий каждой из них метод проверки гипотез и построения доверительных интервалов:

1. Дисперсии генеральных совокупностей известны.

2. Дисперсии не известны, но у нас есть основания считать их равными.

3. Дисперсии неизвестны и мы не можем считать их равными.

Ситуация 1. Дисперсии известны. В этом случае для проверки гипотезы применяется статистика:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

где $\bar{x}_1, \bar{x}_2$ – выборочные средние,
μ_1, μ_2 – гипотетические генеральные средние,
n_1, n_2 – объемы выборок,
σ_1^2, σ_2^2 – известные генеральные дисперсии.

Знаменатель представляет собой стандартную ошибку для разности средних двух выборок. Формула для стандартной ошибки получена следующим образом:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Статистика имеет стандартное нормальное распределение. Критические значения находим по таблице А-2.

Ситуация 2. Дисперсии не известны, но равны. В этом случае для проверки гипотезы применяется статистика:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

где $\bar{x}_1, \bar{x}_2$ – выборочные средние,
μ_1, μ_2 – гипотетические генеральные средние,
n_1, n_2 – объемы выборок,
s_p^2 – смешанная выборочная дисперсия.

Смешанная выборочная дисперсия вычисляется по формуле:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Статистика имеет t-распределение с числом степеней свободы df $= n_1 + n_2 - 1$.

Ситуация 3. Дисперсии не известны и не предполагаются равными. В этом случае для проверки гипотезы применяется статистика:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

где $\bar{x}_1, \bar{x}_2$ – выборочные средние,

μ_1, μ_2 – гипотетические генеральные средние,

n_1, n_2 – объемы выборок,

s_1^2, s_2^2 – выборочные дисперсии.

Статистика имеет t-распределение с числом степеней свободы df = $\min(n_1 - 1, n_2 - 1)$.

Рассмотренные ситуации позволяют решать несколько вариантов задач. Мы решим для примера только одну, когда по условию дисперсии равны, но не известны.

Пример. Девочки прогуливают не чаще. Исследователь предполагает, что среди учеников средней школы девочки чаще, чем мальчики, прогуливают занятия. Выборочное исследование 16-ти девочек показало, что их не бывает в школе примерно 3,9 дня в году, а мальчиков (22 человека) 3,6 дня. Стандартные отклонения составили 0,6 и 0,8 дня соответственно. Проверьте предположение исследователя на уровне значимости α=0,01. Предполагается, что дисперсии равны.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

где μ_1 – генеральное среднее для прогулов девочек,

μ_2 – генеральное среднее для прогулов мальчиков

ШАГ 2. Задан уровень значимости $\alpha = 0{,}01$.

ШАГ 3. Найдем критическое значение и построим критическую область. Поскольку дисперсии генеральных совокупностей неизвестны, но предполагаются равными, мы находимся во второй ситуации и применяем правосторонний t-критерий. Число степеней свободы: $df = n_1 + n_2 - 1 = 16 + 22 - 1 = 37$. Критическое значение по таблице А-3 равно 2,423. Правосторонняя критическая область $t > 2{,}423$.

ШАГ 4. По выборке вычисляем сначала смешанную выборочную дисперсию, а затем значение статистики:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} =$$

$$= \frac{(16-1)\cdot 0{,}6^2 + (22-1)\cdot 0{,}8^2}{(16-1)+(22-1)} =$$

$$= \frac{15 \cdot 0{,}36 + 21 \cdot 0{,}64}{36} = 0{,}523$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} =$$

$$= \frac{3{,}9 - 3{,}6}{\sqrt{\frac{0{,}523}{16} + \frac{0{,}523}{22}}} = \frac{0{,}3}{0{,}238} = 1{,}263$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики 1,263 не попадает в критическую область.

ШАГ 6. Формулируем ответ. Поскольку в результате эксперимента на уровне значимости 1% мы не получили оснований отклонить нулевую гипотезу, у нас нет оснований думать, что девочки прогуливают школьные занятия больше мальчиков.

Доверительный интервал для разности средних. Независимые выборки

Имеются две случайные независимые выборки объема n_1 и n_2 из двух генеральных совокупностей. Генеральные совокупности имеют нормальный закон распределения с параметрами μ_1,σ_1 и μ_2,σ_2 либо объемы обеих выборок $\geq$ 30. Мы хотим оценить разницу между средними двух генеральных совокупностей (μ_1 - μ_2).

Для этого построим доверительный интервал для разности средних в следующем виде:

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

Точность оценки E будем определять тремя способами, в зависимости от значения дисперсий генеральных совокупностей.

Ситуация 1. Если стандартные отклонения σ_1 и σ_2 известны, тогда точность оценки находится по формуле:

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Ситуация 2. Если стандартные отклонения σ_1 и σ_2 неизвестны, но подразумеваются равными, тогда точность оценки находится по формуле:

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Количество степеней свободы: df = $n_1 + n_2 - 1$.

Ситуация 3. Если стандартные отклонения σ_1 и σ_2 неизвестны и не подразумеваются равными, то точность оценки находится по формуле:

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Количество степеней свободы: df = min(n_1–1, n_2–1).

Эти ситуации напрямую связаны с теми, которые мы разобрали при проверке гипотез о равенстве средних.

Пример. Снова о пропусках занятий. Построим доверительный интервал для рассмотренного нами примера с посещением школьных занятий мальчиками и девочками.

ШАГ 1. По условию, выборочные средние равны 3,9 для девочек, 3,6 для мальчиков. Стандартные отклонения 0,6 и 0,8 соответственно. Объемы выборок 16 и 22.

ШАГ 2. Находим t-значение. В заголовке таблицы А-3 пользуемся значениями для двусторонней области. Доверительная вероятность 99% и количество степеней свободы df = 16+ 22 – 1 = 37 соответствуют t-значению 2,704.

ШАГ 3. Вычисляем смешанную выборочную дисперсию, а затем точность интервальной оценки:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = 0{,}523$$

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = 2{,}704 \cdot \sqrt{\frac{0{,}523}{16} + \frac{0{,}523}{22}} = 0{,}643$$

ШАГ 4. Подставить полученные значения в формулу для доверительного интервала:

$$(3{,}9 - 3{,}6) - 0{,}643 < (\mu_1 - \mu_2) < (3{,}9 - 3{,}6) + 0{,}643$$

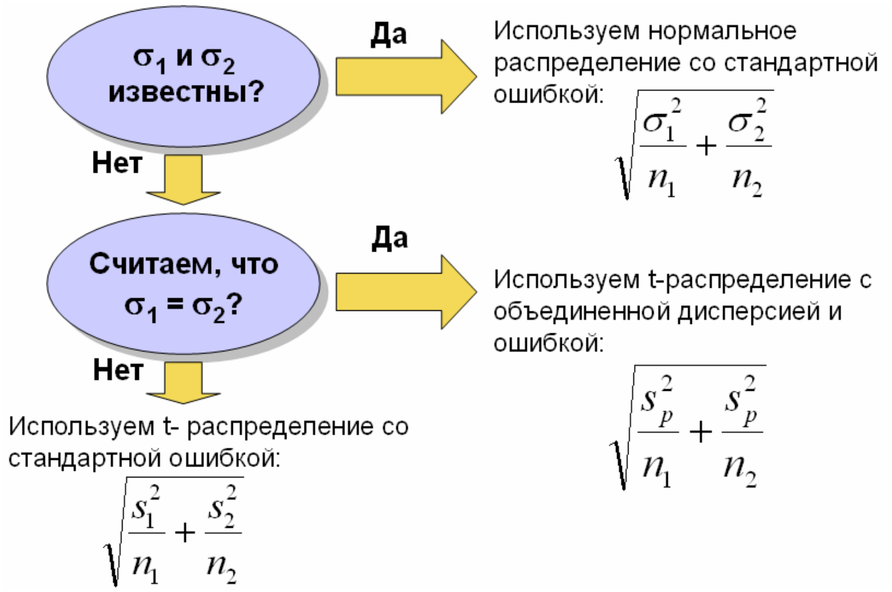


Рисунок 11-6.Три стандартных ошибки для сравнения средних

ШАГ 5. Формулируем ответ. Разница между средним числом прогулов девочек и мальчиков с доверительной вероятностью 99% находится в интервале между -0,343 и 0,943, что может быть записано условием.

$$-0{,}343 < (\mu_1 - \mu_2) < 0{,}943$$

Резюме

В этом параграфе мы рассмотрели критерии проверки гипотез и построение доверительных интервалов для сравнения средних двух нормально распределенных генеральных совокупностей. Для этого мы использовали характеристики двух независимых случайных выборок из этих совокупностей.

Построение доверительных интервалов и проверка гипотез, связанных со сравнением двух средних зависит от генеральных дисперсий (или стандартных отклонений). Можно кратко представить алгоритм решения таких задач в виде схемы, представленной на рисунке 11-6. Согласно представленной схеме, первым шагом нам следует установить, известны ли значения σ_1 и σ_2. Если да, то мы пользуемся z-распределением со стандартной ошибкой, выраженной через значения σ_1 и σ_2. Если нет, существует две возможности. Нам следует проверить, можем ли мы полагать неизвестные нам σ_1 и σ_2 равными или нет. Для поверки гипотезы о равенстве дисперсий можно обратиться к материалу параграфа 11-5.

Если у нас имеются некоторые основания считать неизвестные генеральные дисперсии (стандартные отклонения) равными, то мы применяем t-распределение с объединенной выборочной дисперсией s_p^2. Если у нас нет оснований полагать их равными – пользуемся t-распределением со стандартной ошибкой, выраженной через выборочные дисперсии s_1^2 и s_2^2.

11-3 Сравнение двух средних. Парные выборки

В этом параграфе мы рассмотрим проверку гипотезы о равенстве средних в случае парных выборок, а затем построение доверительного интервала для разности средних также для парных выборок.

Гипотеза о равенстве средних. Парные выборки

В случае парных выборок гипотеза о равенстве средних проверяется следующим образом. В распоряжении исследователя имеются две простые случайные выборки, полученные из двух генеральных совокупностей. Выборки являются парными (зависимыми). Обе выборки имеют объем $n \geq 30$, либо обе взяты из нормально распределенных генеральных совокупностей. Требуется проверить гипотезу о разности средних двух генеральных совокупностей:

$$H_0: \mu_d = 0$$
$$H_1: \mu_d \neq 0$$

Для проверки гипотезы применяется следующая статистика:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

где d - разность между двумя значениями в одной паре,
$\bar{d}$ - выборочное среднее для парных разностей,
μ_d - среднее для парных разностей генеральной совокупности,
s_d - стандартное отклонение разностей для выборки,
n - количество пар.

Таблица 11-1 Результаты теста студентов до и после тренинга

СТУДЕНТ	«ДО»	«ПОСЛЕ»	РАЗНОСТИ	КВАДРАТЫ
1	90	93	-3	9
2	91	90	1	1
3	93	89	4	16
4	89	88	1	1
5	85	88	-3	9
6	89	86	3	9
7	83	84	-1	1
8	88	83	5	25
9	84	83	1	1
10	82	80	2	4
11	83	77	6	36
12	81	76	5	25
13	72	74	-2	4
14	70	70	0	0
15	71	69	2	4
			$\Sigma d=21$	$\Sigma d^2=145$

Стандартное отклонение разностей для выборки вычисляется по формуле:

$$s_d = \sqrt{\frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n}}{n-1}}$$

Статистика имеет t-распределение с числом степеней свободы df = n – 1. Критические значения находятся при помощи таблицы А-3.

Пример. Тренинг для студентов. Группа из 15 студентов прошла тест до тренинга и после. Результаты теста представлены в таблице 11-1. Проверим гипотезу об отсутствии влияния тренинга на подготовку студентов на уровне значимости 0,05. В данном случае выборки парные.

ШАГ 1. Сформулируем проверяемую и альтернативную гипотезу:

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

ШАГ 2. Задан уровень значимости α =0,05.

ШАГ 3. Найдем критическое значение и построим критическую область. Так как число степеней свободы df = 15 – 1 = 14 и α = 0,05, критическое значение t = 2,145. Критическая область t > 2,145.

ШАГ 4. По выборке вычисляем среднее значение и стандартное отклонение выборочных разностей, а затем выборочное значение статистики:

$$\bar{d} = \frac{\sum d}{n} = \frac{21}{15} = 1{,}4$$

$$s_d = \sqrt{\frac{145 - \frac{21^2}{15}}{15-1}} = 2{,}87$$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{1{,}4 - 0}{\frac{2{,}87}{\sqrt{15}}} = 1{,}889$$

ШАГ 5. Сравним полученное значение статистики с критической областью. Значение статистики не попадает в критическую область: 1,889 < 2,145.

ШАГ 6. Формулируем ответ. Мы не имеем достаточных оснований, чтобы отвергнуть нулевую гипотезу. Это означает, что влияние тренинга не является значимым на уровне 0,05.

Доверительный интервал для среднего разностей. Парные выборки

Предположим у нас имеются две случайные парные (зависимые) выборки объема n из двух генеральных совокупностей. Генеральные совокупности имеют нормальный закон распределения с параметрами μ_1, σ_1 и μ_2, σ_2 либо объемы обеих выборок n ≥ 30. Мы хотим оценить среднее значение парных разностей для двух генеральных совокупностей, для этого построить доверительный интервал для среднего в следующем виде:

$$\bar{d} - E < \mu_d < \bar{d} + E$$

Точность оценки находится по формуле:

$$E = t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

где d - разность между двумя значениями в одной паре,
$\bar{d}$ - среднее для парных разностей для выборки,
μ_d - среднее для парных разностей генеральной совокупности,
s_d - стандартное отклонение разностей для выборки,
n - количество пар.

Построение доверительного интервала имеет тесную связь с задачей проверки гипотезы о разности средних для парных выборок.

Пример. Снова тренинг для студентов. Для примера мы построим доверительный интервал в уже рассмотренной задаче про тренинг студентов (таблица 11-1).

ШАГ 1. Вычисляем необходимые нам значения среднего значения и стандартного отклонения для выборочных разностей. Объем выборки n = 15.

$$\bar{d} = \frac{\sum d}{n} = \frac{21}{15} = 1{,}4$$

$$s_d = \sqrt{\frac{145 - \frac{21^2}{15}}{15 - 1}} = 2{,}87$$

ШАГ 2. Находим t-значение. В заголовке таблицы А-3 пользуемся значениями для двусторонней области. Доверительная вероятность 95% и количество степеней свободы df = 15 – 1 = 14 соответствуют t-значению 2,145.

ШАГ 3. Вычисляем точность интервальной оценки:

$$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}} = 2{,}145 \cdot \frac{2{,}87}{\sqrt{15}} = 1{,}590$$

ШАГ 4. Подставить полученные значения в формулу для доверительного интервала:

$$1{,}4 - 1{,}59 < \mu_d < 1{,}4 + 1{,}59$$

ШАГ 5. Формулируем ответ. Средняя разница между результатами теста до тренинга и после тренинга с доверительной вероятностью 95% находится в пределах от - 0,19 до 2,99:

$$-0{,}19 < \mu_d < 2{,}99$$

Резюме

В этом параграфе мы рассмотрели проверку гипотез и доверительные интервалы для разности средних в случае парных выборок. Статистика построена на основе разностей между парными значениями выборки, среднего значения этих разностей и их стандартного отклонения. Для нахождения оценок используется t-распределение.

11-4 Сравнение двух долей

Проверим гипотезу о равенстве двух долей, а затем построим доверительный интервал для разности долей признака двух генеральных совокупностей.

Гипотеза о равенстве долей

Имеются две генеральные совокупности, в каждой из которых доля исследуемого признака нам не известна, но мы предполагаем, что равна p_1 в первой и p_2 во второй генеральной совокупности. В нашем распоряжении имеются две простые случайные выборки, объема n_1 и n_2, соответственно, полученные из этих генеральных совокупностей.

Задача состоит в том, чтобы проверить гипотезу о равенстве долей:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

Необходимым условием является независимость выборок. Кроме этого, для выборок должны быть выполнены условия: np ≥ 5 и nq ≥ 5. Это означает, что, как минимум, пять элементов для каждой выборки имеют рассматриваемое значение признака, и, по крайней мере, пять не имеют.

Будем использовать следующие обозначения:

$p_1,\ p_2$ - генеральные доли признака,

$n_1,\ n_2$ - объемы выборок,

$m_1,\ m_2$ - число «успехов» в каждой из выборок,

$\hat{p}_1 = \dfrac{m_1}{n_1}$ - доля «успехов» в первой выборке,

$\hat{p}_2 = \dfrac{m_2}{n_2}$ - доля «успехов» во второй выборке,

$\overline{p} = \dfrac{m_1 + m_2}{n_1 + n_2}$ - доля «успехов» в двух выборках.

В качестве статистики используют следующую случайную функцию:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n_1} + \dfrac{\overline{p}(1-\overline{p})}{n_2}}}$$

Эта статистика имеет нормальное распределение, для нахождения границ критической области мы будем пользоваться таблицами нормального закона (таблицы А-2). По виду альтернативной гипотезы можем заключить, что критическая область является двусторонней.

Пример. Посещение спецкурсов. Предположим, что из 100 случайно отобранных студентов социологического факультета 43 посещают спецкурсы, а из 200 студентов-экономистов - 90 человек. На уровне значимости α = 0,05 проверить гипотезу о том, что нет значимого различия между долей посещающих спецкурсы на этих факультетах.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

ШАГ 2. Задан уровень значимости α =0,05.

ШАГ 3. По таблице нормального закона находим критические значения: z = - 1,96 и z = 1,96. Критическая область двусторонняя.

ШАГ 4. По выборке вычисляем значение статистики:

$$\hat{p}_1 = \frac{m_1}{n_1} = \frac{43}{100} = 0{,}43$$
$$\hat{p}_2 = \frac{m_2}{n_2} = \frac{90}{200} = 0{,}45$$
$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{43+90}{100+200} = \frac{133}{300} = 0{,}443$$

$$z = \frac{(0{,}43 - 0{,}45) - 0}{\sqrt{\frac{0{,}443 \cdot 0{,}577}{100} + \frac{0{,}443 \cdot 0{,}577}{200}}} = -0{,}33$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики -0,33 не попадает в критическую область.

ШАГ 6. Формулируем ответ. Считаем, что нет статистически значимого различия в доле студентов, посещающих спецкурсы, на двух факультетах.

Доверительный интервал для разности долей

Оценим генеральную долю при помощи доверительного интервала. Построение доверительного интервала связано с проверкой гипотезы о доле.

Доля значений признака в генеральной совокупности с надежностью 1-α/2 находится в доверительном интервале:

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

Точность оценки в этом случае находится по формуле:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Построим доверительный интервал для задачи о посещении спецкурсов студентами-социологами и экономистами.

ШАГ 1. Вычисляем значения выборочных долей:

$$\hat{p}_1 = \frac{m_1}{n_1} = \frac{43}{100} = 0{,}43$$

$$\hat{p}_2 = \frac{m_2}{n_2} = \frac{90}{200} = 0{,}45$$

ШАГ 2. Находим z-значение. Доверительная вероятность 95% соответствует значению 1,96.

ШАГ 3. Вычисляем точность интервальной оценки:

$$E = 1{,}96 \cdot \sqrt{\frac{0{,}43 \cdot 0{,}57}{100} + \frac{0{,}45 \cdot 0{,}55}{200}} = 0{,}119$$

ШАГ 4. Подставим полученные значения в формулу для доверительного интервала:

$$(0{,}43 - 0{,}45) - 0{,}119 < (p_1 - p_2) < (0{,}43 - 0{,}45) + 0{,}119$$

ШАГ 5. Формулируем ответ. Разница между генеральными долями студентов двух факультетов, посещающих спецкурсы, с доверительной вероятностью 95% находится в пределах от минус 13,9% до плюс 9,9%:

$$-0{,}139 < (p_1 - p_2) < 0{,}099$$

Резюме

В этом параграфе мы рассмотрели проверку гипотез и построение доверительных интервалов для оценки разности генеральных долей на основе сравнения выборочных долей. Для этого нам понадобилось z-распределение.

11-5 Сравнение двух дисперсий

В части гипотез, которые мы проверили выше, предполагалось, что дисперсии двух генеральных совокупностей равны. У нас имеется возможность проверить гипотезу о равенстве дисперсий.

В случае, когда генеральные совокупности имеют нормальное распределение, для этого существует F-критерий, называемый также критерием Фишера. Итак, мы располагаем двумя простыми случайными выборками, полученными из двух нормально распределенных генеральных совокупностей. Выборки являются независимыми. Мы хотим проверить гипотезу о равенстве дисперсий генеральных совокупностей:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

Обозначения:

σ_1^2, σ_2^2 - генеральные дисперсии,

s_1^2, s_2^2 - выборочные дисперсии,

n_1, n_2 - объемы выборок.

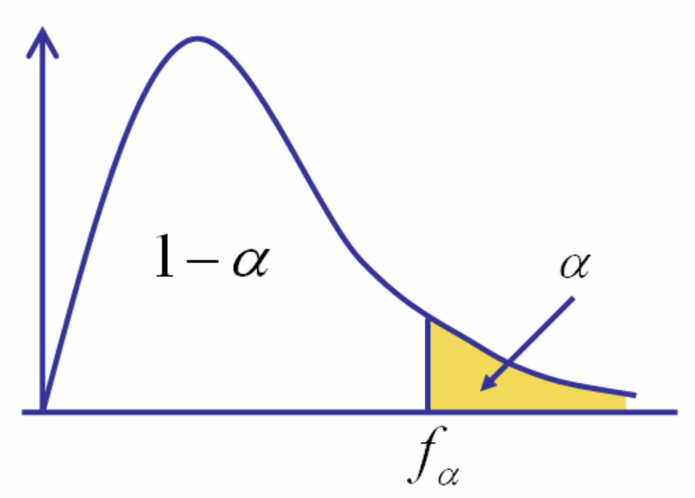


Рисунок 11-7.Критическая область для F-критерия

Для того, чтобы воспользоваться критерием, необходимо, чтобы:

$$s_1^2 > s_2^2$$

Если это условие не выполнено, мы просто должны поменять нумерацию генеральных совокупностей местами, и тогда условие окажется выполнено. Условие требуется потому, что в этом критерии других альтернативных гипотез не рассматривается и это объясняется свойствами распределения Фишера. Итак, **первая выборочная дисперсия обязана быть больше второй**. Если же дисперсии совпадают, то у нас нет оснований думать, что дисперсии исследуемых генеральных совокупностей отличаются.

Для проверки гипотезы используется статистика:

$$F = \frac{s_1^2}{s_2^2}$$

Эта статистика имеет F-распределение с числом степеней свободы: числителя $df_1 = n_1 - 1$ и знаменателя $df_2 = n_2 - 1$.

Мы встречались с F-распределением и знаем, что оно характеризуется двумя параметрами: числом степеней свободы числителя и знаменателя. Таблицы критических значений для этого распределения являются «трехмерными», поскольку каждое критическое значение определяется в зависимости от двух значений степеней свободы и значения α. В приложении А-5 эти таблицы приведены для $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

Уравнение критической области:

$$P(F > f_{\alpha}) = \alpha$$

Пример. Курить или не курить? Исследователь-медик хочет убедиться, что имеется различие между частотой биения сердца у курящих и некурящих пациентов (кол-во ударов в минуту). Результаты двух случайно отобранных групп:

Курящие:	$s_1^2 = 36$	$n_1 = 26$
Некурящие:	$s_2^2 = 10$	$n_2 = 18$

Требуется выяснить, прав ли медик. Уровень значимости $\alpha = 0{,}05$.

ШАГ 1. Сформулируем основную и альтернативную гипотезы:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

ШАГ 2. Задан уровень значимости $\alpha = 0{,}05$.

ШАГ 3. Найдем критическое значение и построим критическую область. Вычислим степени свободы:

$$\mathrm{df}_1 = n_1 - 1 = 26 - 1 = 25$$
$$\mathrm{df}_2 = n_2 - 1 = 18 - 1 = 17$$

Так как $\alpha = 0{,}05$, то находим в таблице А-5, что критическое значение равно 2,19.

ШАГ 4. По выборке вычисляем значение статистики:

$$F = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3{,}6$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики 3,6 > 2,19, попадает в критическую область.

ШАГ 6. Формулируем ответ. Медик абсолютно прав – различие имеется и является статистически значимым.

Используем компьютер

Сравнение двух выборок – одна из типичных задач, которую часто решают при помощи программного обеспечения – статистических пакетов и электронных таблиц. Поэтому следует изучить возможности компьютерных программ и решить несколько задач. Помните, что в статистические пакеты следует вводить первичные данные, а не результаты их обработки. Например, для задачи о генеральных долях предстоит ввести информацию обо всех 300 студентах, а не выборочные доли, поскольку программа сама проведет необходимую обработку данных.

Что означают термины

Независимые выборки

Гипотеза о равенстве средних

Гипотеза о равенстве долей

Парные выборки

Гипотеза о равенстве дисперсий

Смешанная выборочная дисперсия

Символы и формулы

$\bar{x}_1, \bar{x}_2$	Выборочные средние
μ_1, μ_2	Генеральные средние
n_1, n_2	Объемы выборок
s_1^2, s_2^2	Выборочные дисперсии
σ_1^2, σ_2^2	Генеральные дисперсии

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Критерий проверки гипотезы о равенстве средних, генеральные дисперсии известны

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$$df = n_1 + n_2 - 1$$

Критерий проверки гипотезы о равенстве средних, генеральные дисперсии не известны, но предполагаются равными

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Смешанная выборочная дисперсия

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\mathrm{df} = min(n_1 - 1, n_2 - 1)$$

Критерий проверки гипотезы о равенстве средних, генеральные дисперсии не известны и не предполагаются равными

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

Доверительный интервал для разности средних

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Точность интервальной оценки разности средних, генеральные дисперсии известны

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$df = n_1 + n_2 - 1$$

Точность интервальной оценки разности средних, генеральные дисперсии неизвестны, но предполагаются равными

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\mathrm{df} = min(n_1 - 1, n_2 - 1)$$

Точность интервальной оценки разности средних, генеральные дисперсии неизвестны и не предполагаются равными

d — Разность между значениями при парных сравнениях

$\bar{d}$ — Выборочное среднее парных разностей

μ_d	Среднее для парных разностей генеральной совокупности
$s_d = \sqrt{\dfrac{\sum d^2 - \dfrac{\left(\sum d\right)^2}{n}}{n-1}}$	Стандартное отклонение разностей для выборки
$t = \dfrac{\overline{d} - \mu_d}{\dfrac{s_d}{\sqrt{n}}}$ $df = n - 1$	Критерий для проверки гипотезы о разности средних для парных выборок
$\overline{d} - E < \mu_d < \overline{d} + E$	Доверительный интервал для среднего парных разностей
$E = t_{\alpha/2} \dfrac{s_d}{\sqrt{n}}$	Точность оценки
p_1, p_2	Генеральные доли признака
m_1, m_2	число «успехов» в каждой из выборок
n_1, n_2	объемы выборок
$\hat{p}_1 = \dfrac{m_1}{n_1}$	доля «успехов» в первой выборке
$\hat{p}_2 = \dfrac{m_2}{n_2}$	доля «успехов» во второй выборке
$\overline{p} = \dfrac{m_1 + m_2}{n_1 + n_2}$	доля «успехов» в двух выборках
$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n_1} + \dfrac{\overline{p}(1-\overline{p})}{n_2}}}$	Критерий для проверки гипотезы о равенстве долей
$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$	Доверительный интервал для разности долей
$E = z_{\alpha/2} \sqrt{\dfrac{\hat{p}_1 \hat{q}_1}{n_1} + \dfrac{\hat{p}_2 \hat{q}_2}{n_2}}$	Точность оценки для доверительного интервала для разности долей
$F = \dfrac{s_1^2}{s_2^2}$ $\mathrm{df}_1 = n_1 - 1$ $\mathrm{df}_2 = n_2 - 1$	F-критерий для проверки гипотезы о равенстве дисперсий

Задачи и упражнения

11-1. Таксисты и полицейские. Исследователь хочет сравнить скорость реакции таксистов и полицейских. Полученные им результаты представлены ниже. Может ли он при α=0,02 сделать вывод о том, что таксисты обладают меньшей скоростью реакции, чем полицейские. Предполагается, что генеральные совокупности распределены нормально.

Таксисты	**Полицейские**
$\bar{x}_1 = 0{,}38$	$\bar{x}_2 = 0{,}37$
$n_1 = 10$	$n_2 = 9$
$\sigma_1 = 0{,}02$	$\sigma_1 = 0{,}01$

11-2. Кто быстрее. Преподаватель считает, что студенты факультета ВМиК могут написать компьютерную программу быстрее, чем студенты мехмата. Двенадцать студентов ВМиК, попавшие в выборку, потратили в среднем по 36 минут на то, чтобы написать и отладить определенную программу. Восемнадцать студентов мехмата справились с тем же заданием в среднем за 39 минут каждый. Стандартное отклонение каждой группы равно 4 и 9 минут соответственно. При α=0,10 проверьте предположение преподавателя, считая, что дисперсии не равны.

11-3. Заполнение налоговых деклараций. Отделение налоговой инспекции потратило примерно по 27 минут на то, чтобы помочь каждому из 14-ти человек заполнить налоговую декларацию. Стандартное отклонение равно 4,3 минуты. Аудиторская фирма при заполнении налоговых деклараций потратила на каждого из 10 человек по 21 минуте. Стандартное отклонение равно 5,6 минуты. При α=0,02 найдите, есть ли разница во времени. Предполагается, что дисперсии равны.

11-4. Семинарские занятия. Преподаватель утверждает, что когда преподавание курса включает семинары, то дисперсия успеваемости больше, чем когда курс идет без семинаров. Случайным образом были выбраны две группы студентов. Дисперсия успеваемости первой группы (с семинарами) равна 103, а дисперсия второй группы (без семинаров) равна 73. В каждой группе учатся 20 студентов. При α = 0,05 проверить предположение преподавателя.

11-5. Управление стрессом. Социологу интересно узнать, повлияет ли показ фильма об управлении стрессом, на установки двенадцати человек, участвующих в исследовании. Результаты в таблице. Здесь большим числовым значениям соответствует более позитивное отношение к управлению стрессом. При α=0,05 проверьте утверждение, что просмотр данного фильма меняет установки испытуемых.

До	12	15	6	20	2	5	9	16	14	17	8	5
После	11	13	3	21	5	7	6	10	9	72	4	1

11-6. Нужен ли компьютер. Офис-менеджер хочет узнать, можно ли увеличить скорость печатания десяти секретарей, заменив печатные машинки компьютерами. В таблице число слов в минуту. На уровне значимости α =0,10 проверьте утверждение, что, используя компьютер, секретари могут печатать большее количество слов в минуту.

Секретарь	Печатная машинка	Компьютер
1	63	68
2	72	80
3	85	95
4	97	93
5	82	80
6	101	106
7	73	82
8	62	78
9	58	65
10	75	83

11-7. Государственный контроль. Из 200 хирургов 15% считают, что государство не должно контролировать здравоохранение. А из 200 терапевтов так думает 21%. Существует ли различие в долях на уровне значимости α = 0,05?

11-8. Американцы и европейцы. Из 80 американцев 55% хотели бы разбогатеть. Из 90 европейцев, хотели бы разбогатеть 45%. При α = 0,01 есть ли различие в долях?

11-9. За безопасность движения. Из 200 мужчин 130 сказали, что пользуются ремнями безопасности. Из 300 женщин отметили, что пользуются ремнями безопасности, 63 человека. При $\alpha = 0{,}01$ проверьте утверждение, что мужчины более осторожны, нежели женщины.

11-10. Цирк и зоопарк. В первой выборке из 100 человек 30% были в цирке, а во второй (тоже из 100 человек) 24% посещали зоопарк. Отличаются ли доли? Возьмите $\alpha = 0{,}02$.

11-11. Нет войне! Из 200 подростков 59 считают, что война неизбежна. А из 300 человек старше 60-ти, так думают 93. Отличается ли доля подростков, считающих войну неизбежной, от доли людей старше 60-ти лет? Возьмите $\alpha = 0{,}01$?

11-12. Автомобиль - не роскошь. В выборке из 59 второкурсников у восьми оказался свой собственный автомобиль, а среди 75 третьекурсников свои машины есть у 20 человек. Можно ли на уровне значимости $\alpha = 0{,}05$ сделать вывод, что доля третьекурсников с собственными автомобилями выше?

11-13. Мужчины и женщины. Найдите 95%-ый доверительный интервал для действительной разности долей по результатам исследования, в котором 40% из 200 мужчин и 56% из 100 женщин высказались против смертной казни.

11-14. Монетизация льгот. Найдите 99%-ый доверительный интервал для разности генеральных долей по результатам исследования, в котором 80% из 150-ти коммунистов и 60% из 200 либералов проголосовали против монетизации льгот.

Глава 12. Критерий согласия и таблицы сопряженности

В этой главе мы рассмотрим критерий согласия и еще несколько критериев, в основе которых лежит хи-квадрат распределение. В частности, во втором параграфе при помощи хи-квадрат критерия мы научимся проверять неизвестное распределение на нормальность. Далее мы будем изучать таблицы сопряженности, проверять независимость признаков (параграф 12-3) и однородность долей (параграф 12-4). В завершение познакомимся с несколькими важными коэффициентами связи.

12-1 Критерий согласия

Австрийский монах Грегор Мендель (1822-1884) в свободное время в монастыре выращивал горох. В одном из своих экспериментов он скрестил разные виды гороха – с гладкими желтыми горошинками и со сморщенными зелеными горошинками. Он заметил, что результаты были систематическими, то есть некоторые из них имели гладкие желтые горошины, другие – гладкие зеленые горошины, третьи – сморщенные желтые горошины, а четвертые – сморщенные зеленые горошины. Более того, после нескольких экспериментов процентное соотношение каждого вида оставалось практически неизменным.

Мендель сформулировал теорию, основанную на предположении доминантных и рецессивных признаков, и попытался предсказать результат. Затем он скрестил горох и исследовал 556 горошин следующего поколения, сравнивая полученные результаты с теоретическими предсказаниями. Так он пришел к подтверждению правильности теории, которая легла в основу современной генетики.

Для сравнения теории с результатами эксперимента Мендель использовал тест хи-квадрат, который мы сейчас рассмотрим.

Пример. Вкусовые предпочтения покупателей. Маркетолог хочет узнать, какому из пяти вкусов нового напитка отдают предпочтение покупатели. В таблице приведены данные опроса 100 человек:

Вишня	Клубника	Апельсин	Лимон	Виноград
32	28	16	14	10

Если нет каких-либо особых вкусовых предпочтений, то каждый вид напитка покупают с одинаковой частотой. В таком случае каждая частота должна быть равна 100/5 = 20, то есть *приблизительно* по 20 человек выберут каждый вид сока.

Вишня	Клубника	Апельсин	Лимон	Виноград
20	20	20	20	20

Первая таблица содержит наблюдаемые частоты, а вторая – ожидаемые. Дадим формальное определение наблюдаемых и ожидаемых частот.

Наблюдаемые частоты (observed frequency) – частоты, полученные по выборке.

Ожидаемые частоты (expected frequency) - частоты, полученные путем вычисления на основе теоретических представлений о предполагаемом распределении.

Критерий согласия позволяет выяснить, насколько согласуются между собой наблюдаемые частоты и ожидаемые, иными словами, являются ли различия между ними статистически значимыми. Гипотезы для примера с предпочтениями запишутся так:

H_0: У покупателей нет предпочтений по поводу видов сока.
H_1: У покупателей есть предпочтения.

Для проверки гипотезы при помощи хи-квадрат критерия необходимо соблюдение двух условий: (1) Выборка должна быть случайна; (2) Наблюдаемая частота должна быть не меньше 5.

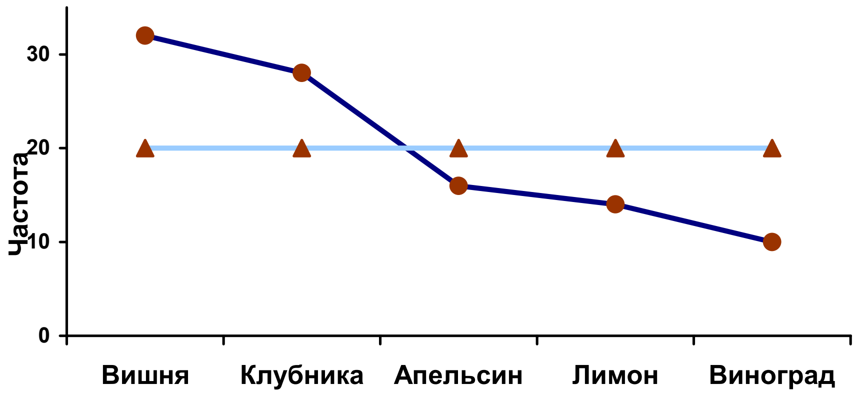


Рисунок 12-1. Согласование теоретических и наблюдаемых частот

Хи-квадрат критерий с числом степеней свободы df = n – 1 выражается в виде формулы:

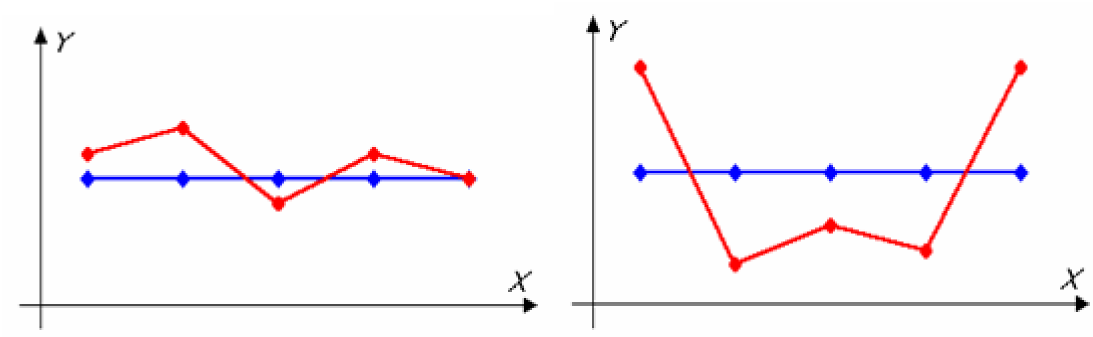
$$\chi^2 = \sum \frac{(H - O)^2}{O}$$

где H – наблюдаемая частота,
O – ожидаемая частота

Критерий согласия проверяет, насколько хорошо «согласуются» частоты. Поясним это при помощи рисунков. На рисунке 12-1 показаны две линии частот. Одна, прямая, показывает ожидаемые частоты (на уровне 20). Другая, ломаная линия, показывает наблюдаемые частоты (32 для вишни, 28 для клубники и так далее). «Согласие» означает не слишком большое различие и расхождение между двумя линиями.

На рисунке 12-2 мы намеренно в первом случае «улучшили» согласие, сблизив значения теоретические и наблюдаемые. При таком изображении у маркетолога не возникнет мыслей о неоднородности предпочтений покупателей. Во втором случае (правая картинка 12-2) мы намеренно увеличили расхождения между частотами. Маркетолог поверит по виду этого рисунка, что явные предпочтения действительно имеются. Явно видны предпочтения в пользу вишни и винограда, в то время как клубника и лимон имеют наименьшую популярность.

Хи-квадрат критерий необходим, чтобы делать выводы в неочевидных ситуациях, таких, например, как распределение частот в нашем примере. Мы не можем визуально определить, является ли расхождение статистически значимым или нет.



а) согласуются хорошо б) согласуются плохо

Рисунок 12-2. «Хорошее» и «плохое» согласование частот

Проведем проверку гипотезы об отсутствии вкусовых предпочтений покупателей при помощи критерия хи-квадрат.

ШАГ 1. Формулируем (еще раз) основную и альтернативную гипотезы:

H_0: У покупателей нет вкусовых предпочтений
H_1: У покупателей есть предпочтения

ШАГ 2. Задаем уровень значимости, например, $\alpha = 0{,}05$.

ШАГ 3. Число степеней свободы $df = 5 - 1 = 4$ и $\alpha = 0{,}05$. Критическое значение по таблице А-4 равно 9,488.

ШАГ 4. По выборке вычисляем значение статистики:

$$\chi^2 = \sum \frac{(H - O)^2}{O} =$$
$$= \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} +$$
$$+ \frac{(10-20)^2}{20} = 18$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики попадает в критическую область (18 > 9,488).

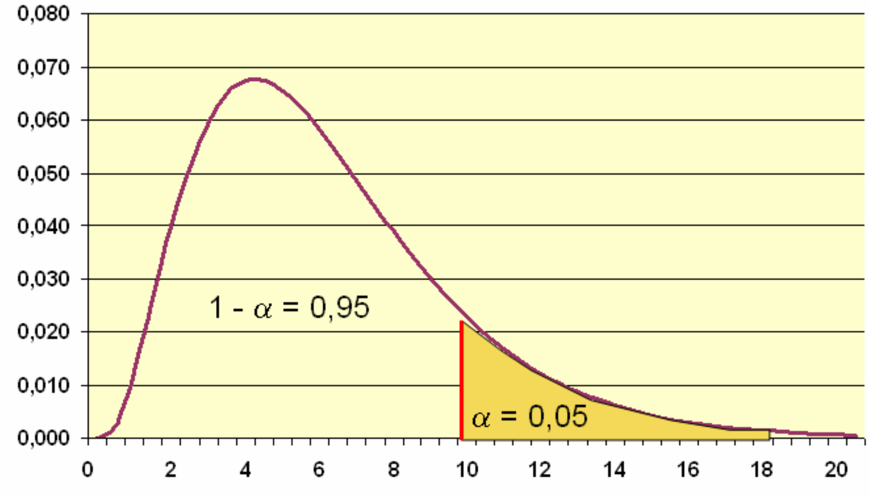


Рисунок 12-3. Критическая область хи-квадрат распределения

ШАГ 6. Формулируем ответ. Существуют значимые предпочтения покупателей по поводу вида напитка.

12-2 Проверка нормальности

Критерий согласия использовался нами в примере предыдущего параграфа для проверки гипотезы о согласовании наблюдаемого и теоретического распределений.

Второй часто используемый вид проверки - проверка гипотезы о совпадении законов распределения двух генеральных совокупностей. Предположение о виде теоретического распределения (теоретическая модель данных) в этом случае не требуется. Критерий дает нам представление о «расстоянии между двумя наборами данных» и на основе значения этого расстояния позволяет делать вывод о «согласии» между двумя распределениями.

В этом параграфе мы рассмотрим проверку гипотезы о нормальности эмпирического распределения. По имеющейся случайной выборке мы проверим, имеет ли исследуемый признак нормальное распределение.

Основная и альтернативные гипотезы:

H_0: признак имеет нормальное распределение
H_1: признак не имеет нормального распределения

Таблица 12-1. Распределение признака

ИНТЕРВАЛ	ЧАСТОТА
89,5 - 104,5	24
104,5 - 119,5	62
119,5 - 134,5	72
134,5 - 149,5	26
149,5 - 164,5	12
164,5 - 179,5	4
	200

Для проверки гипотезы используется хи-квадрат критерий с числом степеней свободы df = $n - k - 1$:

$$\chi^2 = \sum \frac{(H - O)^2}{O}$$

где H – наблюдаемая частота
O – ожидаемая частота
k – количество параметров распределения

Пример. Проверка на нормальность. В результате проведенных исследований получено интервальное распределение признака (таблица 12-1). Проверить, является ли распределение нормальным.

Прежде чем перейти к пошаговому решению задачи, проведем необходимые вычисления. Вычислим среднее и дисперсию для выборочного распределения, представленного в таблице 12-1. Для этого дополним таблицу вспомогательными столбцами (таблица 12-2).

Таблица 12-2. Дополним таблицу тремя столбцами для вычислений

ИНТЕРВАЛ	F	X	F · X	F · X²
89,5 - 104,5	24	97	2 328	225 816
104,5 - 119,5	62	112	6 944	777 728
119,5 - 134,5	72	127	9 144	1 161 288
134,5 - 149,5	26	142	3 692	524 264
149,5 - 164,5	12	157	1 884	295 788
164,5 - 179,5	4	172	688	118 336
	Σ=200		Σ=24 680	Σ=3 103 220

Таблица 12-3. Вычисление теоретически частот

ИНТЕРВАЛ	F	Z	Z-ЗНАЧЕНИЕ	P	P · N
- ∞ – 104,5	24	-1,11	0,1335	0,1335	26,7
104,5 – 119,5	62	-0,23	0,4090	0,2755	55,1
119,5 – 134,5	72	0,65	0,7422	0,3332	66,6
134,5 – 149,5	26	1,53	0,9370	0,1948	39,0
149,5 – 164,5	12	2,41	0,9920	0,0550	11,0
164,5 – + ∞	4	3,29	1,0000	0,0080	1,6
	Σ=200				Σ=199,8

В третьем столбце вычислим середину каждого интервала. В четвертом – произведение середины на частоту, в пятом – произведение квадрата середины на частоту. Просуммируем столбцы и воспользуемся результатами для вычислений:

$$\bar{x} = \frac{24680}{200} = 123,4$$

$$s = \sqrt{\frac{3103220 - \frac{24680^2}{200}}{199}} \approx 17,03$$

Рассматриваем полученные среднее и стандартное отклонение как оценки параметров нормального распределения. Вычислим теоретические частоты (таблица 12-3). В третьем столбце вычисляем z, затем по таблице находим z-значения, на их основы заполняем столбец вероятностей p, находим произведение вероятности на объем выборки, чтобы определить теоретическую частоту.

Как получены значения z? Эти значения вычисляются для каждого интервала подстановкой границы в формулу стандартизации для нормального закона:

$$z = \frac{значение - \bar{x}}{s}$$

Следуя этой формуле, например, для первого интервала получим:

$$z = \frac{104,5 - 123,5}{17,03} = -1,11$$

Таблица 12-4. Вычисление значения статистики хи-квадрат

ИНТЕРВАЛ	НАБЛЮДАЕМЫЕ ЧАСТОТЫ	ТЕОРЕТИЧЕСКИЕ ЧАСТОТЫ	РАЗНИЦА	ХИ-КВАДРАТ
–∞ – 104,5	24	26,7	+2,7	0,27
104,5 – 119,5	62	55,1	-6,9	0,86
119,5 – 134,5	72	66,6	-5,4	0,44
134,5 – 149,5	26	39,0	13,0	4,33
149,5 – +∞	16	12,6	-3,4	0.92
	Σ=200	Σ=199,8		Σ=6,83

Как находить вероятности (столбец р), если заполнен предыдущий столбец z-значений? Следует вычитать предыдущее z-значение из текущего. Например, 0,3332 = 0,7422 - 0,4090. Получаем вероятности того, что случайная величина попадет в заданный интервал. Умножаем эти вероятности на объем выборки и получим теоретические частоты.

Переходим к таблице 12-4, которая содержит наблюдаемые и теоретические частоты. Количество интервалов нам пришлось сократить, поскольку последний (шестой) интервал содержал небольшую частоту – 4, и мы объединили его с предыдущим. Такое объединение возможно только для первых и последних интервалов и не допустимо для других. Последний столбец таблицы вычислен по формуле для критерия хи-квадрат. Сумма по столбцу дает нам значение критерия:

$$\chi^2 = \sum \frac{(H-O)^2}{O} = 6{,}83$$

Все наши действия имели единственную цель – вычислить значение статистики хи-квадрат в нашем примере. Нам потребовались три вспомогательные таблицы. Теперь перейдем к пошаговому решению задачи.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: признак имеет нормальное распределение
H_1: признак не имеет нормального распределения

ШАГ 2. Задаем уровень значимости, например, α =0,05.

ШАГ 3. Число степеней свободы df = 4 – 1 = 4 и $\alpha = 0{,}05$. Критическое значение по таблице А-4 равно 9,488.

ШАГ 4. Вычисляем значение статистики:

$$\chi^2 = \sum \frac{(H - O)^2}{O} = 6{,}83$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики не попадает в критическую область (6,83 < 9,488).

ШАГ 6. Формулируем ответ. Представленное интервальное распределение на уровне значимости 5% можно признать нормальным с параметрами: средним значением 123,4 и стандартным отклонением 17,03.

Резюме

В этом параграфе мы рассмотрели критерий для проверки нормальности выборочного распределения. Нам понадобились три вспомогательные таблицы, поскольку ручной счет при вычислении статистики хи-квадрат требует большого количества арифметических операций. Таким же способом можно проверять гипотезы о других распределениях. При этом для определения теоретических частот нам потребовались бы таблицы, соответствующие проверяемому типу распределения.

12-3 Проверка независимости признаков

Критерий хи-квадрат используется для проверки независимости двух признаков, данные о которых содержатся в таблице сопряженности. Напомним, что таблицей сопряженности признаков мы называли таблицу специального вида, в которой строки и столбцы соответствуют градациям двух признаков.

Данные эксперимента

Номер респондента	Признак 1 Пол?	Признак 2 Курит?
1	Мужчина	Курит
2	Женщина	Не курит
3	Женщина	Курит
4	Мужчина	Курит
5	Мужчина	Не курит
6	Женщина	Не курит
7	Мужчина	Не курит
8	Мужчина	Курит
9	Женщина	Не курит
10	Женщина	Не курит

Таблица сопряженности

	Курит	Не курит
Мужчина	3	2
Женщина	1	4

Рисунок 12-4. Построение таблицы сопряженности признаков по данным эксперимента

На рисунке 12-4 представлены данные эксперимента, в котором опрашивались 10 человек: мужчины и женщины на предмет курения. Затем полученные данные были представлены в виде таблицы сопряженности. Таблицы могут быть не только размера 2 х 2, а любого другого, в зависимости от количества градаций каждого из двух признаков. В общем виде размер таблицы записывается (Row x Columns). Число строк соответствует количеству градаций первого признака, а число столбцов – второго.

Пример. Отношение к нововведению в клинике. Рассмотрим исследование, в котором персонал клиники высказал свое мнение о применении нового лекарственного препарата. Предположим, мы хотим проверить, имеется ли связь между категорией персонала и их мнением относительно нововведения. Первый признак «Персонал» имеет две категории: врачи и медсестры. Второй признак «отношение к новому препарату» имеет три градации: «согласен», «не согласен», «воздержался». Эта таблица сопряженности имеет размер 2 х 3.

Судя по таблице, опрошены были 400 человек, из них 200 медсестер и 200 врачей. Согласны с применением нового препарата 150 респондентов, 200 не согласны, 50 не имеют определенного мнения. В таблице указаны наблюдаемые частоты. Например, 30 врачей не имеют определенного мнения относительно препарата, 100 медсестер согласны с нововведением. Визуально сложно сказать, имеются ли связь между признаками. Для проверки нам потребуется хи-квадрат критерий.

Признак 1.	Признак 2. Отношение к новому препарату			
Категория персонала	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры	100	80	20	200
Врачи	50	120	30	200
ВСЕГО	150	200	50	400

Рисунок 12-5. Наблюдаемые частоты

Решение. Вычислим ожидаемые (теоретические) частоты. Если 150 из 400 респондентов оказалось в первом столбце, значит, теоретическая вероятность оказаться в первом столбце равна 150/400. Вероятность оказаться в первой строке 200/400. Если признаки независимы, тогда по правилу произведения вероятностей:

P(первый столбец и первая строка) =
= P(первый столбец) x P(первая строка) =
= 150/400 x 200/400 = 0,1875

Если общее число респондентов равно 400, то в первой клетке должно было теоретически оказаться 400 x 0,1875 = 75 респондентов. Вычисляя так же для всех клеток таблицы сопряженности, мы получим теоретические частоты:

строка 1 столбец 2: 200/400 x 200/400 = 100

строка 1 столбец 3: 50/400 x 200/400 = 25

строка 2 столбец 1: 150/400 x 200/400 = 75

строка 2 столбец 2: 200/400 x 200/400 = 100

строка 2 столбец 3: 50/400 x 200/400 = 25

Размещаем теоретические частоты в клетках таблицы. Вычисление одной из клеток показано на рисунке 12-7. Теперь задача критерия состоит в том, чтобы ответить, насколько сильно различаются наблюдаемые частоты и ожидаемые (теоретические).

$$\frac{50}{400} \cdot \frac{200}{400} \cdot 400 = 25$$

	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры	75	100	25	200
Врачи	75	100	25	200
ВСЕГО	150	200	50	400

Рисунок 12-7. Ожидаемые частоты

Итак, критерий согласия используется в данном случае для проверки гипотезы о независимости признаков. Гипотезы выглядят так:

H_0: рассматриваемые признаки независимы
H_1: признаки зависимы

Критерий позволяет оценить, насколько сильно различаются наблюдаемые частоты от ожидаемых. Если сильно, тогда мы признаем наличие зависимости между признаками. Формула для критерия:

$$\chi^2 = \sum \frac{(H-O)^2}{O}$$

Критерий имеет правостороннюю критическую область. Число степеней свободы определяется по формуле: df = (Row – 1) x (Columns – 1).

Проверим гипотезу о независимости признаков.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: рассматриваемые признаки независимы
H_1: признаки зависимы

ШАГ 2. Задаем уровень значимости, например, α =0,05.

ШАГ 3. Число степеней свободы df = (Row – 1) x (Columns – 1) = (2 – 1) x (3 – 1) = 2 и α = 0,05. Критическое значение по таблице А-4 равно 5,991.

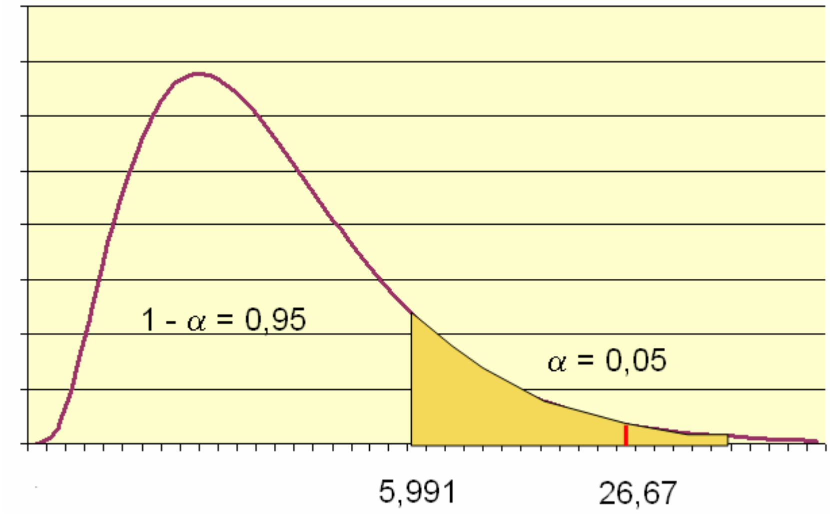


Рисунок 12-8. Значение статистики попало в критическую область

ШАГ 4. Вычисляем значение статистики:

$$\chi^2 = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} + \frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25} = 26{,}67$$

ШАГ 5. Сравниваем полученное значение с критической областью. Значение статистики попало в критическую область (26,67>5,991).

ШАГ 6. Формулируем ответ. Признаки зависимы. Отношение к новому лекарству существенно зависит от категории персонала.

Резюме

В этом параграфе критерий согласия оказался полезен для проверки гипотезы о независимости признаков. Проверка независимости строится на основе расчетов в таблице сопряженности признаков. Вычисленные теоретические частоты сравниваются с наблюдаемыми, а критерий позволяет оценить, насколько далеко «расходятся» между собой эти частоты и сделать вывод о зависимости/независимости.

Используем компьютер

По материалам настоящей главы при помощи компьютера следует научиться строить таблицы сопряженности и проверять гипотезы о независимости признаков, проверять, согласуются ли между собой распределения (Goodness of Fit). Отдельного изучения требует теория и применение компьютера для вычисления и интерпретации коэффициентов связи. В этой главе мы не обсудили коэффициенты связи, хотя на практике их использование оказывается крайне важным.

Что означают термины

Наблюдаемые частоты

Критерий согласия

Проверка независимости признаков

Ожидаемые частоты

Проверка нормальности

Теоретические частоты

Символы и формулы

$\chi^2 = \sum \frac{(H-O)^2}{O}$	Хи-квадрат критерий
H	Наблюдаемая частота
O	Ожидаемая частота

Задачи и упражнения

12-1. Опытные водители знают лучше. Опрос начинающих водителей показал, что 74% респондентов считает, что автомобилисты ездят агрессивнее, чем 5 лет назад, 23% считает, что они ездят точно так же, 3% считает, что автомобилисты ездят менее агрессивно, чем 5 лет назад. В то же время, опрос 180 опытных водителей показал, что 125 из них считают, что автомобилисты ездят агрессивнее, чем 5 лет назад, 36 – примерно одинаково, 19 человек считают, что автомобилист ездят менее агрессивно, чем 5 лет назад. При $\alpha = 0{,}10$ проверьте утверждение, что мнение опытных водителей совпадает с мнением новичков.

12-2. Когда случаются несчастья. Штатный статистик службы скорой помощи желает определить, одинаково ли распределено количество несчастных случаев в течение недели. Была выбрана наугад неделя, и получены следующие данные. Достаточно ли оснований, чтобы отвергнуть гипотезу, утверждающую, что количество несчастных случаев распределено равномерно в течение недели, при $\alpha = 0,05$?

День	Пн	Вт	Ср	Чт	Пт	Сб	Вс
Частота	28	32	15	14	38	43	19

12-3. Покупаем ружья. Владелец магазина «Охота» желает узнать, отдается ли предпочтение какому-то конкретному месяцу при покупке охотничьего ружья. Результаты продаж приведены ниже. При $\alpha = 0,05$ проверьте утверждение, что покупка оружия не зависит от конкретного месяца.

Месяц	Сентябрь	Октябрь	Ноябрь	Декабрь
Частота	18	23	28	15

12-4. Кровь американцев. Американский филиал Красного Креста сообщает о том, что 42% американцев имеют кровь типа O, 44% – типа A, 10% – типа B и 4% – типа AB. Районный медицинский исследователь говорит о том, что распределение типов крови в его регионе соответствует общим показателям в стране. Делается наугад выборка из 200 человек. Данные приведены ниже. При $\alpha = 0,10$, проверьте гипотезу исследователя.

Тип крови	A	O	B	AB
Частота	58	65	55	22

12-5. Чашка кофе к старости. Исследователю интересно узнать, есть ли связь между возрастом респондента и количеством потребляемого кофе. Было опрошено 152 человека, данные приведены ниже в таблице. При $\alpha = 0,01$ определите, есть ли связь между возрастом и количеством потребляемого человеком кофе.

	Потребление кофе		
Возраст	Низкое	Среднее	Высокое
21 – 30	18	16	12
31 – 40	9	15	27
41 – 50	5	12	10
Старше 50	13	9	6

12-6. Машины и возраст. Производитель автомобилей желает узнать, есть ли связь между возрастом покупателей и ценой купленной машины. Было опрошено 222 водителя. Данные приведены ниже в таблице. При $\alpha = 0{,}05$ определите, есть ли зависимость между ценой машины и возрастом водителя?

	Цена		
Возраст	До \$20000	\$20001 – \$30000	Выше \$30000
21 – 30	16	25	3
31 – 40	44	23	15
41 – 50	31	15	18
Старше 50	9	11	12

12-7. Не читайте газеты перед сном. Преподавателю интересно узнать, зависит ли способ получения информации от образования людей. Опрос 400 респондентов показал результаты, приведенные в таблице. При $\alpha = 0{,}05$ проверьте утверждение, что способ получения информации не зависит от образования.

	Телевидение	Газеты	Другое
Среднее	159	90	51
Высшее	27	42	31

12-8. Опасные водители. Страховая компания хочет узнать, как влияет возраст водителя на управление автомобилем в нетрезвом состоянии. Компания опросила 86 водителей четырех возрастных категорий, чтобы узнать, водят ли они машину после употребления спиртного. При $\alpha = 0{,}05$ проверьте утверждение о том, что доля водителей, ответивших утвердительно, одинакова в каждой возрастной группе.

	21-29	30-39	40-49	Старше 50
Да	32	28	26	21
Нет	54	58	60	65
ВСЕГО	86	86	86	86

Глава 13. Корреляция и регрессия

В этой главе мы будем изучать методы, позволяющие измерить силу связи между переменными, а также находить коэффициенты линейного уравнения в случае линейной зависимости двух переменных. Методы регрессии оказываются важными, поскольку позволяют делать прогнозы, интервально и точечно оценивать возможное значение одной переменной, если значение другой задано. В заключение мы обсудим вопросы надежности прогноза, то есть степени доверия полученным нами результатам при прогнозировании переменных.

13-1 Проблемы со связью

Изучение связи между двумя и более переменными является важнейшей частью статистического исследования. Менеджер хочет проверить, зависит ли объем продаж от количества рекламы в определенном периоде. Преподаватель хочет знать, есть ли зависимость между количеством часов, потраченных студентом на занятия, и результатами экзамена. Врач исследует, влияет ли кофеин на сердечные болезни и существует ли связь между возрастом человека и его кровяным давлением. Зоолог стремится узнать, есть ли связь между весом определенного животного при рождении и его продолжительностью жизни. Социолог исследует, какова связь между уровнем преступности и уровнем безработицы в регионе? Есть ли зависимость между расходами на жилье и совокупным доходом семьи? Связаны ли доход от профессиональной деятельности и уровень образования? Эти вопросы существуют вокруг нас и для ответа требуют применения методов корреляционного и регрессионного анализа.

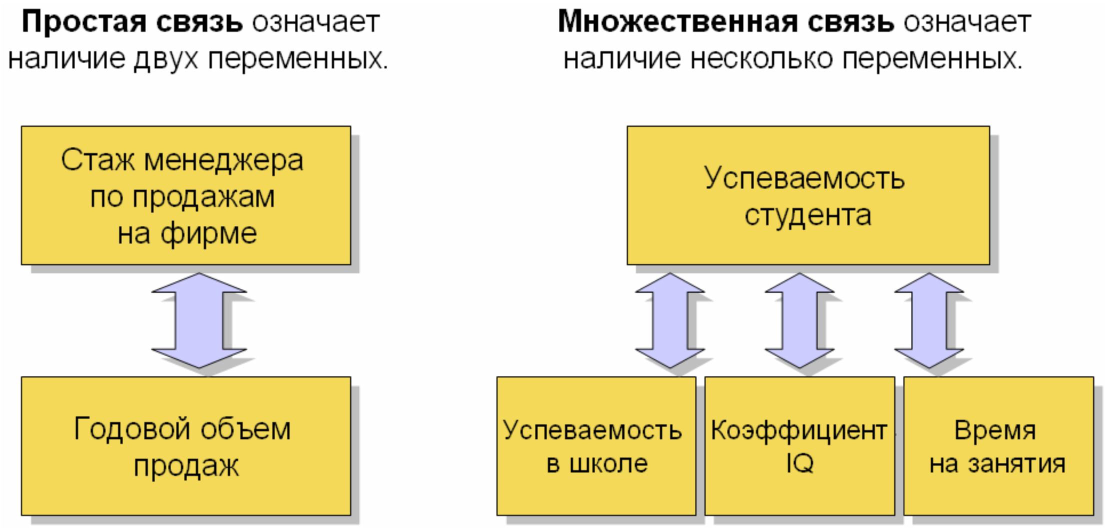


Рисунок 13-1. Простая и множественная связь

Корреляция – статистический метод, позволяющий определить, существует ли зависимость между переменными и насколько она сильна.

Регрессия – статистический метод, который применяется для описания характера связи между переменными (положительная или отрицательная, линейная или нелинейная зависимость).

Наша цель состоит в том, чтобы уметь ответить на четыре вопроса:

(1) **Существует ли связь** между двумя или более переменными?

(2) Какой **тип** имеет эта связь?

(3) Насколько она **сильна**?

(4) Какой **прогноз** можно сделать, основываясь на этой связи?

Мы будем различать ситуации, в которых исследуются простая и множественная связь. Примером простой связи является исследование зависимости годового объема продаж и стажа работы менеджера по продажам на фирме. Связаны ли эти переменные? Примером множественной связи является зависимость успеваемости студента от успеваемости в школе, коэффициента IQ и времени, которое он тратит на обучение. Успеваемость может зависеть от всех трех этих переменных.

Следует отметить, что для исследователя чрезвычайно важно установить направление связи. Какая переменная оказывает влияние на другую, где причина, а где следствие? Мы уже говорили о том, что некоторые переменные мы будем считать зависимыми, а некоторые независимыми. Для целей изучения регрессии будем считать независимой переменную, значение которой можно изменять. Зависимой переменной, напротив, будем считать такую переменную, значение которой не может изменяться исследователем. Например, если рассматривается связь между количеством часов, потраченных на подготовку к экзамену, и полученной оценкой, то количество часов может изменяться, а сама оценка в рассматриваемой модели является следствием продолжительности занятий.

Независимая переменная – переменная, которую можно изменять. **Зависимая переменная** – переменная, которую нельзя изменять по желанию исследователя, и значение которой является следствием некоторого числа явных или скрытых причин.

Как мы уже сказали, в примере со студентом переменная «Количество часов» является независимой, а переменная «Оценка на экзамене» - зависимой. Первая обозначается *x,* вторая – *y*. В рассматриваемой модели предполагается, что оценка, которую получает студент, зависит от количества часов, которые он посвятил занятиям. Студенты могут регулировать количество часов, которое они тратят на подготовку к тесту.

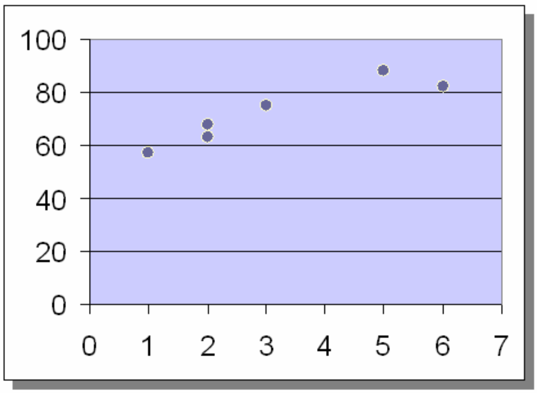
Для выявления зависимости имеет смысл строить графическое представление данных и визуально определять, имеется линейная зависимость или нет. На рисунке 13-2 показаны примеры для трех различных вариантов.

Вариант а) показывает связь между временем, которое студент потратил на подготовку к экзамену, и оценкой за этот экзамен (по 100-бальной шкале). Визуально видно, что имеется положительная линейная зависимость – чем больше часов, тем выше оценка.

Положительная линейная зависимость – при возрастании значений одной переменной увеличиваются значения второй переменной.

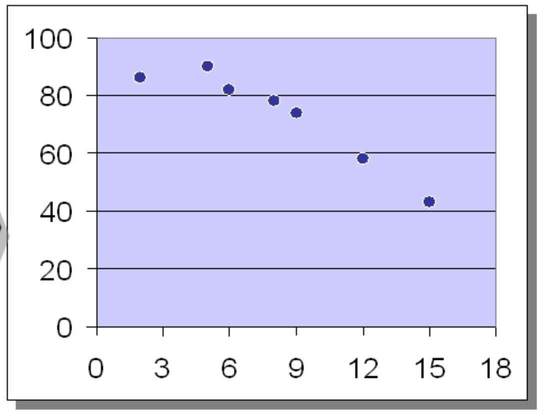
Вариант б) показывает связь между количеством пропущенных семинарских занятий и оценкой за экзамен. Визуально видно, что имеется отрицательная линейная зависимость – чем больше пропустил, тем ниже оценка.

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75



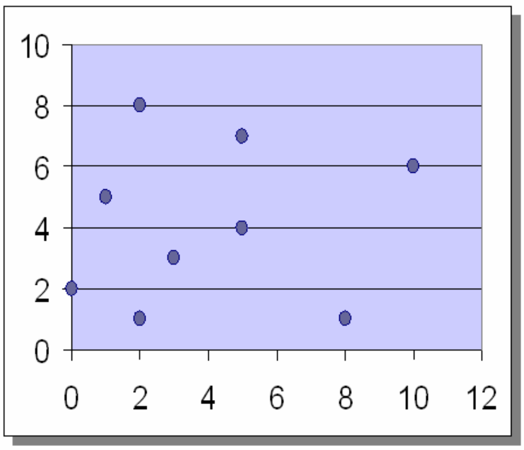
а) Положительная линейная зависимость. Чем больше часов на подготовку к экзамену, тем выше результат.

Студент	Пропустил *x*	Оценка *y*
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78



б) Отрицательная линейная зависимость. Чем больше пропустил, тем ниже оценка.

Студент	Часы *x*	Вопросы *y*
A	3	3
B	0	2
C	2	1
D	5	7
E	8	1
F	5	4
G	10	6
H	2	8
I	1	5



в) Отсутствие зависимости. Время подготовки не связано с количеством вопросов, которые задаст преподаватель.

Рисунок 13-2. Виды зависимости, определяемые визуально

Отрицательная линейная зависимость – при возрастании значений одной переменной уменьшаются значения второй переменной.

В третьем случае показана связь между количеством часов подготовки к экзамену и количеством вопросов, которые задал преподаватель на экзамене, перед тем, как поставить оценку. По графику видно, что эти переменные не связаны, зависимость не просматривается. В этом случае говорят об отсутствии зависимости между переменными. Выявление отсутствия зависимости также является неплохим результатом исследования.

И, наконец, ситуация, которая не показана на рисунках, когда зависимость между переменными просматривается, но не линейная, а какая-то другая. Следует иметь ввиду, что кроме линейных зависимостей возможны любые другие, включая квадратичную, экспоненциальную и так далее. Установление более сложных зависимостей, чем линейные, в нашем курсе не предусмотрено. После визуального анализа можно переходить к следующим стадиям исследования.

13-2 Корреляция

В этом параграфе мы познакомимся с коэффициентом корреляции, научимся его вычислять, а затем рассмотрим критерий проверки значимости коэффициента. В завершение обсудим виды причинно-следственных связей между переменными.

Коэффициент корреляции Пирсона

Для измерения силы и направления связи между переменными используется коэффициент корреляции Пирсона.

Коэффициент корреляции измеряет силу и направление связи между двумя переменными.

Мы будем обозначать выборочный коэффициент корреляции r, а коэффициент корреляции для генеральных совокупностей ϱ (ро). Коэффициент вычисляется по формуле:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Рисунок 13-3. Значения коэффициента корреляции

После несложных преобразований, из первой формулы можно получить другую формулу для коэффициента.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Как мы увидим позже, она более пригодна для вычисления коэффициента корреляции при ручном счете.

Коэффициент корреляции изменяется на отрезке от –1 до +1. Если между переменными существует сильная положительная связь, то значение r будет близко к плюс 1. Если между переменными существует сильная отрицательная связь, то значение r будет близко к минус 1. Когда между переменными нет линейной связи или она очень слабая, значение r будет близко к нулю.

Существует более детальная шкала для интерпретации значений коэффициента корреляции.

Значение r	**Уровень связи между переменными**
0,75 – 1.00	Очень высокая положительная
0,50 – 0.74	Высокая положительная
0,25 – 0.49	Средняя положительная
0,00 – 0.24	Слабая положительная
0,00 – -0.24	Слабая отрицательная
-0,25 – -0.49	Средняя отрицательная
-0,50 – -0.74	Высокая отрицательная
-0,75 – -1.00	Очень высокая отрицательная

Пример. Оценка и время подготовки к экзамену. Вычислим для примера коэффициент корреляции для данных на рисунке 13-2 а). Тем самым, ответим на вопрос о силе и направлении связи между временем на подготовку и экзаменационной оценкой.

СТУДЕНТ	ЧАСЫ X	ОЦЕНКА Y	XY	X^2	Y^2
A	6	82	492	36	6724
B	2	63	126	4	3969
C	1	57	57	1	3249
D	5	88	440	25	7744
E	2	68	136	4	4624
F	3	75	225	9	5625
	Σ=19	Σ=433	Σ=1476	Σ=79	Σ=31935

Таблица 13-1. Вычисление коэффициента корреляции в таблице

Решение. Строим вспомогательную таблицу. Дополняем таблицу данных тремя новыми столбцами: xy, x^2, y^2. В результате суммирования по столбцам, получим все необходимые для вычисления коэффициента суммы:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} =$$

$$= \frac{6 \cdot 1476 - 19 \cdot 433}{\sqrt{6 \cdot 79 - 19^2}\sqrt{6 \cdot 31935 - 433^2}} = 0{,}922$$

Значение коэффициента корреляции равно 0,922. Это означает, что существует сильная положительная связь. Мы уже выявили эту связь визуально, а теперь измерили количественно.

Значимость коэффициента корреляции

После того, как мы вычислили коэффициент корреляции для выборочных данных, возникает вполне законный вопрос, какое значение принимает коэффициент корреляции для генеральной совокупности, из которой получена выборка. Другими словами, если для шести объектов коэффициент оказался равен 0,632, то может ли быть, что для генеральной совокупности он окажется равен нулю?

Коэффициент корреляции генеральной совокупности ρ – это корреляция, вычисленная с использованием всевозможных пар значений признаков (x,y) объектов генеральной совокупности.

Нам требуется оценить коэффициент корреляции генеральной совокупности ϱ на основе выборочного значения коэффициента корреляции *r*.

Будем считать выполненными следующие **условия**: переменные *x* и *y линейно* зависимы, обе являются случайными величинами и имеют *нормальное распределение*.

Проверяются следующие гипотезы:

Основная гипотеза H_0: $\varrho = 0$
Альтернативная гипотеза H_1: $\varrho \neq 0$

Основная гипотеза утверждает, что не существует корреляции между признаками *x* и *y* в генеральной совокупности. Альтернативная гипотеза утверждает, что корреляция между признаками *x* и *y* в генеральной совокупности значима.

Когда основная гипотеза отвергается на определенном уровне значимости, это значит, что существует значимое различие между значением *r* и 0. Когда основная гипотеза принимается, это значит, что значение *r* не сильно отличается от 0 и является случайным.

Для проверки гипотезы используется t-критерий с df = n – 2 степенями свободы:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Границы двусторонней критической области находятся при помощи таблиц значений t-распределения.

Пример. Проверка значимости. Рассчитан коэффициент корреляции и его значение оказалось равно 0,897. Выборка содержала 6 пар. На уровне значимости 0,05 проверить гипотезу о значимости коэффициента корреляции.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: $\varrho = 0$
H_1: $\varrho \neq 0$

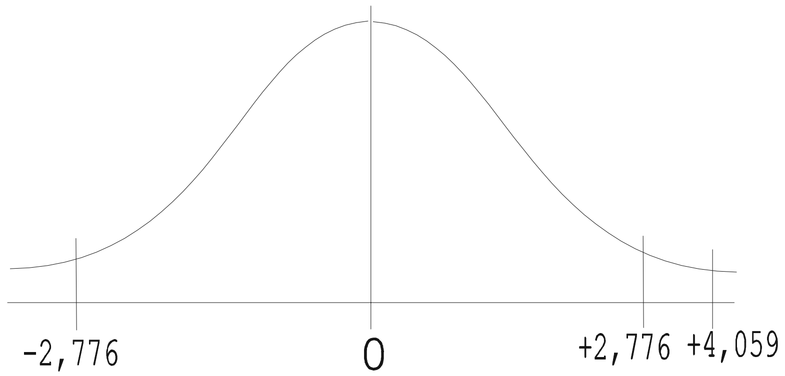


Рисунок 13-4. Значение статистики попало в критическую область

ШАГ 2. Задан уровень значимости α =0,05.

ШАГ 3. Число степеней свободы df = 6 – 2 = 4. По таблице А-3 находим критические значения: t = ±2,776. Критическая область двусторонняя.

ШАГ 4. По выборке вычисляем значение статистики:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}632 \cdot \sqrt{\frac{6-2}{1-(0{,}632)^2}} = 1{,}631$$

ШАГ 5. Сравним полученное значение с критической областью. Значение статистики 1,631 не попадает в критическую область.

ШАГ 6. Формулируем ответ. Не можем отвергнуть нулевую гипотезу. Считаем, что полученное значение выборочного коэффициента не позволяет сделать вывод о наличии связи между признаками в генеральной совокупности на уровне значимости 5%.

Корреляция и причинная связь

Когда проверка гипотезы показывает, что существует значимая линейная связь между переменными, исследователи должны рассмотреть возможные виды связи и выбрать ту, которая диктуется логикой данного исследования. Назовем несколько известных видов связи.

Ситуация 1. Прямая причинно-следственная связь между переменными. В этом случае переменная x определяет переменную y. Наличие воды ускоряет рост растений, яд вызывает смерть, жара – таяние льда.

Ситуация 2. Обратная причинно-следственная связь между переменными В этом случае переменная y определяет значение переменной x. Исследователь может думать, что чрезмерное потребление кофе вызывает нервозность. Но, может быть, очень нервный человек хочет выпить кофе, чтобы успокоить нервы? В этом случае рассматривается не прямая, а обратная связь.

Ситуация 3. Связь между переменными вызвана третьей переменной. Исследователь установил, что существует некая зависимость между числом утонувших людей и числом выпитых безалкогольных напитков в летнее время. А может быть, обе переменные растут при наступлении жары и потребностью людей во влаге?

Ситуация 4. Взаимосвязь между несколькими переменными. Исследователь может обнаружить значимую связь между оценками студентов в университете и оценками в школе. Но, возможно, действуют и другие переменные: IQ, количество часов занятий, влияние родителей, мотивация, возраст, авторитет преподавателей.

Ситуация 5. Зависимость случайна. Исследователь может найти значимую зависимость между ростом числа людей, которые занимаются спортом и ростом числа людей, которые совершают преступления. Но здравый смысл говорит о том, что связь между этими переменными случайна.

Резюме

В этом параграфе мы познакомились с коэффициентом корреляции, а также с критерием для проверки значимости коэффициента. Как мы увидели на примере, если выборочный коэффициент равен 0,6, это еще не гарантирует наличие связи в генеральной совокупности. Проверка гипотезы о равенстве нулю генерального коэффициента корреляции означает проверку значимости выборочного коэффициента.

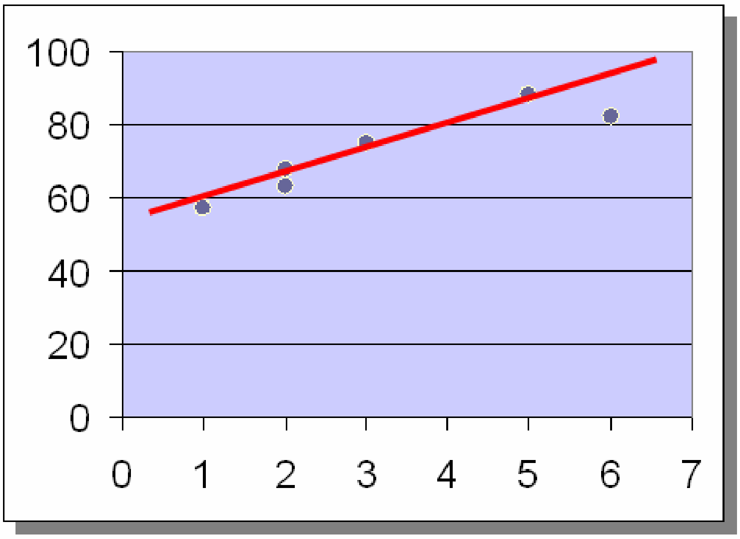


Рисунок 13-5. Линия регрессии показывает зависимость

13-3 Регрессия

На рисунке 13-5 видно, что с увеличением роста увеличивается и вес. Зависимость имеет приближенно линейный характер. Значения переменных колеблются вокруг некоей гипотетической прямой линии, которая называется *линией регрессии*. Как её построить? Существует несколько способов. Один из них – метод натянутой нити. Если взять в руки нить и приложить к рисунку, то можно выбрать наиболее подходящее с визуальной точки зрения ее положение. Если после этого зафиксировать нить в двух точках, то по ним можно определить уравнение полученной прямой линии. Это будет грубая, но пригодная в некоторых случаях оценка. Существует еще несколько методов, из которых мы рассмотрим только один, называемый методом наименьших квадратов.

Нахождение коэффициентов линейной регрессии

Наша задача состоит в том, чтобы простроить наилучшую линию. Каким образом мы ее найдем? Уравнение гипотетической линии:

$$y = ax + b$$

Задача состоит в вычислении неизвестных коэффициентов a и b.

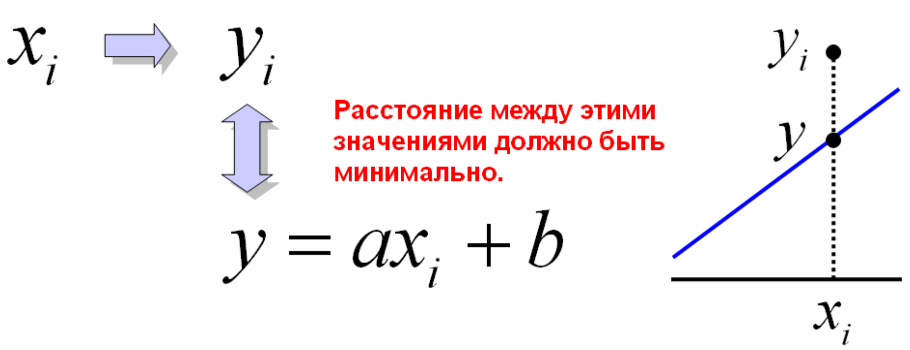


Рисунок 13-6. Каждому значению x соответствует два значения y

Наши данные представляют собой пары (x_i, y_i). Для каждого значения x_i в выборке имеется значение y_i (наблюдаемое значение). Кроме того, для каждого x существует еще одно значение y (предсказываемое значение), которое может быть получено из уравнения y = ax_i + b, если подставить в это уравнение x_i.

Линия будет наилучшей, если сумма квадратов разностей между y_i и y = ax_i + b будет минимальна.

$$\sum_{\substack{\text{по всем} \\ \text{парам (x,y)}}} (y - y_i)^2 \to \min$$

Из курса математического анализа известно, что функция от двух переменных имеет экстремум, если обе ее частные производные равны нулю.

$$f(a,b) = \sum_{i=1}^{n} (y - y_i)^2 = \sum_{i=1}^{n} ((ax_i + b) - y_i)^2$$

$$\frac{\partial f(a,b)}{\partial a} = 0 \qquad \frac{\partial f(a,b)}{\partial b} = 0$$

Решение этой системы двух уравнений с двумя неизвестными приводит к следующему ответу:

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Это и есть **коэффициенты линейной регрессии**. Коэффициент a есть наклон прямой, а коэффициент b – ее смещение вдоль оси Y.

Для вычислений ручным способом с использованием вспомогательных расчетных таблиц более просты и пригодны две другие формулы, которые эквиваленты предыдущим:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Решим задачу про экзамен и время подготовки. Таблица 13-1 содержит все необходимые данные и вычисления, поэтому, не повторяя таблицу, мы используем результаты расчетов и подставляем их в формулы для нахождения коэффициентов линейной регрессии:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{6 \cdot 1476 - 19 \cdot 433}{6 \cdot 79 - 19^2} = 5{,}6$$

$$b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{433 \cdot 79 - 19 \cdot 1476}{6 \cdot 79 - 19^2} =$$
$$= 54{,}5$$

Получили уравнение «наилучшей прямой»:

y = 5,6 x + 54,5

Что дает нам такое уравнение? Какие выводы мы можем сделать?

Вывод 1. Увеличение времени подготовки на 1 час приводит к улучшению результата на 5,6 балла.

Вывод 2. Чтобы улучшить результат на 10 баллов, нужно заниматься на 1,8 часа больше.

Вывод 3. Если не заниматься вообще ($x = 0$) – получишь 54,5 балла.

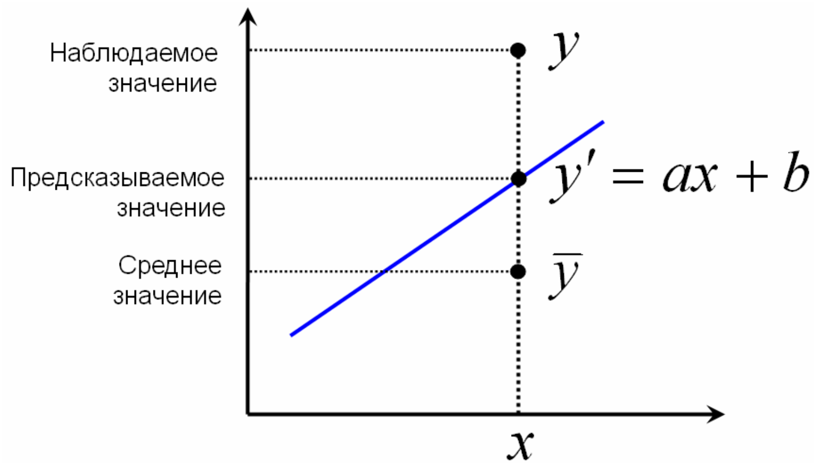


Рисунок 13-7. Три значения y, соответствующие одному x

Вывод 4. Чтобы получить 100 баллов ($y = 100$), нужно заниматься 8,1 часов.

Стоп! Два последних вывода некорректны!

В третьем и четвертом случае мы вышли за границу анализируемой области. Все наши выводы имеют силу, если мы находимся в области исследуемых данных. Часы изменяются от 1 до 6 и оценки от 57 до 88. Интерполяция на внешнюю область опасна и может приводить к необоснованным заключениям. Будьте бдительны!

Надежность прогноза

Мы уже рассмотрели несколько методов, связанных с регрессионным анализом. Если нашей задачей является установление и измерение связи между двумя переменными, то мы можем:

ШАГ 1. Графически изобразить пары значений (x, y).

ШАГ 2. Если визуально просматривается связь, найти коэффициент корреляции.

ШАГ 3. Оценить значимость коэффициента корреляции.

ШАГ 4. Если коэффициент значим, то найти уравнение регрессии.

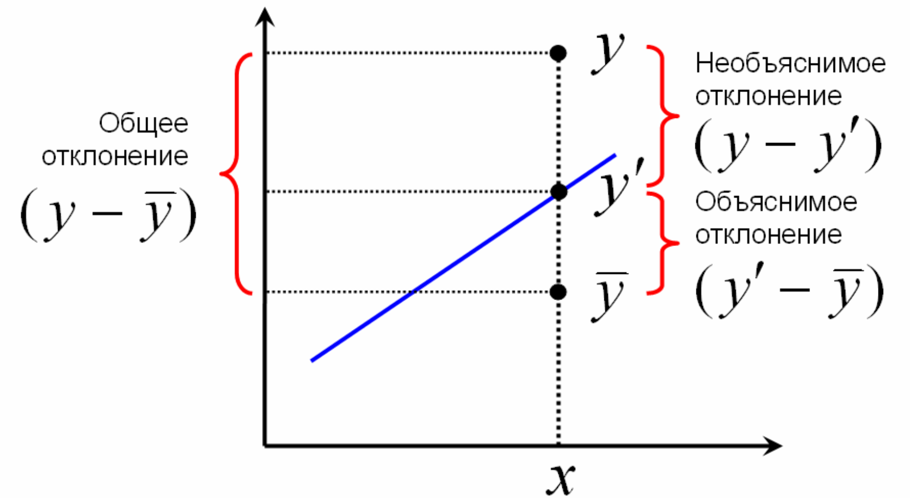


Рисунок 13-8. Необъяснимое и объяснимое отклонение

ШАГ 5. Построить разумные прогнозы: для значения независимой переменной x предсказать значение зависимой переменной y.

Нам осталось в завершение главы научиться выполнять шестой этап регрессионного исследования:

ШАГ 6. Оценить надежность прогноза: найти коэффициент детерминации, стандартную ошибку оценки и интервал предсказания.

Кроме двух значений у, которые нам известны, наблюдаемого и предсказываемого, рассмотрим среднее значение $\bar{y}$. Отклонение наблюдаемого значения y от среднего $\bar{y}$ назовем общим отклонением.

Отклонение наблюдаемого значения y от предсказываемого y' назовем необъяснимым отклонением. Отклонение предсказываемого значения y от среднего $\bar{y}$ назовем объяснимым отклонением.

Общее отклонение - отклонение наблюдаемого значения у от среднего $\bar{y}$. **Объяснимое отклонение** - отклонение предсказываемого значения y' от среднего $\bar{y}$. **Необъяснимое отклонение** - отклонение наблюдаемого значения у от предсказываемого y'.

Имеется очень важное свойство, которым обладают указанные отклонения, а именно, общая вариация есть сумма объяснимой и необъяснимой вариации:

$$\sum(y-\bar{y})^2=\sum(y'-\bar{y})^2+\sum(y-y')^2$$

СТУДЕНТ	ЧАСЫ X	ОЦЕНКА Y	y'	$\sum(y'-\bar{y})^2$	$\sum(y-y')^2$	$\sum(y-\bar{y})^2$
A	6	82	87,9	248,7	35,2	96,7
B	2	63	65,7	42,2	7,1	84,0
C	1	57	60,1	145,5	9,6	230,0
D	5	88	82,4	104,1	31,7	250,7
E	2	68	65,7	42,2	5,4	17,4
F	3	75	71,2	0,9	14,2	8,0
	Σ=19	Σ=433		Σ=583,5	Σ=103,3	Σ=686,8

Таблица 13-2. Вычисление объяснимой и необъяснимой вариации

Вернемся к рассматриваемому нами примеру и вычислим объяснимую и необъяснимую вариацию. Прежде всего, вычисляем среднее:

$$\bar{y} = 433/6 = 72{,}2$$

После получения среднего дополняем таблицу несколькими вспомогательными столбцами (таблица 13-2). Столбец y' содержит предсказываемые значения y для каждого значения x. Следующие три столбца содержат вычисления для $\sum(y'-\bar{y})^2$, $\sum(y-y')^2$ и $\sum(y-\bar{y})^2$. Суммы по этим столбцам дают нам значения объяснимой, необъяснимой и общей вариации, соответственно.

Рассмотренные выше вариации необходимы для вычисления коэффициента детерминации.

Коэффициент детерминации – это мера вариации зависимой переменной, которая определяется линией регрессии и независимой переменной.

Коэффициент детерминации r^2 вычисляется как отношение объяснимой вариации к общей вариации:

$$r^2 = \frac{\text{объяснимая вариация}}{\text{общая вариация}}$$

Для рассматриваемого нами примера:

$$r^2 = \frac{\text{объяснимая вариация}}{\text{общая вариация}} = \frac{583{,}5}{686{,}8} = 0{,}85$$

Значение коэффициента детерминации можно получить, если возвести в квадрат коэффициент корреляции.

В нашей задаче, если $r = 0{,}922$, то $r^2 = 0{,}85$ или 85%. Это означает, что 85% вариации зависимой переменной определяется вариацией независимой переменной. Оставшиеся 15% составляет необъяснимая или случайная вариация. Это значение называется **коэффициентом недетерминации** и находится вычитанием коэффициента детерминации из единицы.

Коэффициент недетерминации – разность между единицей и коэффициентом детерминации. Характеризует долю необъяснимой вариации.

По мере того, как r приближается к нулю, значение r^2 уменьшается еще быстрее. Например, если $r = 0{,}6$, то $r^2 = 0{,}36$, то есть только 36% вариации зависимой переменной могут быть связаны с вариацией независимой переменной.

Кроме коэффициента детерминации надежность прогноза характеризует **стандартная ошибка оценки и интервал предсказания**.

Стандартная ошибка оценки – это стандартное отклонение наблюдаемых значений y от предсказываемых значений y.

Стандартная ошибка оценки вычисляется по формуле:

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

Существует вторая формула для стандартной ошибки оценки, которая больше подходит для ручного счета:

$$s_{\text{est}} = \sqrt{\frac{\sum y^2 - b\sum y - a\sum xy}{n - 2}}$$

Стандартная ошибка оценки похожа на стандартное отклонение выборки, но не использует среднее значение. Чем ближе наблюдаемые значения к предсказываемым, тем меньше стандартная ошибка оценки.

Вычислим стандартную ошибку оценки в нашем примере:

$$s_{\text{est}} = \sqrt{\frac{\sum(y-y')^2}{n-2}} = \sqrt{\frac{103{,}3}{6-2}} \approx 5{,}08$$

Если значение x подставить в уравнение регрессии, мы получаем для него предсказанное значение y', которое является точечной оценкой для y фактического. Кроме этого, мы можем получить интервальную оценку - построить интервал предсказания.

Интервал предсказания – интервал, который с заданной вероятностью (1 – α) содержит фактическое значение y.

Выбрав значение α, строим интервал, который с вероятностью (1 – α) содержит фактическое значение y.

Интервал предсказания имеет следующий вид:

$$y' - E < y < y' + E$$

Точность интервальной оценки находится по формуле:

$$E = t_{\frac{\alpha}{2}} \cdot s_{\text{est}} \sqrt{1 + \frac{1}{n} + \frac{n(x-\bar{x})^2}{n\sum x^2 - \left(\sum x\right)^2}}$$

В качестве примера, ответим на вопрос, сколько баллов получит студент на экзамене, если он готовился 4 часа.

ШАГ 1. Проведем необходимые вычисления в таблице и найдем уравнение линейной регрессии.

ШАГ 2. Подставим в уравнение регрессии значение $x = 4$. Мы получим предсказываемое значение:

$$y' = 5{,}6 \cdot 4 + 54{,}5 = 76{,}9$$

ШАГ 3. Вычислим стандартную ошибку оценки. уже нашли:

$$s_{est} = 5{,}08.$$

ШАГ 4. Найдем t-значение. Для α=0,95 и числа степеней свободы df = 6 – 2 = 4 в таблице А-3 находим t=2,776.

ШАГ 5. Вычислим точность интервальной оценки E и подставим в интервал:

$$E = 2{,}776 \cdot 5{,}08 \cdot \sqrt{1 + \frac{1}{6} + \frac{6 \cdot (4 - 3{,}17)^2}{6 \cdot 79 - 19^2}} = 15{,}5$$

$$76{,}9 - 15{,}5 < y < 76{,}9 + 15{,}5$$

ШАГ 6. Напишем ответ. Прогнозируемое значение баллов, которое может получить студент при 4 часах подготовки, находится с вероятностью 95% в интервале от 61,4 до 92,4:

$$61{,}4 < y < 92{,}4$$

Резюме

В этом параграфе мы рассмотрели оценку надежности прогноза значений, получаемых на основе уравнения линейной регрессии. Коэффициент детерминации, равный квадрату коэффициента корреляции, позволяет определить, насколько вариация зависимой переменной определяется вариацией независимой переменной. Кроме точечной оценки прогнозируемого значения, мы рассмотрели интервальную оценку, которая называется интервалом предсказания и находится через вычисление стандартной ошибки оценки с использованием t-распределения.

Используем компьютер

Вычисление коэффициента корреляции, коэффициентов линейной регрессии возможно ручным способом только в задачах, имеющих учебный характер. В действительности, с ростом количества исследуемых данных такие расчеты становятся громоздкими и поэтому часто выполняются при помощи компьютера. Самым простым является использование электронных таблиц, поскольку в них можно автоматически производить вычисления рассмотренными нами методами. Мы намеренно используем вспомогательные расчетные таблицы в таком виде, в котором их легко воспроизвести в EXCEL. Следует познакомиться также и решить несколько учебных задач в соответствующих разделах SPSS, поскольку статистические пакеты имеют хороший инструментарий для корреляционного и регрессионного анализа.

Что означают термины

Корреляция	**Независимая переменная**	**Общее отклонение**
Регрессия	**Зависимая переменная**	**Объяснимое отклонение**
Сила связи между переменными	**Положительная линейная зависимость**	**Необъяснимое отклонение**
Коэффициент корреляции выборки	**Отрицательная линейная зависимость**	**Коэффициент детерминации**
Коэффициент корреляции генеральной совокупности	**Интервал предсказания**	**Стандартная ошибка оценки**

Символы и формулы

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Коэффициент корреляции

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Коэффициент корреляции (вторая формула)

$$y = ax + b$$

Уравнение линейной регрессии

$$a=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$b=\bar{y}-a\bar{x}$$

Формулы для нахождения коэффициентов линейной регрессии

$$a=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right)-\left(\sum x\right)^2}$$

Вторая формула для нахождения коэффициента a

$$b=\frac{\left(\sum y\right)\left(\sum x^2\right)-\left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right)-\left(\sum x\right)^2}$$

Вторая формула для нахождения коэффициента b

$$\sum\left(y'-\bar{y}\right)^2$$

Объяснимая вариация

$$\sum\left(y-y'\right)^2$$

Необъяснимая вариация

$$\sum\left(y-\bar{y}\right)^2$$

Общая вариация

$$r^2=\frac{\text{объяснимая вариация}}{\text{общая вариация}}$$

Коэффициент детерминации

$$s_{\text{est}}=\sqrt{\frac{\sum\left(y-y'\right)^2}{n-2}}$$

Стандартная ошибка оценки

$$s_{\text{est}}=\sqrt{\frac{\sum y^2-b\sum y-a\sum xy}{n-2}}$$

Стандартная ошибка оценки (вторая формула)

$$y'-E<y<y'+E$$

Интервал предсказания

$$E=t_{\frac{\alpha}{2}}\cdot s_{\text{est}}\sqrt{1+\frac{1}{n}+\frac{n(x-\bar{x})^2}{n\sum x^2-\left(\sum x\right)^2}}$$

Точность оценки для интервала предсказания

Задачи и упражнения

13-1. Кто больше смотрит телевизор. Исследователь хочет определить, существует ли связь между возрастом человека и тем, сколько часов в день он или она смотрит телевизор.

Возраст	18	24	36	40	58
Количество часов	3,9	2,6	2	2,3	1,2

13-2. Больничные листы сотрудников. Менеджер магазина хотел бы узнать существует ли какая-либо связь между возрастом работников и количеством больничных, которые они берут каждый год.

Возраст	18	26	39	48	53	58
Дни болезни	16	12	9	5	6	2

13-3. Успеваемость не зависит от интеллекта. Преподавателю необходимо узнать, какова связь между IQ студента и его успеваемостью.

IQ	98	105	100	100	106	95	116	112
Средний балл	2,1	2,4	3,2	2,7	2,2	2,3	3,8	3,4

13-4. Ремонт ксероксов. Офис-менеджер хочет определить, есть ли связь между тем, сколько лет уже прослужила копировальная машина, и тем, во сколько обходится ее ремонтное обслуживание в течение месяца.

Возраст ксерокса	3	5	2	1	2	4	3
Стоимость обслуживания	80	100	75	60	80	93	84

13-5. Проверка значимости. Вычислите значение коэффициента корреляции для следующих данных и проверьте гипотезу о значимости. Нарисуйте график. Проинтерпретируете результаты.

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

В задачах 13-6 по 13-9 проведите регрессионный анализ:

а) Нарисуйте график.
б) Вычислите значение коэффициента корреляции.
в) Сформулируйте нулевую и альтернативную гипотезы.
г) Проверьте их на уровне значимости $\alpha = 0{,}05$.
д) Найдите уравнение регрессии.
е) Нарисуйте линию регрессии на графике рассеивания.
ж) Сделайте выводы.

13-6. Курение вредит вашему здоровью. Было проведено исследование легочных заболеваний. Полученные данные содержат информацию о том, сколько лет человек курит и насколько сильно повреждены его легкие (в процентах). Сделайте прогноз относительно того, насколько будут повреждены легкие человека, который курит уже в течение 30-ти лет.

Кол-во лет	22	14	31	36	9	41	19
Повреждение легких	20	14	54	63	17	71	23

13-7. Пропуски занятий и оценки. Преподаватель стремится понять, как число пропущенных студентом занятий влияет на результаты успеваемости.

Количество пропусков	10	12	2	0	8	5
Итоговый балл	70	65	96	94	75	82

13-8. Любите ли вы отдыхать. Было проведено исследование относительно связи между ежемесячным доходом человека и расходами на развлечения (в долларах).

Доход	800	1200	1000	900	850	907	1100
Расходы на развлечения	60	200	160	135	45	90	150

13-9. Опять про телевизор. Для задачи 13-1 найдите уравнение регрессии и предскажите значение для возраста 38 лет. Найдите стандартную ошибку предсказания и найдите 90% интервал предсказания при $x = 20$ лет.

13-10. Опять про болезни. Для задачи 13-2 найдите уравнение регрессии и предскажите значение для 28 лет. Найдите стандартную ошибку предсказания и найдите 98% интервал предсказания при $x = 47$ лет.

Глава 14. Дисперсионный анализ

Эта глава посвящена методам дисперсионного анализа. В первом параграфе мы используем F-критерий для проверки гипотезы о равенстве средних трех и более совокупностей. Метод основан на разделении общей вариации данных на внутригрупповую и межгрупповую. Отношение этих вариаций позволяет делать выводы о наличии или отсутствии влияния исследуемого фактора. Во втором параграфе рассмотрен двухфакторный анализ – модель, при которой исследуется влияние двух независимых переменных на одну зависимую.

14-1 Однофакторный дисперсионный анализ

F-ґритерий, который мы использовали при сравнении дисперсий, будет применен для сравнения трех и более средних. Метод, который мы рассмотрим, называется *дисперсионным анализом* или в англоязычной аббревиатуре *ANOVA* (Analysis of Variance). F-критерий можно использовать для сравнения двух средних, но в этом случае он становится идентичным t-критерию.

Дисперсионный анализ (ANOVA, Analysis of Variance) – метод проверки гипотез о равенстве трех и более средних, основанный на F-критерии.

Дисперсионный анализ, который рассматривает только одну переменную, называется *однофакторным дисперсионным анализом*, в случае двух переменных - это *двухфакторный дисперсионный анализ.*

Однофакторный дисперсионный анализ (One-Way ANOVA) – метод, проверяющий влияние на зависимую переменную одной независимой переменой (фактора).

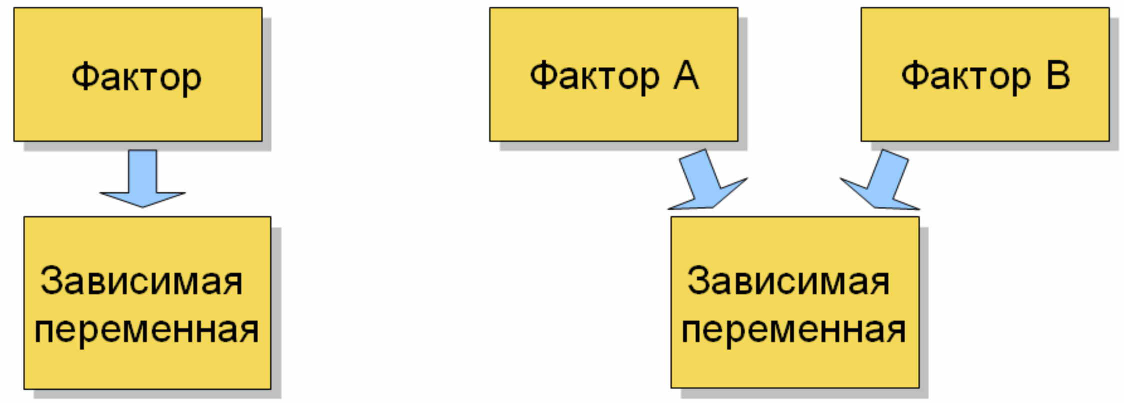


Рисунок 14-1. Однофакторный и двухфакторный анализ

Двухфакторный дисперсионный анализ (Two-Way ANOVA) – метод, проверяющий влияние на зависимую переменную двух независимых переменных (факторов).

Рассмотрим пример. В школе работают сотрудники трех разных категорий: учителя, администрация и обслуживающий персонал. У городского отдела образования возник вопрос, есть ли разница в среднем возрасте сотрудников разных категорий. Иными словами, имеется ли категория, которая старше других, или все они одинаковы в смысле среднего возраста. Вместо изучения большого количества личных дел и проведения сплошного исследования было решено взять выборки из трех генеральных совокупностей. Эти выборки представлены в таблице 14-1.

Мы имеем дело с типичной задачей однофакторного анализа. Исследуется зависимая переменная – возраст сотрудников. Рассматривается только один воздействующий фактор – категория персонала, который имеет три *уровня*: учителя, администрация, обслуживающий персонал.

Таблица 14-1. Выборки возраста трех категорий сотрудников

УЧИТЕЛЯ	АДМИНИСТРАЦИЯ	ОБСЛУЖИВАЮЩИЙ ПЕРСОНАЛ
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

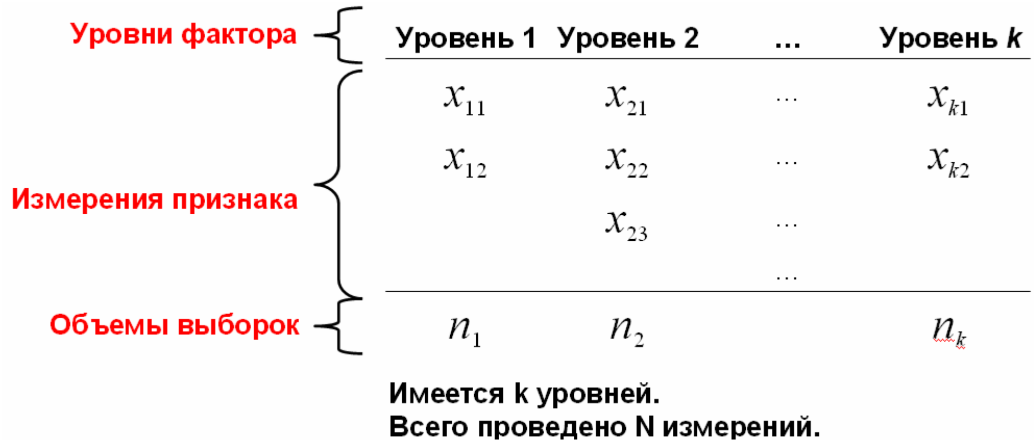


Рисунок 14-2. Представление данных в однофакторном анализе

Уровнями фактора мы будем называть градации (или значения, категории) зависимой переменной.

Для применения метода необходимо выполнение нескольких условий. Генеральные совокупности, из которых формируются выборки, должны быть распределены нормально. Выборки должны быть независимыми, не обязаны иметь одинаковый объем. Дисперсии рассматриваемых генеральных совокупностей должны быть равны.

Данные для проведения факторного дисперсионного анализа обычно представляют в виде таблицы, общий вид которой представлен на рисунке 14-2. В верхней строке в качестве заголовков столбцов записываются названия уровней фактора, в основном поле таблице – результаты наблюдений, выборки, взятые на каждом уровне фактора. В нижней строке удобно указать объем каждой выборки, поскольку они имеют (могут иметь) разный объем.

Для выявления различий между тремя и более средними, выдвигаются следующие гипотезы:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$$
$$H_1 : \text{не все средние равны}$$

Идея метода состоит в следующем. Для проверки гипотезы вычисляются две различные оценки дисперсии генеральной совокупности: *межгрупповая дисперсия* и *внутригрупповая дисперсия.*

Если нет разницы в средних, то оценки межгрупповой и внутригрупповой дисперсий приблизительно равны, значение критерия близко к 1, поэтому нулевая гипотеза принимается.

Если различие в средних значительно, межгрупповая дисперсия будет гораздо больше, чем внутригрупповая. Значение *F*-критерия будет существенно больше 1 и нулевая гипотеза отвергается.

Тем самым, при проверке гипотезы о равенстве средних, мы используем сравнение дисперсий. Собственно поэтому метод получил такое название – *дисперсионный анализ.*

Степени свободы F-распределения задаются двумя значениями:

Числителя:	df = k – 1
Знаменателя:	df = N – k

В последней строке *N* обозначает общий размер всех выборок: $N = n_1 + n_2 + \ldots + n_k$. Уравнение критической области, которая всегда является правосторонней, запишется в следующем виде:

$$P(F > f_\alpha) = \alpha$$

Для вычисления выборочного значения критерия требуется сделать следующее. Сначала вычисляются средние для каждой выборки $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$ и среднее по таблице $\bar{\bar{x}}$. Затем последовательно вычисляются следующие величины.

Межгрупповая сумма квадратов отклонений (**S**um **S**quare **B**etween Groups):

$$SS_b = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2$$

Внутригрупповая сумма квадратов отклонений (**S**um **S**quare **W**ithin Groups):

$$SS_w = \sum (x - \bar{x}_i)^2$$

Общая сумма квадратов отклонений (Sum Square):

$$SS = \sum (x - \bar{x})^2 = SS_b + SS_w$$

Межгрупповая (факторная) дисперсия (Mean Square Between Groups):

$$MS_B = \frac{SS_B}{k-1}$$

Таблица 14-2. Вспомогательная таблица для вычисления F-значения

	Сумма квадратов	df	Среднее квадратичное	*F*
Между группами	SS_B	$k-1$	MS_B	*F*-значение
Внутри групп	SS_W	$N-k$	MS_W	
Итого	$SS_B + SS_W$	$N-1$	$MS_B + MS_W$	

Внутригрупповая (остаточная) дисперсия (**M**ean **S**quare **W**ithin Groups):

$$MS_W = \frac{SS_W}{N-k}$$

F-критерий для проверки гипотезы:

$$F = \frac{MS_B}{MS_W}$$

Результаты вычисления удобно размещать в специальной таблице (рисунок 14-3). Как мы увидим в примере, она устроена так, что помогает делать необходимые вычисления и получить выборочное F-значение.

Рассмотрим стандартную последовательность действий для проверки гипотезы о равенстве средних.

ШАГ 1. Сформулировать нулевую и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_m$$
$$H_1 : \text{не все средние равны}$$

ШАГ 2. Задать уровень значимости α.

ШАГ 3. Найти критическое F-значение по заданному уровню значимости и степеням свободы числителя df = $k-1$ и знаменателя df = $N-k$, используя таблицу А-5. Построить критическую область.

ШАГ 4. Найти среднее для каждой выборки ($\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$) и среднее по таблице $\bar{\bar{x}}$.

ШАГ 5. Вычислить межгрупповую и внутригрупповую сумму квадратов.

ШАГ 6. Составить вспомогательную таблицу. Записать в таблицу полученные суммы квадратов. В соседнем столбце записать соответствующие степени свободы (df = k – 1 и df = N – k).

ШАГ 7. В таблице вычислить межгрупповую и внутригрупповую дисперсию и их сумму, а затем F-значение.

ШАГ 8. Сравнить полученное выборочное значение F-критерия с критической областью и сделать вывод.

ШАГ 9. Написать ответ.

Пример. Задача про школьный персонал. Вернемся к задаче про школьный персонал и решим ее в указанной последовательности шагов.

ШАГ 1. Формулируем нулевую и альтернативную гипотезы:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_m$$
$$H_1 : \text{не все средние равны}$$

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. По заданному уровню значимости α=0,05 и степеням свободы:

числителя df = k – 1 = 3 – 1 = 2
знаменателя df = N – k = 19 – 3 = 16

в таблице А-5 получаем F-значение 3,63. Критическая область запишется в виде F > 3,63.

ШАГ 4. В таблице 14-3 мы вычислили объемы выборок и средние для каждой выборки. Среднее по таблице $\bar{\bar{x}} = 39$.

Таблица 14-3. Вычисление средних и объема выборок

УЧИТЕЛЯ	АДМИНИСТРАЦИЯ	ОБСЛУЖИВАЮЩИЙ ПЕРСОНАЛ
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	
$n_1 = 6$	$n_2 = 7$	$n_3 = 6$
$\bar{x}_1 = 35{,}8$	$\bar{x}_2 = 44{,}6$	$\bar{x}_3 = 35{,}7$
	$N = n_1 + n_2 + n_3 = 19$	

ШАГ 5. Вычисляем межгрупповую и внутригрупповую сумму квадратов.

$$SS_b = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2 =$$
$$= 6 \cdot (35{,}8 - 39)^2 + 7 \cdot (44{,}6 - 39)^2 + 6 \cdot (35{,}7 - 39)^2 =$$
$$= 344{,}1$$

$$SS_w = \sum (x - \bar{x}_i)^2 =$$
$$= (24 - 35{,}8)^2 + (27 - 35{,}8)^2 + \ldots + (48 - 35{,}8)^2 + (40 - 35{,}8)^2 +$$
$$+ (59 - 44{,}6)^2 + (35 - 44{,}6)^2 + \ldots + (54 - 44{,}6)^2 + (56 - 44{,}6)^2 +$$
$$+ (34 - 35{,}7)^2 + (29 - 35{,}7)^2 + \ldots + (40 - 35{,}7)^2 + (45 - 35{,}7)^2 =$$
$$= 1669{,}9$$

ШАГ 6. Составим вспомогательную таблицу (таблица 14-4). Запишем в таблицу полученные суммы квадратов. В соседнем столбце запишем соответствующие степени свободы (df = 2 и df = 16).

ШАГ 7. В таблице вычислим межгрупповую и внутригрупповую дисперсию и их сумму, а затем F-значение:

$$MS_B = \frac{SS_B}{k - 1} = \frac{344{,}1}{2} = 172{,}06$$

Таблица 14-4. Вычисление F-значения в таблице

	Сумма квадратов	df	Среднее квадратичное	F
Между группами	344,1	2	172,06	1,649
Внутри групп	1669,9	16	104,37	
Итого	2014,0			

$$MS_W = \frac{SS_W}{N-k} = \frac{1669,9}{16} = 104,37$$

$$F = \frac{MS_B}{MS_W} = \frac{172,06}{104,37} = 1,649$$

ШАГ 8. Сравним полученное выборочное значение F-критерия с критической областью: 1,649 < 3,633. Полученное значение статистики не попало в критическую область.

ШАГ 9. Сформулируем ответ. У нас нет оснований думать, что средний возраст персонала разных категорий различен на уровне значимости 5%.

Мы убедились, что применение вычислительной таблицы удобно. Кроме того, эта таблица является привычной для тех, кто пользуется компьютером для вычисления выборочного F-значения. На рисунке 14-2 показан отчет, выдаваемый программой при вычислении F-значения в SPSS. Термины на английском языке полностью соответствуют используемым нами терминам и обозначениям.

ANOVA

AGE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	344,119	2	172,060	1,649	,223
Within Groups	1669,881	16	104,368		
Total	2014,000	18			

Рисунок 14-2. Отчет, полученный при вычислении F-значения в SPSS

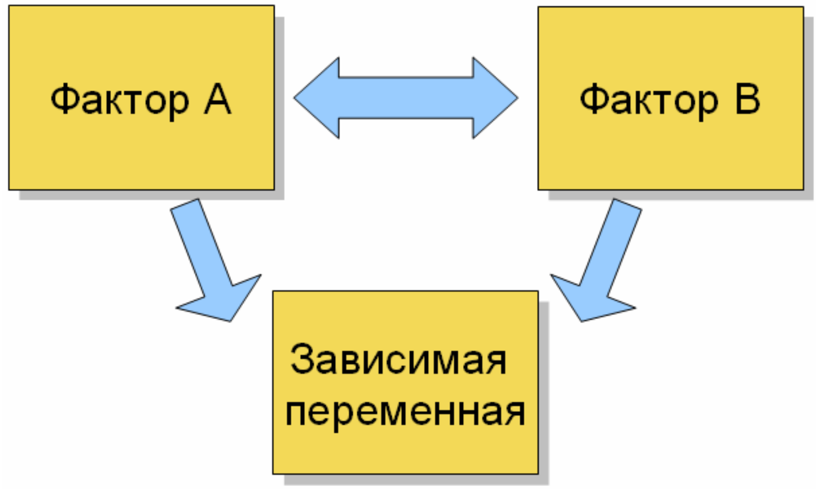


Рисунок 14-3. Двухфакторная модель дисперсионного анализа

Резюме

Рассмотренный нами однофакторный дисперсионный анализ позволяет проверять гипотезу о равенстве средних для трех и более генеральных совокупностей. Эти совокупности должны иметь нормальное распределение и их дисперсии должны быть равны. Идея метода состоит в сравнении внутригрупповой и межгрупповой дисперсии. Если эти дисперсии не отличаются сильно, то можно сделать вывод о том, что значимых различий между средними не выявлено. Проверка проводится при помощи F-критерия, который мы уже использовали для сравнения дисперсий. Решение задачи состоит из 9 шагов, из которых самым трудоемким является вычисление внутригрупповой и межгрупповой сумм квадратов.

14-2 Двухфакторный дисперсионный анализ

Двухфакторным дисперсионным анализом называют метод, проверяющий влияние двух независимых переменных (факторов) на зависимую переменную. Кроме этого, исследуется эффект взаимодействия между двумя независимыми переменными.

В качестве первого примера будем использовать данные, представленные в таблице 14-5. Компания-производитель желает проверить эффективность различных видов рекламы. Для рекламируемого продукта созданы два типа рекламных роликов: серьезный и смешной. Ролики размещены в рабочие и выходные дни.

Таблица 14-5. Влияние типа ролика и дня на эффективность рекламы

Тип ролика	Рабочий день	Выходной день
Смешной	6, 10, 11, 9	15, 18, 14, 16
Серьезный	8, 13, 12, 10	19, 20, 13, 17

Выбраны 16 потенциальных покупателей и случайным образом распределены на 4 группы. После того, как каждый покупатель просмотрел ролик, его просят оценить рекламу по шкале из 20 баллов. Различные баллы даются за привлекательность, ясность, краткость ролика и так далее. При $\alpha = 0{,}01$ требуется проанализировать данные и сделать выводы.

Исследуемые группы называют *эффектами обработки* (treatment groups), в данном случае их четыре:

Группа 1: Смешной ролик, рабочий день
Группа 2: Смешной ролик, выходной день
Группа 3: Серьезный ролик, рабочий день
Группа 4: Серьезный ролик, выходной день

Описание метода и пример

В рассматриваемом примере каждая независимая переменная имеет два *уровня*, или два варианта обработки. Таблица данных имеет размер 2×2., всего четыре клетки. Метод применим и к другим схемам, например, 3 x 2 или 3 x 3.

Для применения метода необходимо выполнение нескольких **условий**. (1) Генеральные совокупности, из которых извлечены выборки, имеют нормальное распределение. (2) Выборки независимы. (3) Дисперсии генеральных совокупностей равны. (4) Выборки (группы) имеют одинаковый объем.

Прежде всего, мы будем проверять гипотезу об *эффекте взаимодействия* между двумя переменными.

H_0: Тип ролика и день не имеют эффекта взаимодействия на эффективность рекламы
H_1: Тип ролика и день имеют эффект взаимодействия на эффективность рекламы

Наличие значимого эффекта будет означать, что тип ролика по-разному влияет на эффективность рекламы в зависимости от типа дня. Если в результате проверки будет выявлено значимое взаимодействие, другие гипотезы не проверяются. Если взаимодействия не выявлено, можно проверить две нулевых гипотезы, по одной для каждой независимой переменной. В нашем примере эти гипотезы будет сформулированы следующим образом:

H_0: Эффективность рекламы не зависит от типа дня
H_1: Эффективность рекламы зависит от типа дня

H_0: Эффективность рекламы не зависит от типа ролика
H_1: Эффективность рекламы зависит от типа ролика

Опишем последовательность проверки указанных гипотез и одновременно решим учебную задачу про рекламные ролики.

ШАГ 1. *Формулируем три вида гипотез* так, как это сделано выше: гипотезы о взаимодействии, гипотезы о типе ролика и гипотезы о типе дня.

ШАГ 2. *Задаем уровень значимости*, например, α=0,05.

ШАГ 3. *Находим степени свободы* для каждого фактора и для взаимодействия:

Фактор A:	df.N = $a - 1$
Фактор B:	df.N = $b - 1$
Взаимодействие ($A\times B$):	df.N = $(a - 1)(b - 1)$
Ошибка:	df.D = $ab\,(n - 1)$

где a – количество уровней фактора A,
b – количество уровней фактора B,
n – число наблюдений в каждой группе (клетке).

В нашем примере каждая независимая переменная, или фактор, имеет два уровня (принимает два значения). Фактор A – тип дня: рабочий или выходной, $a = 2$. Фактор B – тип ролика: смешной или серьезный, $b = 2$, число наблюдений в группе $n = 4$. Степени свободы:

Таблица 14-6. Расчетная таблица для нахождения сумм квадратов

ТИП РОЛИКА	РАБОЧИЙ ДЕНЬ	СУММА	ВЫХОДНОЙ ДЕНЬ	СУММА	ИТОГО
Смешной	6 10 11 9	36	15 18 14 16	63	99
Серьезный	8 13 12 10	43	19 20 13 17	69	112
ИТОГО		79		132	211

Фактор A:	df.N = 2 – 1 = 1
Фактор B:	df.N = 2 – 1 = 1
Взаимодействие ($A\times B$):	df.N = (2 – 1)(2 – 1) = 1×1 = 1
Ошибка	df.D = ab (n – 1) = 2×2×(4 – 1) = 12

ШАГ 4. По уровню значимости и числу степеней свободы *находим критические значения* в таблице А-5:

FA	$\alpha = 0{,}05$	df.N = 1	df.D = 12	FA = 4,75
FB	$\alpha = 0{,}05$	df.N = 1	df.D = 12	FB = 4,75
FAxB	$\alpha = 0{,}05$	df.N = 1	df.D = 12	FAxB =4,75

Критическая область запишется в виде F > 4,75.

ШАГ 5. *Вычисляем суммы квадратов*. Для этого строим вспомогательную таблицу 14-6.

Суммы в клетках T_{cell}	36, 43, 63, 69
Суммы по столбцам T_{row}	79, 132
Суммы по строкам T_{column}	99, 112
Общая сумма T	211
Общее число наблюдений N	16

Сумма квадратов всех значений в таблице:

$$\sum x^2 = 6^2 + 10^2 + 11^2 + 9^2 + 15^2 + ... + 13^2 + 17^2 = 3035$$

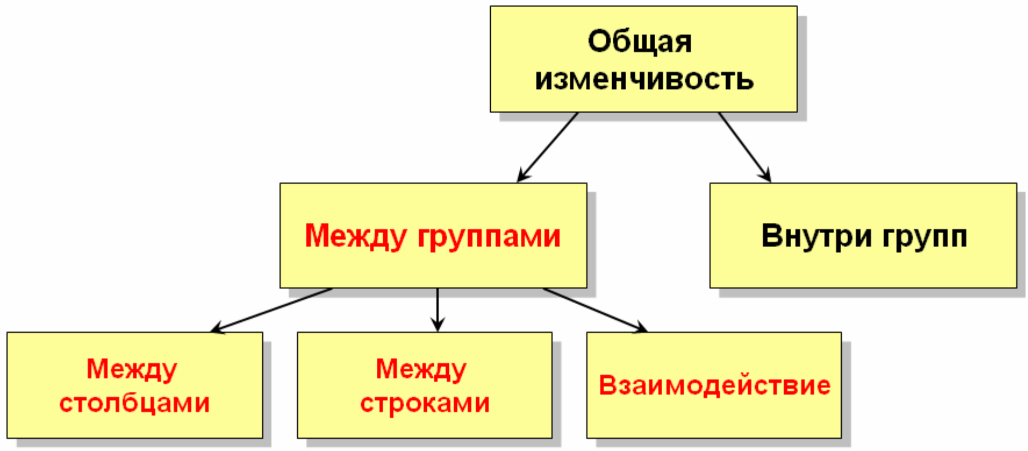


Рисунок 14-4. Источники изменчивости в модели двухфакторного анализа

Формулы для вычисления сумм квадратов:

$$SS_{total} = \sum x^2 - \frac{T^2}{N}$$

$$SS_{between} = \sum \frac{T_{cell}^2}{n} - \frac{T^2}{N}$$

$$SS_{within} = \sum x^2 - \sum \frac{T_{cell}^2}{n}$$

$$SS_{column} = \sum \frac{T_{column}^2}{n_{column}} - \frac{T^2}{N}$$

$$SS_{row} = \sum \frac{T_{row}^2}{n_{row}} - \frac{T^2}{N}$$

$$SS_{\text{int}eraction} = SS_{between} - \left(SS_{column} + SS_{row}\right)$$

Мы будем обозначать сумму квадратов по столбцам SS_A, сумму квадратов по строкам SS_B, сумму квадратов для взаимодействия факторов SS_{AxB}, внутригрупповую сумму квадратов SS_{error}:

$$SS_{column} = SS_A$$
$$SS_{row} = SS_B$$
$$SS_{\text{int}eraction} = SS_{A\times B}$$
$$SS_{within} = SS_{error}$$

Таблица 14-7. Вспомогательная таблица для вычисления F-значений

	Сумма квадратов	*df*	Среднее квадратичное	*F*
Фактор А	SS_A	$a-1$	MS_A	F_A
Фактор В	SS_B	$b-1$	MS_B	F_B
Взаимодействие, AxB	SS_{AxB}	$(a-1)(b-1)$	MS_{AxB}	F_{AxB}
Ошибка	SS_{error}	$ab(n-1)$	MS_{error}	
ИТОГО	…	…	…	

Такая двойная система обозначений выбрана, чтобы в первом случае указать на происхождение соответствующих сумм, а во втором на их содержательный смысл и назначение. Иногда во втором случае применяются обозначения с прямым указанием фактора, например, SS (день) или SS (ролик).

Для проверки правильности подсчета сумм квадратов следует использовать следующую формулу:

$$SS_{total} = SS_{column} + SS_{row} + SS_{interaction} + SS_{within}$$

Вычислим суммы квадратов по приведенным формулам для нашего примера

$$SS_{total} = 3035 - \frac{211^2}{16} = 252{,}44$$

$$SS_{between} = \left(\frac{36^2}{4} + \frac{43^2}{4} + \frac{63^2}{4} + \frac{69^2}{4}\right) - \frac{211^2}{16} = 186{,}19$$

$$SS_{within} = 3035 - \left(\frac{36^2}{4} + \frac{43^2}{4} + \frac{63^2}{4} + \frac{69^2}{4}\right) = 66{,}25$$

$$SS_{column} = \frac{79^2}{8} + \frac{132^2}{8} - \frac{211^2}{16} = 175{,}56$$

$$SS_{row} = \frac{99^2}{8} + \frac{112^2}{8} - \frac{211^2}{16} = 10{,}56$$

$$SS_{interaction} = 186{,}19 - (175{,}56 + 10{,}56) = 0{,}07$$

Таблица 14-8. Вспомогательная таблица для вычисления F-значений

	Сумма квадратов	df	Среднее квадратичное	F
День	175,56	1	175,56	31,8
Ролик	10,56	1	10,56	1,9
Взаимодействие	0,07	1	0,07	0,01
Ошибка	66,25	12	5,52	
Итого	252,44	15		

Проверка:

$$SS_{total} = SS_{column} + SS_{row} + SS_{\text{int}\,eraction} + SS_{within}$$
$$252{,}44 = 175{,}56 + 10{,}56 + 0{,}07 + 66{,}25$$
$$252{,}44 = 252{,}44$$

ШАГ 6. Строим вспомогательную расчетную таблицу (таблица 14-8) и *вычисляем F-значения*. Общие формулы для вычислений средних квадратичных:

$$MS_A = \frac{SS_A}{a-1}$$
$$MS_B = \frac{SS_B}{b-1}$$
$$MS_{A\times B} = \frac{SS_{A\times B}}{(a-1)(b-1)}$$
$$MS_{error} = \frac{SS_{error}}{ab(n-1)}$$

Общие формулы для подсчета F-значений:

$$F_A = \frac{MS_A}{MS_{error}}$$
$$F_B = \frac{MS_B}{MS_{error}}$$
$$F_{A\times B} = \frac{MS_{A\times B}}{MS_{error}}$$

Мы не приводим детальных вычислений. Размещаем полученные результаты в таблице. Таблица составлена очень удобно, поскольку в точности соответствует логике и последовательности вычислений F-значений на основе полученных предварительно сумм квадратов. На рисунке 14-5 приведен аналогичный отчет, полученный в SPSS.

ШАГ 7. *Сравниваем полученные F-значения с критической областью.* Начинаем проверку с гипотезы о взаимодействии факторов. Затем, если не будет отвергнута нулевая гипотеза об отсутствии взаимодействия, проверяем последовательно две другие нулевые гипотезы.

В нашем примере в критическую область попало F-значение для типа дня, а два других не попали. Принимаем нулевые гипотезы об отсутствии взаимодействия между факторами и об отсутствии влияния типа ролика. Отклоняем нулевую гипотезу об отсутствии влияния типа дня.

ШАГ 8. *Формулируем ответ.* Взаимодействие между типом ролика и типом дня отсутствует. При анализе влияния двух факторов на эффективность рекламы выявлено, что тип дня оказывает существенное влияние, а тип ролика - нет. Заключения сделаны на уровне 5% после анализа данных, представленных в условии задачи.

Tests of Between-Subjects Effects

Dependent Variable: GRADE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	2968,750[a]	4	742,188	134,434	,000
DAY	175,563	1	175,563	31,800	,000
TRAILER	10,563	1	10,563	1,913	,192
DAY * TRAILER	6,250E-02	1	6,250E-02	,011	,917
Error	66,250	12	5,521		
Total	3035,000	16			

a. R Squared = ,978 (Adjusted R Squared = ,971)

Рисунок 14-5. Отчет, полученный при вычислении F-значений в SPSS

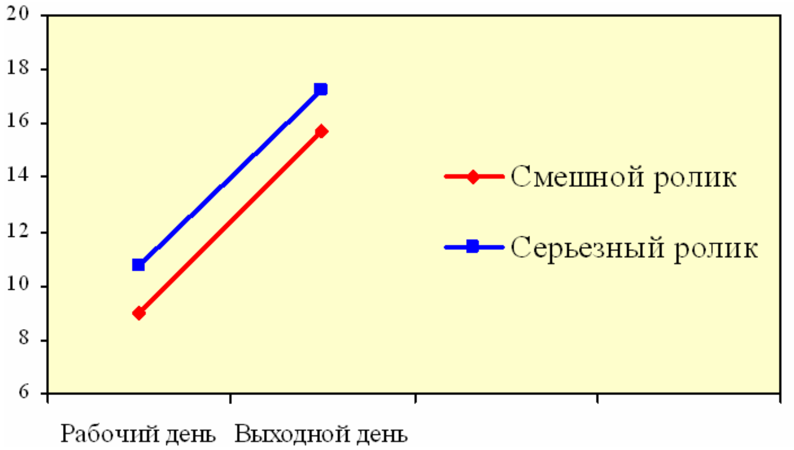


Рисунок 14-6. Отрезки почти параллельны, что означает отсутствие взаимодействия между факторами

Предварительный анализ взаимодействия

Существует визуальный метод для предварительного анализа наличия взаимодействия между факторами, который мы сейчас представим. Поскольку проверка взаимодействия проводится до проверки гипотез о влиянии каждой независимой переменной, при визуальном выявлении взаимодействия дальнейших вычислений можно уже не делать.

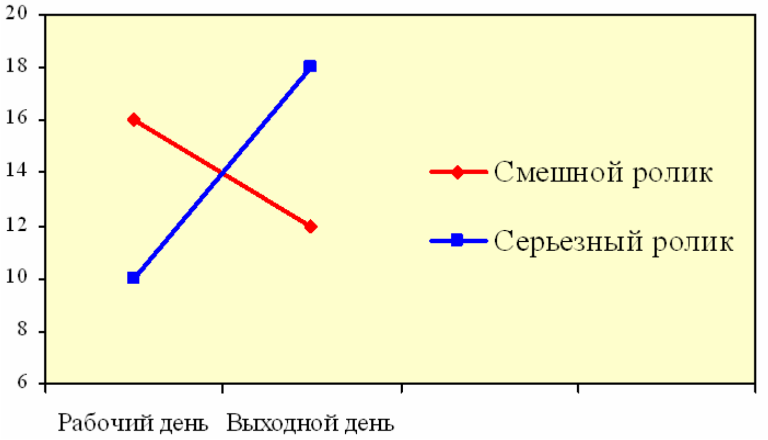
Для начала требуется вычислить средние для каждой группы:

Группа 1: Смешной ролик, рабочий день	36/4 = 9
Группа 2: Смешной ролик, выходной день	63/4 = 15,75
Группа 3: Серьезный ролик, рабочий день	43/4 = 10,75
Группа 4: Серьезный ролик, выходной день	69/4 = 17,25

Затем для каждого уровня фактора *B* (тип ролика) построить отрезок, как показано на рисунке 14-6. Если отрезки почти параллельны, это означает *отсутствие взаимодействия* между факторами. В этом случае мы уверены, что имеет смысл анализировать влияние каждого фактора по отдельности.

Если отрезки пересекаются (рисунок 14-7 а), такое взаимодействие называется *беспорядочным*. В случае беспорядочного взаимодействия не имеет смысла интерпретировать основные эффекты без учета эффекта взаимодействия. В случае, если прямые не пересекаются (рисунок 14-7 б), а значение *F*-критерия для взаимодействия оказывается значимым, тогда взаимодействие называется *порядковым*. В этом случае основные эффекты можно интерпретировать отдельно друг от друга.

а) Прямые пересекаются, взаимодействие между факторами имеется



б) Прямые не параллельны, но и не пересекаются. Требуется счетная проверка гипотезы о значимости взаимодействия между факторами

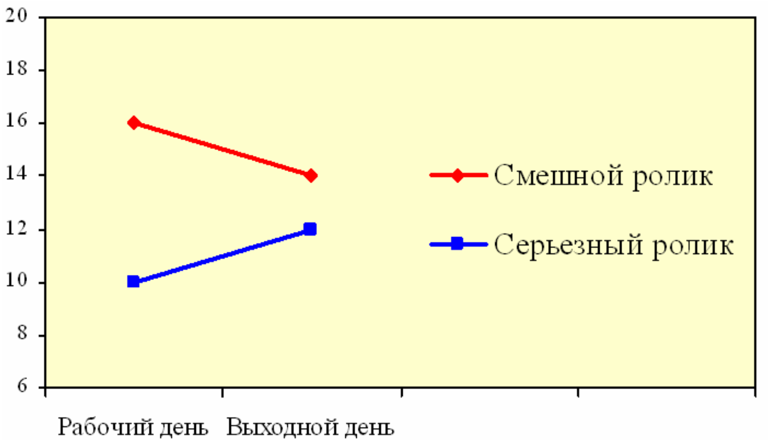


Рисунок 14-7. Две другие возможные ситуации

Резюме

Мы изучили модель двухфакторного дисперсионного анализа, который позволяет исследовать влияние двух независимых переменных на зависимую. Проверка гипотез об отсутствии влияния проводится при помощи F-критерия. Метод позволяет проверять также наличие взаимодействия между независимыми переменными. Вычисления для метода являются громоздкими для ручного счета и, как правило, выполняются при помощи компьютера.

Используем компьютер

Модели однофакторного и двухфакторного анализа довольно просты для понимания, но чрезвычайно сложны для ручного счета. Для решения задач необходимо пользоваться статистическими пакетами. В случае двухфакторного анализа следует хорошо понимать модель, которую вы намерены использовать, поскольку в статистических пакетах имеется несколько моделей, в то время как мы рассмотрели лишь одну из них. В меню SPSS она соответствует разделу Analyze → General Linear Model → Univariate…

Что означают термины

Дисперсионный анализ	**Уровни фактора**	**Общая сумма квадратов отклонений**
Однофакторный дисперсионный анализ	**Межгрупповая (факторная) дисперсия**	**Межгрупповая сумма квадратов отклонений**
Двухфакторный дисперсионный анализ	**Внутригрупповая (остаточная) дисперсия**	**Внутригрупповая сумма квадратов отклонений**
	Эффекты обработки	**Беспорядочное взаимодействие**
	Эффект взаимодействия	**Порядковое взаимодействие**

Символы и формулы

$SS_b = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$	Межгрупповая сумма квадратов отклонений
$SS_w = \sum (x - \bar{x}_i)^2$	Внутригрупповая сумма квадратов отклонений
$SS = \sum (x - \bar{x})^2 = SS_b + SS_w$	Общая сумма квадратов отклонений
$MS_B = \dfrac{SS_B}{k-1}$	Межгрупповая (факторная) дисперсия
$MS_W = \dfrac{SS_W}{N-k}$	Внутригрупповая (остаточная) дисперсия
$F = \dfrac{MS_B}{MS_W}$	F-критерий для проверки гипотезы

$df = k - 1$	Число степеней свободы числителя
$df = N - k$	Число степеней свободы знаменателя
T_{cell}	Суммы в клетках
T_{row}	Суммы по строкам
T_{column}	Сумма по столбцам
T	Сумма по таблице
N	Общее число наблюдений
$SS_{total} = \sum x^2 - \frac{T^2}{N}$	Сумма квадратов общая
$SS_{between} = \sum \frac{T^2_{cell}}{n} - \frac{T^2}{N}$	Сумма квадратов для клеток
$SS_{within} = \sum x^2 - \sum \frac{T^2_{cell}}{n}$ $SS_{within} = SS_{error}$	Сумма квадратов для ошибки
$SS_{column} = \sum \frac{T^2_{column}}{n_{column}} - \frac{T^2}{N}$ $SS_{column} = SS_A$	Сумма квадратов для столбцов
$SS_{row} = \sum \frac{T^2_{row}}{n_{row}} - \frac{T^2}{N}$ $SS_{row} = SS_B$	Сумма квадратов для строк
$SS_{interaction} = SS_{between} - \left(SS_{column} + SS_{row}\right)$ $SS_{interaction} = SS_{A\times B}$	Сумма квадратов для взаимодействия
$SS_{total} = SS_{column} + SS_{row} + SS_{interaction} + SS_{within}$	Равенство между суммами
$MS_A = \frac{SS_A}{a-1}$	Среднее квадратичное для фактора А
$MS_B = \frac{SS_B}{b-1}$	Среднее квадратичное для фактора В
$MS_{A\times B} = \frac{SS_{A\times B}}{(a-1)(b-1)}$	Среднее квадратичное для взаимодействия факторов
$MS_{error} = \frac{SS_{error}}{ab(n-1)}$	Среднее квадратичное для ошибки

$F_A = \dfrac{MS_A}{MS_{error}}$	Критерий для проверки влияния фактора А
$df.N = a - 1$	Число степеней свободы для фактора А
$F_B = \dfrac{MS_B}{MS_{error}}$	Критерий для проверки влияния фактора В
$df.N = b - 1$	Число степеней свободы для фактора В
$F_{A\times B} = \dfrac{MS_{A\times B}}{MS_{error}}$	Критерий для проверки влияния взаимодействия между факторами А и В
$df.N = (a - 1)(b - 1)$	Число степеней свободы для взаимодействия
$df.D = ab\,(n - 1)$	Число степеней свободы для ошибки

Задачи и упражнения

14-1. Диета для спортсменов. Исследователь хочет узнать, если ли разница в прибавке веса у спортсменов, следующих специальной диете. Спортсмены разделены на три группы случайным образом, каждая группа следует определенной диете 6 недель. Прибавка в весе указана в таблице. На уровне значимости $\alpha = 0{,}05$ определить, может ли исследователь утверждать, что есть разница в диетах?

Диета №1	Диета №2	Диета №3
3	10	8
6	12	3
7	11	2
4	14	5
	8	
	6	

14-2. Посудомоечные машины. Журнал для массового потребителя оценил посудомоечные машины как отличные, хорошие и средние. Исследователь хочет узнать, есть ли разница в цене на эти посудомоечные машины. При уровне значимости $\alpha = 0{,}10$ определить, существует ли значимая разница в средней цене на эти посудомоечные машины?

Данные на след странице.

Отличные	Хорошие	Средние
$565	$330	$350
400	840	379
369	510	280
550	470	320
460	380	
400	375	
400	450	
	290	
	319	

14-3. Вес газонокосилок. Исследователь хочет узнать, есть ли разница в весе четырех типов газонокосилок. При уровне значимости $\alpha = 0{,}10$ можно ли утверждать, что веса различны?

Газовая 1	Газовая 2	Электрич.	Ручная
95	73	55	37
101	69	52	24
108	72	51	25
107	71	37	29
97	67	57	22
101	62	54	17
	68	34	17
	71	45	22
		41	20
		53	18
			21

14-4. Высота домов. Высота домов в нескольких городах. При уровне значимости $\alpha = 0{,}01$ существует ли разница в средней высоте домов?

Сиэтл	Питтсбург	Н.Орлеан	Монреаль
943	841	697	669
740	725	645	640
735	635	531	630
722	616	530	624
609	615	481	612
605	582	478	590
580	535	450	498
574	520	442	428
569	511	439	355

(продолжение)

Сиэтл	Питсбург	Н.Орлеан	Монреаль
543	485	438	480
543	475	407	479
520	445		450
514	442		450
514	430		449
499	428		425
498	424		425
493	410		440
487			
480			
466			
456			
454			
409			

14-5. Уровень глюкозы в крови. Медработник хочет проверить влияние двух разных диет и двух типов упражнений на уровень глюкозы в крови. Уровень глюкозы измеряется в миллиграммах на децилитр (mg/dl). Для этого в каждую из групп были распределены по три человека.

Проанализируйте нижеприведенные данные, используя двухмерный дисперсионный анализ при $\alpha = 0{,}05$.

Упражнения	Диета А	Диета В
I	62, 64, 66	58, 62, 53
II	65, 68, 72	83, 85, 91

Расчет сумм квадратов уже выполнен:

Фактор	*SS*	df	*MS*	*F*
Упражнения	816,750			
Диета	102,083			
Взаимодействие	444,083			
Внутригрупповая	108			
Общая	1470,916			

Глава 15. Непараметрические критерии. Проверка однородности

При помощи непараметрических критериев проверяются гипотезы, которые не связаны с конкретным видом распределения, либо определенными значениями его параметров. Типичными непараметрическими критериями являются критерии для проверки однородности двух выборок, что означает совпадение законов распределения генеральных совокупностей, из которых они получены. Мы рассмотрим четыре критерия: знаковый, знако-ранговый, Манна-Уитни и Вилкоксона. Некоторые из них работают для независимых выборок, а другие для парных.

15-1 Непараметрические методы

Такие статистические критерии, как z, t и F называются параметрическими. *Параметрические критерии* предназначены для проверки гипотез о параметрах генеральной совокупности - среднем, дисперсии, доле признака; либо гипотез о типе распределения. Кроме этого, статистики разработали направление, которое развивает *непараметрические критерии.* В этом случае вид и параметры распределения не рассматриваются.

Очевидным *преимуществом* непараметрических критериев является возможность использования их для проверки гипотез, когда переменная не распределена нормально. Эти критерии применимы, в частности, для номинальных и порядковых данных. Они более понятны и просты, чем параметрические.

Таблица 15-1. Сравнение эффективности непараметрических методов

ПРИЛОЖЕНИЯ	ПАРАМЕТРИ-ЧЕСКИЙ ТЕСТ	НЕПАРАМЕТРИ-ЧЕСКИЙ ТЕСТ	ЭФФЕКТИВ-НОСТЬ
Парные выборки	t-тест или z-тест	Критерий знаков Знако-ранговый критерий	0,63 0,95
Две независимые выборки	t-тест или z-тест	Критерий Вилкоксона	0,95
Несколько независимых выборок	Дисперсионный анализ (F-тест)	Критерий Краскела-Уоллиса	0,95
Корреляция	Линейная корреляция	Ранговая корреляция	0,91

Следует назвать также *три недостатка* непараметрических критериев. Во-первых, они менее точны, чем соответствующие параметрические методы. Как следствие, требуются значительные отклонения, чтобы отвергнуть нулевую гипотезу.

Во-вторых, они менее информативны, чем параметрические критерии. Например, критерий знаков позволяет исследователю определить, превосходит значение данных медиану или нет, но не отвечает – насколько именно превосходит.

В-третьих, они менее эффективны, чем соответствующие параметрические критерии. Например, непараметрический критерий знаков дает лишь 60% эффективности от того, что можно получить, используя его параметрическое соответствие – z-критерий. Требуется больший объем выборки, чтобы компенсировать утрату информации: нужна выборка из 100 человек для критерия знаков, в то время, как для аналогичных результатов при использовании z-критерия достаточно было бы выборки из 60 человек. Сопоставление разных методов: параметрических и непараметрических, приводится в таблице 15-1. Эффективность непараметрических тестов оценивалась в сравнении с параметрическими для нормально распределенной генеральной совокупности.

Поскольку, непараметрические критерии широко используются, мы рассмотрим последовательно критерии проверки однородности (глава 15), критерии ранговой корреляции (глава 16) и, наконец, факторный непараметрический анализ (глава 17).

15-2 Критерий знаков

Рассмотрим для примера три задачи. Каждая из них позднее будет использована для иллюстрации трех различных применений критерия знаков.

Задача 1. Строительство башни. Несколько детей попросили из предоставленных им кубиков собрать башню. Эксперимент повторили с этими же детьми через месяц, результаты времени (в секундах) представлены в таблице ниже. На уровне значимости α=0,05 проверить предположение о том, что *нет существенной разницы между результатами.*

Ребенок	A	B	C	D	E	F	G	I	K	L	M	N	O	P	R
Испытание 1	30	19	19	23	29	44	42	20	12	39	14	81	17	31	52
Испытание 2	30	13	14	16	14	52	14	22	17	12	11	30	14	17	15

Задача 2. В день 40 леденцов. Владелец продуктового магазина высказал гипотезу о том, что медианное количество продаваемых им за день леденцов равно 40. Случайная выборка за 20 дней дает следующие данные по количеству леденцов, продаваемых каждый день. При *a* = 0,05 *проверить гипотезу владельца магазина.*

18	43	40	16	22
30	29	32	37	36
39	34	39	45	28
36	40	34	39	52

Задача 3. Кадровые предпочтения. Руководство сети ресторанов быстрого обслуживания обратило внимание, что кадровая служба при приеме на работу на должность менеджера отдает большее предпочтение девушкам, нежели, чем юношам. Среди менеджеров оказалось 30 юношей и 70 девушек. Усомнившись в разумности сложившихся пропорций, руководство запросило объяснений. Кадровая служба объяснила сложившуюся пропорцию результатом случайности. Требуется проверить на уровне значимости α=0,05, *может ли такая пропорция оказаться результатом случайности.*

Нам предстоит изучить применение критерия для трех типов гипотез и в качестве примера решить эти задачи.

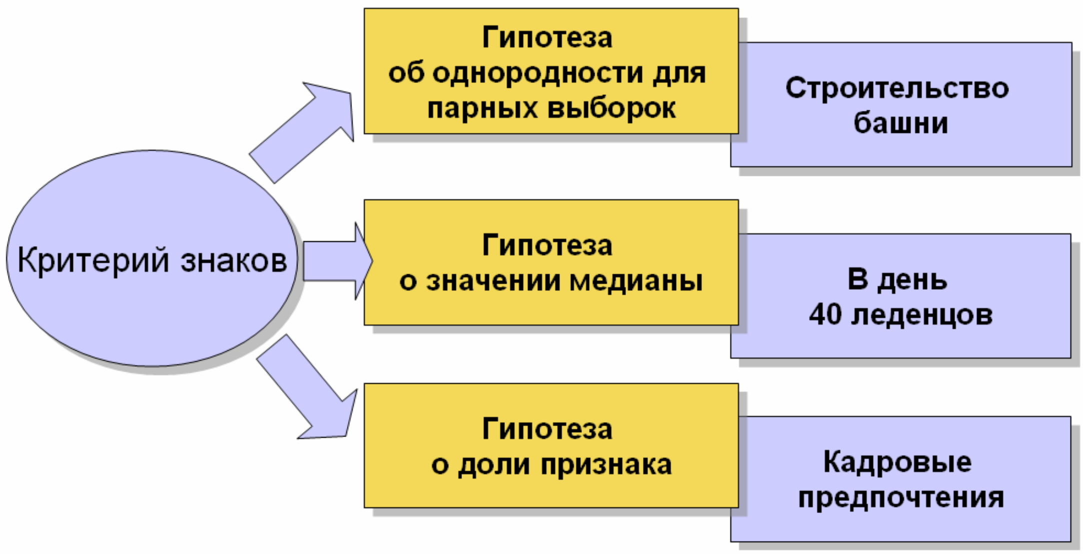


Рисунок 15-1. Три варианта гипотез для критерия знаков

Гипотеза об однородности для парных выборок

Суть критерия состоит в следующем. Предположим, мы имеем парные выборки, включающие n чисел:

$$\begin{array}{ll} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_n & y_n \end{array}$$

Если эти пары получены случайным образом из двух генеральных совокупностей, то мы можем проверить гипотезу об однородности – предположение о совпадении законов распределения этих совокупностей. В этом случае можно сказать, что выборки являются однородными, получены из одной и той же генеральной совокупности.

Гипотеза об однородности есть предположение о совпадении законов распределения двух генеральных совокупностей из которых получены выборки. Если распределения одинаковы, то **выборки однородны**.

Итак, проверяемые основная и альтернативная гипотезы:

H_0: выборки однородны
H_1: выборки не однородны

Метод состоит в следующем. Вычтем в каждой паре одно число из другого. Результат может быть положительным или отрицательным, в зависимости от того, какое число больше.

Для каждой пары проставим знак разности и затем подсчитаем количество плюсов и количество минусов. Пары, в которых числа совпали, отбрасываем. Если выборки однородны, то количество плюсов должно быть приблизительно равно количеству минусов. Это следует из того, что в этом случае плюс или минус мы получаем с равными вероятностями:

$$P(\text{"}плюс\text{"}) = P(\text{"}минус\text{"}) = 0,5$$

Если в результате эксперимента плюсов будет существенно больше (или существенно меньше), чем минусов, мы сделаем вывод о неоднородности выборок. Это будет означать, что распределение первой генеральной совокупности отличается от распределения второй.

Поскольку в основе критерия лежит нахождение и анализ количества знаков при сравнении чисел, он и получил название *критерия знаков*. Для применения критерия требуется только, чтобы выборки были получены случайным образом. При этом нет никаких требований относительно закона распределения генеральных совокупностей, из которых эти данные получены.

Ситуация 1. Объем выборки $n \le 25$. В качестве статистики (критерия) для проверки гипотезы мы выбираем минимальное число из количества плюсов и количества минусов:

$$x = \min \text{(количество минусов, количество плюсов)}$$

Критические значения находятся по таблице А-6. Если x окажется меньше или равен критическому значению из таблицы, то гипотеза отвергается. Следует заметить, что в качестве статистики можно выбрать и другую случайную функцию, например, количество плюсов. Тогда критическая область будет двусторонней. Слишком большое или слишком малое количество плюсов будет означать отклонение нулевой гипотезы.

Ситуация 2. Объем выборки $n > 25$. В этом случае в качестве статистики выбираем:

$$z = \frac{(x + 0,5) - (n/2)}{\sqrt{n}/2}$$

Таблица 15-2. Проверка однородности критерием знаков

Ребенок	Испытание 1	Испытание 2	Знак
A	30	30	
B	19	13	+
C	19	14	+
D	23	16	+
E	29	14	+
F	64	52	+
G	42	14	+
I	20	22	-
K	12	17	-
L	39	12	+
M	14	11	+
N	81	30	+
O	17	14	+
P	31	17	+
R	52	15	+

Это возможно, поскольку при возрастании объема выборки распределение количества плюсов (и минусов) стремится к нормальному закону. Критические z-значения находятся по таблице А-2.

Решение задачи 1. Составляем таблицу 15-2 и подсчитываем количество плюсов, минусов. Отбрасывая одно совпадение, получили, что объем выборок равен 14.

ШАГ 1. Формулируем гипотезы:

H_0: существенных изменений не произошло
H_1: произошли существенные изменения

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. По таблице А-6 находим для n=14 критическое значение 2. Критическая область $x \leq 2$.

ШАГ 4. Находим значение статистики. Получили 12 плюсов, 2 минуса и 1 совпадение. Совпадение отбрасываем. Статистика:

$$x = \min(2,12) = 2$$

18	43	40	16	22
30	29	32	37	36
39	34	39	45	28
36	40	34	39	52

Данные

-	+	0	-	-
-	-	-	-	-
-	-	-	+	-
-	0	-	-	+

3 плюса, 15 минусов, 2 совпадения

Рисунок 15-2. Проверка про 40 леденцов (медиана = 40)

ШАГ 5. Сравниваем. Статистика попала в критическую область.

ШАГ 6. Пишем ответ. Выборки неоднородны. Это означает, что результаты во втором испытании отличаются существенно на уровне значимости 5%.

Гипотеза о значении медианы

Критерий знаков применяется для проверки гипотезы о значении медианы. Метод идентичен предыдущему критерию. Как и прежде, в качестве статистики используется минимальное число из количества плюсов и минусов.

Решение задачи 2. Решаем задачу про сорок леденцов (рисунок 15-2).

ШАГ 1. Формулируем гипотезы:

H_0: медиана равна 40
H_1: медиана значимо отличается от 40

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. По таблице А-6 находим для n =18 критическое значение 4. Критическая область $x \leq 4$.

ШАГ 4. Находим значение статистики. Получили 3 плюса и 15 минусов, 2 совпадения отбросили. Статистика равна:

$$x = \min(3, 15) = 3$$

ШАГ 5. Сравниваем значение статистики с критической областью. Статистика попала в критическую область.

ШАГ 6. Пишем ответ. У нас достаточно оснований, чтобы отказаться от заявления, о том, что медиана продаваемых в день леденцов равна 40.

Гипотеза о доле признака

Критерий знаков применяется также для проверки гипотезы о доле признака в генеральной совокупности. Его использование не отличается от двух разобранных выше критериев. Выборка должна быть получена из генеральной совокупности случайным образом. В качестве статистки используется минимальная частота:

x = min (число успехов; n - число успехов)

Напомним, что успехом мы называем ситуацию, при которой признак принимает интересующее нас значение (долю которого мы рассматриваем).

Решение задачи 3. Решаем для примера задачу о кадровых предпочтениях.

ШАГ 1. Формулируем гипотезы:

H_0: доля юношей в генеральной совокупности 0,5
H_1: доля юношей значимо отличается от 0,5

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Находим критическое значение. Выборка составила 100 человек: $n = 100$. Поскольку $n > 25$, пользуемся нормальным приближением. Для α=0,05 находим z = - 1,96.

ШАГ 4. Находим значение статистики. Получили 12 плюсов, 2 минуса и 1 совпадение, которое отбрасываем. Статистика:

x = min (30, 70) = 30

$$z = \frac{(x + 0{,}5) - (n/2)}{\sqrt{n}/2} = \frac{(30 + 0{,}5) - (100/2)}{\sqrt{100}/2} = -3{,}90$$

ШАГ 5. Сравниваем. z-значение попало в критическую область.

ШАГ 6. Пишем ответ. Поскольку значение статистики попало в критическую область, мы отвергаем основную гипотезу и на уровне значимости 5% считаем, что кадровые предпочтения имеются.

15-2 Знако-ранговый критерий

Знако-ранговый критерий проверяет гипотезу об однородности для парных выборок. Предположим, нам требуется проверить, совпадают ли законы распределения генеральных совокупностей, из которых взяты две парные (зависимые) выборки. Часто такие задачи формулируются в терминах проверки эффекта обработки: проверяется совпадение распределений «до» и «после» обработки.

Для применения метода необходимо выполнение нескольких **условий**: (1) Исследуются парные (зависимые) выборки. (2) Данные должны быть получены случайным образом. (3) Генеральная совокупность разностей между значениями в паре должна иметь симметричное распределение, в том смысле, что правая часть графика является зеркальным отражением левой. При этом не требуется, чтобы данные имели нормальное распределение.

Метод заключается в следующем. Для каждой пары *(x, y)* вычисляются разности между значениями $d = x - y$. Затем этим разностям присваиваются ранги. Ранжирование происходит без учета знака. Затем отдельно суммируются ранги отрицательных разностей, отдельно положительных. Если выборки однородны, то эти суммы не могут сильно отличаться. На этом построена проверка. Наименьшая из сумм обозначается T. Статистика зависит от объема выборки.

Ситуация 1. Если $n \leq 30$, то применяется статистика:

T = min (сумма рангов разностей с плюсом, сумма рангов разностей с минусом)

Таблица 15-3. Проверка гипотезы об однородности при помощи знако-рангового критерия (задача о строительстве башни)

РЕБЕНОК	ИСП. 1	ИСП. 2	РАЗНОСТИ D	РАНГИ	РАНГИ МИНУС	РАНГИ ПЛЮС
A	30	30	0			
B	19	13	6	6		6
C	19	14	5	4,5		4,5
D	23	16	7	7		7
E	29	14	15	10		10
F	64	52	12	8		8
G	42	14	28	12		12
I	20	22	-2	1	1	
K	12	17	-5	4,5	4,5	
L	39	12	27	11		11
M	14	11	3	2,5		2,5
N	81	30	51	14		14
O	17	14	3	2,5		2,5
P	31	17	14	9		9
R	52	15	37	13		13
					Σ=5,5	Σ=99,5

Ситуация 2. Если $n > 30$, то тогда применяется статистика:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Критические значения в первом случае находятся по таблице А-7, во втором – по таблице А-2.

Решение задачи 1. Решим при помощи знако-рангового критерия задачу про детей и башню. Необходимые расчеты приведены в таблице 15-3.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: выборки однородны

H_1: выборки не однородны

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Как и прежде, объем выборки корректируется на число совпадений: $n = 15 - 1 = 14$. Поскольку $n \leq 30$, по таблице А-7 находим критическое значение, равное 21. Критическая область $T \leq 21$.

ШАГ 4. Вычисляем значение статистики. В четвертом столбце таблицы записываем разности, ранжируем их по абсолютной величине (игнорируя знаки), а затем складываем отдельно ранги отрицательных разностей, отдельно положительных. Получаем, что сумма отрицательных рангов равна 5,5, а сумма положительных 99,5. Критерий $T = \min(5{,}5;\ 99{,}5) = 5{,}5$.

ШАГ 5. Сравниваем. Т-значение попало в критическую область.

ШАГ 6. Поскольку значение статистики попало в критическую область, мы отклоняем основную гипотезу на уровне значимости 5%.

15-3 Критерий Манна-Уитни

Критерий Манна-Уитни проверяет гипотезу об однородности для двух независимых выборок: совпадают ли законы распределения генеральных совокупностей, из которых получены эти выборки.

Рассматриваются основная и альтернативная гипотезы:

H_0: выборки однородны
H_1: выборки не однородны

В критерии Манна-Уитни сравниваются все элементы первой выборки со всеми элементами второй. Всего есть $m \times n$ пар для сравнений. Отдельно подсчитывается число пар, в которых $x > y$, и отдельно, в которых $y > x$. Если значения в паре равны, она считается за 0,5. Минимальное из этих чисел есть U-статистика Манна-Уитни. Если выборки однородны, U-значение не должно быть слишком малым. Критическая область $U < U_{\text{критич}}$.

Подсчет числа сравнения удобно выполнять на компьютере, но мы, тем не менее, приведем удобный способ ручного подсчета на примере следующей учебной задачи.

Таблица 15-4. Сравнение длины клюва самок двух разных пород при помощи критерия Манна-Уитни

ПОРОДА А X	СЧЕТ ПАР ДЛЯ А	ПОРОДА В Y	СЧЕТ ПАР ДЛЯ В
1,5	1	1,3	0
2,1	1	2,4	3
2,3	1	2,5	3
2,7	3	2,8	4
3,4	5	3,2	4
3,5	5,5	3,5	5,5
4,1	6	4,2	7
4,6	8,5	4,5	7
6,3	9	4,6	7,5
6,5	9		
7,1	9		
n = 11	Σ=58	n = 9	Σ=41

Пример. Длина клюва. Орнитолог заинтересовался длиной клюва у самок двух определенных пород. Данные, которые им были получены:

Порода А	Порода В
2,3	1,3
3,5	2,4
4,6	4,5
2,1	3,2
3,4	2,5
6,3	4,2
1,5	3,5
2,7	4,6
6,5	2,8
4,1	
7,1	

Его желание состоит в том, чтобы убедиться в отсутствии значимых различий в длине клюва самок этих пород. Как мы видим, в критерии Манна-Уитни не требуется, чтобы выборки имели одинаковый объем.

Решение. Для проверки гипотезы строим вспомогательную таблицу 15-4.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: выборки однородны
H_1: выборки не однородны

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Используем для этого таблицы критических значений Манна-Уитни (таблицы А-8). Объемы выборок 11 и 9. Критическое значение равно 23. Критическая область $T \leq 23$.

ШАГ 4. Вычисляем значение статистики при помощи вспомогательной таблицы. В ней обе выборки для удобства подсчетов упорядочены в порядке возрастания. Во втором столбце ведем для каждого значения x подсчет числа значений y, которые меньше его. Например, для числа 1,5 только одно значение 1,3 меньше его, а для числа 2,1 тоже одно и так далее. Заполнив второй и третий столбец, мы получаем суммы 58 и 41 соответственно. Критерий U = min (58; 41) = 41.

ШАГ 5. Сравниваем. U-значение не попало в критическую область.

ШАГ 6. Поскольку значение статистики не попало в критическую область, мы не выявили значимых различий в длине клюва самок двух разных пород.

15-4 Критерий Вилкоксона

Критерий Вилкоксона проверяет гипотезу об однородности для двух независимых выборок: совпадают ли законы распределения генеральных совокупностей, из которых взяты эти выборки. Последовательность проверки гипотезы об однородности следующая.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: выборки однородны
H_1: выборки не однородны

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Используем для этого таблицы критических значений Вилкоксона (таблицы А-9) или таблицы А-2, в зависимости от объема выборок. Для применения таблиц нормального закона достаточно, чтобы каждая выборка имела объем больше 10.

ШАГ 4. Вычисляем значение статистики при помощи вспомогательной таблицы. Ранжируем обе выборки так, как если бы мы их перемешали. Затем находим сумму рангов для первой выборки и сумму рангов для второй. Если выборки однородны, то эти суммы не должны сильно отличаться друг от друга.

Если объем выборок меньше 10, то критерием в этом случае является W = {меньшая сумма рангов}.

Если объем каждой выборки больше 10, то в качестве критерия будем использовать следующую случайную функцию:

$$z = \frac{R - \mu_R}{\sigma_R},$$

где R - есть меньшая из полученных сумм рангов,

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} \text{ -}$$

есть среднее значение R, при условии, что две генеральные совокупности имеют одинаковый закон распределения,

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \text{ -}$$

есть стандартное отклонение R, при условии, что две генеральные совокупности имеют одинаковый закон распределения.

ШАГ 5. Далее, как обычно, сравниваем значение статистики с критической областью и делаем вывод.

ШАГ 6. Пишем ответ.

Пример. Пьер или Гарри. Студенты решили сравнить сложность чтения текстов двух известных произведений: «Гарри Поттера» и «Войны и мира». Сложность текста измерялась по 100-бальной шкале. Вот результаты:

J.K.Rowling	Лев Толстой
85,3	69,4
84,3	64,2
79,5	71,4
82,5	71,6
80,2	68,5
84,6	51,9
79,2	72,2
70,9	74,4
78,6	52,8
86,2	58,4
74,0	65,4
83,7	73,6
71,4	

Требуется проверить гипотезу об однородности двух независимых выборок. Тем самым, мы хотим получить ответ на вопрос, можно ли считать, что простота (или сложность) чтения одинакова для произведений двух исследуемых писателей?

Составляем вспомогательную расчетную таблицу 15-5. В первом и третьем столбце расположены результаты измерений – две выборки объемом 13 и 12 измерений соответственно. Во втором и четвертом столбце проставлены ранги каждого значения, при этом ранжирование производилось, как будто бы обе выборки мы перемешали. Сумма рангов первой выборки составила 236,5, второй – 88,5.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: книги не имеют отличий в сложности текста
H_1: книги отличаются сложностью текста

Таблица 15-5. Сравнение простоты чтения «Гарри Потера» и «Войны и мира» при помощи критерия Вилкоксона

J.K.ROWLING	РАНГИ	ЛЕВ ТОЛСТОЙ	РАНГИ
85,3	24	69,4	7
84,3	22	64,2	4
79,5	18	71,4	9,5
82,5	20	71,6	11
80,2	19	68,5	6
84,6	23	51,9	1
79,2	17	72,2	12
70,9	8	74,4	15
78,6	16	52,8	2
86,2	25	58,4	3
74,0	14	65,4	5
83,7	21	73,6	13
71,4	9,5		
Всего 13	Σ=236,5	Всего 12	Σ=88,5

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Поскольку наши выборки имеют объем больше 10 каждая, используем таблицы А-2. Критическая область односторонняя и при α=0,05 критическая точка z = - 1,96.

ШАГ 4. Вычисляем значение статистики при помощи вспомогательной таблицы. Меньшая сумма рангов равна 88,5. Вычисляем среднее и стандартное отклонение суммы рангов *R*:

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{13(13+12+1)}{2} = 169$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{13 \cdot 12(13+12+1)}{12}} = 18{,}385$$

Ranks

	AUTOR	N	Mean Rank	Sum of Ranks
READING	1	13	18,19	236,50
	2	12	7,38	88,50
	Total	25		

Test Statistics[b]

	READING
Mann-Whitney U	10,500
Wilcoxon W	88,500
Z	-3,672
Asymp. Sig. (2-tailed)	,000
Exact Sig. [2*(1-tailed Sig.)]	,000[a]

Рисунок 15-3. Проверка гипотезы о сложности текстов двух книг

Значение статистики равно:

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{88{,}5 - 169}{18{,}385} = -3{,}672$$

ШАГ 5. Сравниваем значение статистики с критической областью. Полученное нами значение попадает в критическую область. Основную гипотезу следует отклонить.

ШАГ 6. Пишем ответ. Выборки не однородны, получены из разных генеральных совокупностей. Сложность текстов существенно различается.

Резюме

Критерий Вилкоксона для независимых выборок завершает главу, в которой мы рассмотрели четыре критерия проверки однородности – два для независимых и два для парных выборок. Все эти критерии являются непараметрическими. Их использование не предполагает знание типа распределения и не связано с изучением параметров генеральной совокупности. Мы всего лишь отвечали на вопрос, одинаково ли распределение генеральных совокупностей, из которых взяты выборки, или нет.

Используем компьютер

После изучения критериев проверки однородности и решения нескольких задач при помощи ручного счета, следует познакомиться с возможностями компьютера для проведения вычислений. Самым простым является применение электронных таблиц, которые могли бы полностью имитировать наши действия на листе бумаги. Тем не менее, удобнее использовать специальные статистические пакеты, которые имеют специальные разделы и функции для проверки однородности.

Что означают термины

Параметрические критерии	**Гипотеза об однородности выборок**	**Знако-ранговый критерий**
Непараметрические критерии	**Однородные выборки**	**Критерий Манна-Уитни**
Эффективность критерия	**Критерий знаков**	**Критерий Вилкоксона**

Символы и формулы

$x = \min$ *(количество минусов, количество плюсов)*	Критерий знаков (n ≤ 25)
$z = \dfrac{(x+0{,}5)-(n/2)}{\sqrt{n}/2}$	Критерий знаков (n > 25)
$T = \min$ *(сумма рангов разностей с плюсом, сумма рангов разностей с минусом)*	Знако-ранговый критерий (n ≤ 30)
$z = \dfrac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$	Знако-ранговый критерий (n > 30)
$U = \{$*количество пар при сравнениях значений* x *и* y *каждого с каждым, для которых* $x > y\}$	Критерий Манна-Уитни
$W = \{$ *меньшая сумма рангов при ранжировании всех значений двух выборок*$\}$	Критерий Вилкоксона (n ≤ 10)

$z = \dfrac{R - \mu_R}{\sigma_R}$	Критерий Вилкоксона (n < 10)
R	Меньшая из полученных сумм рангов
$\mu_R = \dfrac{n_1(n_1 + n_2 + 1)}{2}$	Среднее значение R
$\sigma_R = \sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$	Стандартное отклонение R

Задачи и упражнения

15-1. Кворум избирателей. При опросе 1002 человек 701 из них сказал, что собирается идти голосовать на выборах президента. Можно ли полагать, что большинство населения придет на выборы?

15-2. Аренда квартиры. Агент по продаже недвижимости предполагает, что средняя арендная плата за однокомнатную квартиру в городе составляет $325 в месяц. Выборка 12 однокомнатных квартир показала следующие месячные расценки. При α=0,05 достаточно ли у нас оснований, чтобы опровергнуть заявление агента по продаже недвижимости?

420	460	514	405
320	435	531	450
560	309	312	350

15-3. Один дома. Из 50 опрошенных студентов 29 предпочитали бы жить в общежитиях в одноместной комнате. При α=0,02 проверьте гипотезу о том, что более 50% студентов предпочитают жить в общежитиях в одиночку. Применить критерий знаков.

15-4. Повлияло ли лечение? Было проведено исследование, чтобы выяснить, повлияют ли новые диетические медикаменты на женщин, желающих сбросить вес. Вес 8 пациенток был измерен до лечения и через 6 недель ежедневного применения лечения. Данные приведены ниже. При α = 0,05 можно ли сделать вывод, что лечение повлияло (увеличило или уменьшило) на вес этих женщин? Применить критерий знаков.

	A	B	C	D	E	F	G	H
До	187	163	201	158	139	143	198	154
После	178	162	188	156	133	150	175	150

15-5. Изменения в отношениях. Восьми супружеским парам была предложена анкета на супружескую совместимость. После прохождения парами семинара, им дали вторую анкету, чтобы выяснить, произошли ли какие-либо изменения в их поведении по отношению друг к другу. Данные приведены ниже. При $\alpha = 0{,}10$ есть ли различия в результатах пар?

До	43	52	37	29	51	62	57	61
После	48	59	36	29	60	68	59	72

15-6. Мужчины и женщины. Для сравнения уровня заработной платы были отобраны (в соответствии со стажем) работники-мужчины и работники-женщины. В таблице ниже приведены получившиеся данные (в тысячах рублей). При $\alpha = 0{,}10$ есть ли различие в зарплатах мужчин и женщин?

Мужчины	18	43	32	27	15	45	21	22
Женщины	16	38	35	29	15	46	25	28

15-7. Однородность выборок. Проверить гипотезу об однородности следующих выборок:

Выборка 1	Выборка 2
2,19	1,98
2,26	2,31
2,28	2,25
2,21	1,91
1,87	2,15
2,34	2,12
2,14	1,83
1,93	2,13
2,10	2,11
1,89	2,16
2,36	1,99
2,17	

15-8. Читают ли замужние? Исследователь опросил замужних и одиноких женщин, чтобы проверить, есть ли разница в том, сколько книг те и другие прочитали в течение прошлого года. Данные приведены ниже. При α=0,10 проверьте заявление о том, что обе группы прочли одинаковое количество книг.

Замужние	6	8	7	4	9	12	13	7	10	18	15	
Одинокие	2	3	5	11	3	5	11	12	16	4	0	1

15-9. Кто больше любит свою работу? Двум группам рабочих дали вопросники, чтобы установить степень их удовлетворенности работой. Задавалась шкала диапазоном от 0 до 100. Группы делились по стажу: те, кто работал более 5 лет, и те, кто работал менее 5 лет. Данные приведены ниже. При α=0,10, проверьте заявление о том, что между удовлетворенностью работой двух групп нет разницы.

До 5	78	98	83	86	75	77	72	68	56	93	97	99	93
5 и более	94	79	82	85	73	66	64	59	52	58	63	68	88

15-10. Как работают одинокие? Инспекторам было поручено оценить продуктивность работы служащих. Исследователь хочет узнать, у кого она выше: у людей, живущих в браке, или у одиноких? Диапазон шкалы оценки продуктивности составляет от 1 до 50. Данные приведены ниже. При α=0,01 достаточно ли у нас оснований, чтобы подтвердить это заявление?

Одинокие	48	46	2	50	38	36	40	31	28	24	49	34
В браке	44	35	41	37	42	43	29	31	37	32	36	

Глава 16. Непараметрические критерии. Ранговая корреляция

В главе 13 мы рассматривали коэффициент связи для числовых (интервальных) данных. В этой главе нам предстоит познакомиться еще с двумя коэффициентами: Спирмена и Кендалла. Они применимы для порядковой шкалы, но построены на альтернативных подходах. Установление наличия связи, измерение силы связи и проверка гипотез о значимости связи при помощи коэффициента Спирмена и коэффициента Кендалла составляют содержание настоящей главы.

16-1 Коэффициент ранговой корреляции Спирмена

Изучение корреляции и регрессии в 13 главе относилось к интервальным переменным. Коэффициент корреляции Пирсона, который мы использовали для измерения силы и направления связи, применим для переменных, имеющих нормальный закон распределения. В ситуации, когда нет нормального распределения, или, более того, переменные измеряются не интервальной, а порядковой шкалой, необходимы другие методы для проверки наличия связи и измерения ее силы.

В этом параграфе мы рассмотрим коэффициент Спирмена. Этот коэффициент предполагает ранжирование каждого набора данных и сопоставление рангов. Отсюда его название – *коэффициент ранговой корреляции*. Метод основан на том, что при отсутствии связи ранги будут хорошо перемешаны и тогда коэффициент будет близок к нулю.

Таблица 16-1. Оценки двух экспертов на конкурсе красоты

ОЦЕНКИ ЭКСПЕРТА-МУЖЧИНЫ	ОЦЕНКИ ЭКСПЕРТА-ЖЕНЩИНЫ	РАЗНОСТИ	КВАДРАТЫ РАЗНОСТЕЙ
4	2	2	4
2	6	-4	16
5	7	-2	4
1	3	-2	4
3	1	2	4
6	10	-4	16
7	4	3	9
8	8	0	0
9	5	4	16
10	9	1	1
			Σ=74

Если для двух наборов данных ранги совпадают, коэффициент будет равен +1, если ранги противоположны (имеется обратная связь), то коэффициент примет значение -1.

Вычисление коэффициента

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$r = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)}$$

где d_i - разность между рангами значений в паре,
n - объем выборки (количество пар).

Будем изучать нахождение коэффициента корреляции Спирмена и проверку его значимости, параллельно рассматривая следующий пример.

Пример. Два эксперта на конкурсе красоты. Два эксперта: мужчина и женщина, оценивают моделей на конкурсе красоты и для этого проставили места для каждой модели. Данные размещены в двух первых столбцах таблицы 16-1.

Решение. Вычислим коэффициент корреляции Спирмена. Дополним таблицу двумя столбцами: вычислим разности и квадраты разностей между значениями в каждой паре. Сумма разностей составила 74. Теперь подставляем это значение в формулу для коэффициента:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\cdot 74}{10(10^2-1)} = 1 - \frac{444}{990} = 0{,}552$$

Полученное значение 0,552 не расположено возле нуля и единицы. Его можно интерпретировать следующим образом: связь в выборке выявлена на среднем уровне. Для получения выводов о наличии связи в генеральной совокупности требуется проверка значимости коэффициента.

Проверка значимости

Выборочный коэффициент корреляции Спирмена r_s позволяет оценивать коэффициент корреляции генеральной совокупности, обозначаемый ρ_s.

Покажем, как можно проверить гипотезу о равенстве нулю коэффициента ранговой корреляции генеральной совокупности на основании полученного значения коэффициента ранговой корреляции выборки. Проверяемые гипотезы:

H_0: $\rho_s = 0$ (связи нет)
H_1: $\rho_s \neq 0$ (связь между переменными значима)

Проверка зависит от объема выборки.

Ситуация 1. Если $n \leq 30$, то критические значения находятся по таблице А-10, которая составлена для критических значений коэффициента при объеме выборки от 5 до 30.

Ситуация 2. Если $n > 30$, то критические значения находятся по формуле:

$$r_s = \frac{\pm z}{\sqrt{n-1}}$$

Мы пользуемся в этом случае нормальным приближением и применяем таблицы нормального закона для нахождения критических значений.

Для задачи о двух экспертах (таблица 16-1) проверим значимость полученного коэффициента, то есть ответим на вопрос, имеется ли связь между переменными в генеральной совокупности. Наличие значимой связи будет означать, что мнения экспертов близки, отсутствие связи – что эксперты абсолютно по-разному оценивают моделей и их мнения не совпадают. Проверим гипотезу.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: $\rho_s = 0$ (нет связи между оценками экспертов)

H_1: $\rho_s \neq 0$ (есть связь)

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Критическое значение для α=0,05 и объема выборки n=10 находим в таблице А-10, оно равно ±0,648.

ШАГ 4. Вычисляем выборочное значение коэффициента корреляции Спирмена. Сумма квадратов разностей рангов равна 74. Значение коэффициента:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 0{,}552$$

ШАГ 5. Сравниваем полученное значение с критической областью. Коэффициент, близкий к ±1 означает наличие связи между переменными. В нашем случае это не так. Мы не попали в критическую область. Это означает, что мы принимаем основную гипотезу.

ШАГ 6. Пишем ответ. Связи между мнением экспертов нет.

Второй критерий проверки значимости

Имеется другой критерий для проверки значимости, который удобно использовать в отсутствии таблиц критических значений коэффициента Спирмена:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

Таблица 16-2. Проверка связи между результатами экзаменов

СТУ-ДЕНТ	МАТЕМА-ТИКА	СТАТИС-ТИКА	РАНГ ПО МАТЕМА-ТИКЕ	РАНГ ПО СТАТИС-ТИКЕ	РАЗНОСТЬ РАНГОВ	КВАДРАТ РАЗНОСТИ
1	22	17	6	8	-2	4
2	49	43	3	1	2	4
3	44	23	4	6	-2	4
4	50	30	2	4	-2	4
5	57	42	1	2	-1	1
6	10	20	8	7	1	1
7	25	32	5	3	2	4
8	17	28	7	5	2	4
						Σ=26

Он построен на t-статистике с числом степеней свободы df = n – 1. Критические значения находятся по таблице А-3, а затем строится двусторонняя критическая область $(-\infty;-t_{\alpha/2}]\cup[t_{\alpha/2};+\infty)$.

Для примера проверим гипотезу о наличии связи для следующей задачи. Анализируется связь между результатами экзаменов по статистке и философии. В выборку попали восемь студентов, их результаты записаны во втором и третьем столбце таблицы 16-2 (оценки по 100-бальной шкале).

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: $\rho_s = 0$ (нет связи между результатами экзаменов)
H_1: $\rho_s \neq 0$ (связь есть)

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Критическое значение для α=0,05, объема выборки n = 8, числа степеней свободы df = 8 – 2 =6 находим в таблице А-3 равно ±2,45.

ШАГ 4. Вычисляем значение статистики по выборке. Находим сначала выборочное значение коэффициента корреляции Спирмена. Сумма квадратов разностей рангов равна 26 (см.таблицу 16-2). Значение коэффициента:

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 26}{8 \cdot (64 - 1)} = 1 - \frac{156}{504} = 0{,}6905$$

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} = 0{,}69 \sqrt{\frac{8-2}{1-(0{,}69)^2}} = 2{,}34$$

ШАГ 5. Сравниваем полученное значение с критической областью. Значение 2,34 не попадает в критическую область.

ШАГ 6. Пишем ответ. Связь между результатами экзаменов в генеральной совокупности не выявлена.

В рассмотренной нами задаче про экспертов мы сразу имели дело с рангами, поскольку места, проставленные фотомоделям, этими рангами и являются. Во второй задаче (про экзамены) нам пришлось ранжировать результаты экзаменов (четвертый и пятый столбец) а затем вычислять коэффициент.

16-2 Коэффициент Кендалла

Коэффициент Кендалла используется в тех же целях, что и коэффициент Спирмена – для проверки наличия связи между двумя переменными, измеряемыми порядковой шкалой. Отличие состоит в методе. Изложим его суть, используя следующий пример.

Пример. Результаты соревнований. Имеются результаты соревнований шести спортсменов в двух видах спорта: плавании и стрельбе (таблица 16-3). Требуется проверить, имеется ли значимая связь между результатами спортсменов в этих видах спорта.

Решение. Необходимо построить всевозможные пары спортсменов, в нашем примере их будет 15:

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

Таблица 16-3. Проверка связи для результатов соревнований

Спортсмен	Место по плаванию	Место по стрельбе
1	2	2
2	1	3
3	3	1
4	4	5
5	5	4
6	6	6

В общем случае, если число наблюдений *n*, то число пар равно:

$$\frac{n\cdot(n-1)}{2}$$

Назовем пару **проверсией,** если в ней большее место по плаванию соответствует большему месту в стрельбе. Будем говорить, что в такой паре **порядок** мест совпадает. В противном случае, когда большее место по плаванию соответствует меньшему месту в стрельбе, будем называть такую пару **инверсией.**

Например, если рассмотреть пару спортсменов (2, 4), то это инверсия, поскольку второй спортсмен по сравнению с четвертым имеет более высокие результаты и в беге, и в стрельбе. Это инверсия. Для пары (4, 5) все наоборот, четвертый спортсмен выше по плаванию, но ниже по стрельбе. Эта пара – инверсия.

Таблица 16-4. Подсчет числа проверсий и инверсий

Спортсмен	Место по плаванию	Место по стрельбе	Проверсий	Инверсий
2	1	3	$\boxed{3}$	$\boxed{2}$
1	2	2		
3	3	1		
4	4	5		
5	5	4		
6	6	6		

Таблица 16-5. Результаты подсчета числа проверсий и инверсий

Спортсмен	Место по плаванию	Место по стрельбе	Проверсий	Инверсий	Всего
2	1	3	3	2	5
1	2	2	3	1	4
3	3	1	3	0	3
4	4	5	1	1	2
5	5	4	1	0	1
6	6	6	0	0	0
Итого			11	4	15

Метод основывается на том, что если связь полная, то все 15 пар будут проверсиями и наоборот, чем меньше проверсий, тем слабее связь.

Вычисление коэффициента

Коэффициент Кендалла находится по формуле:

$$\tau = \frac{P - I}{P + I}$$

где P – число проверсий,
I – число инверсий.

Отметим, что сумма числа проверсий и инверсий равна общему количеству пар:

$$P + I = \frac{n \cdot (n-1)}{2}$$

Это равенство позволяет путем несложных преобразований получить две другие формулы для критерия:

$$\tau = \frac{4P}{n \cdot (n-1)} - 1, \quad \tau = 1 - \frac{4I}{n(n-1)}$$

Очевидно, любая из этих формул даст правильный результат, поэтому следует использовать ту, которую проще использовать. Обсудим последовательность действий для нахождения коэффициента корреляции Кендалла, которая удобна для ручного счета.

ШАГ 1. Упорядочиваем данные по первой переменной. В нашем примере мы располагаем спортсменов в порядке возрастания мест по плаванию.

ШАГ 2. Затем подсчитываем число проверсий и инверсий. В частности, для спортсмена под номером 2, имеющего первое место по плаванью и третье место по стрельбе имеется 3 проверсии и 2 инверсии. Подсчет показан в таблице 16-4. Заполняем так всю таблицу, получив результаты для всех 15 пар спортсменов (таблица 16-5).

ШАГ 3. Вычисляем значение коэффициента Кендалла:

$$\tau = \frac{P - I}{P + I} = \frac{11 - 4}{11 + 4} = \frac{7}{15} = 0{,}47$$

ШАГ 4. Интерпретируем полученный результат. Значение коэффициента на уровне 0,47 свидетельствует о наличии некоторой положительной связи между результатами по плаванию и стрельбе.

Следует иметь в виду, что коэффициент может изменяться от -1 до +1. Значение, близкое к нулю, свидетельствует о приблизительно равном количестве проверсий и инверсий, что означает отсутствие связи. Значение, близкое к -1, означает наличие явной отрицательной связи.

Проверка значимости

Полученное значение является выборочным, то есть оно вычислено для результатов выборки из шести спортсменов. Может существовать другая задача: на основе выборочного значения коэффициента Кендалла сделать вывод о наличии связи в генеральной совокупности. Этот критерий, как и в случае коэффициента Спирмена, называется проверкой значимости связи.

Будем обозначать выборочный коэффициент ранговой корреляции Кендалла $\tau_{выб}$, а коэффициент генеральной совокупности - $\tau_{ген}$. Проверим гипотезу о равенстве нулю коэффициента ранговой корреляции генеральной совокупности $\tau_{ген}$ на основании вычисленного значения коэффициента ранговой корреляции выборки $\tau_{выб}$.

Равенство нулю будет означать отсутствие значимой связи между двумя рассматриваемыми переменными. Для проверки гипотезы используется нормальное приближение и в качестве критерия применяется функция:

$$z = \frac{\tau_{выб}}{\sqrt{\dfrac{2\cdot(2n+5)}{9n\cdot(n-1)}}}$$

Границы двусторонней критической области находятся при помощи таблиц нормального распределения.

Рассмотрим алгоритм проверки гипотезы применительно к задаче о результатах спортсменов.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: $\tau_{ген} = 0$ (нет связи в результатах по двум видам спорта)
H_1: $\tau_{ген} \neq 0$ (связь имеется)

ШАГ 2. Задаем уровень значимости. Например, α=0,05.

ШАГ 3. Определяем критические значения. Для α=0,05 находим в таблице А-2 две границы критической области $\pm$1,96. Критическая область двусторонняя: $(-\infty; -1{,}96] \cup [1{,}96; +\infty)$.

ШАГ 4. Вычисляем значение статистики (критерия). Находим сначала выборочное значение коэффициента корреляции Кендалла (таблица 16-5), а затем выборочное z-значение:

$$\tau = \frac{P-I}{P+I} = \frac{11-4}{11+4} = \frac{7}{15} = 0{,}47$$

$$z = \frac{\tau_{выб}}{\sqrt{\dfrac{2\cdot(2n+5)}{9n\cdot(n-1)}}} = \frac{0{,}47}{\sqrt{\dfrac{2\cdot(2\cdot6+5)}{9\cdot6\cdot(6-1)}}} = 1{,}32$$

ШАГ 5. Сравниваем полученное значение с критической областью. Значение 1,32 не попадает в критическую область.

ШАГ 6. Пишем ответ. Выявленная выборочная связь между результатами спортсменов в двух видах спорта не означает наличие связи в генеральной совокупности. Для генеральной совокупности нет оснований считать, что имеется значимая связь.

Резюме

Два коэффициента ранговой корреляции – Спирмена и Кендалла применяются для проверки наличия связи между переменными, измеряемыми порядковой шкалой.

Для интервальных переменных применение указанных коэффициентов требует перехода от значений к их рангам. Коэффициент Спирмена использует тот же метод, что и коэффициент Пирсона для интервальных переменных, а коэффициент Кендалла основан на сравнении числа проверсий и инверсий, которые характеризуют наличие или отсутствие связи между переменными. Если число проверсий и инверсий приблизительно равно, это означает отсутствие связи.

Используем компьютер

Вычисление коэффициентов ранговой корреляции возможно при помощи компьютера. В пакете SPSS следует воспользоваться разделом Analyze → Correlate → Bivariate … Для указанных переменных могут быть сосчитаны коэффициент Пирсона, коэффициенты Спирмена (ро) и Кендалла (тау). Сопоставление с результатами вычислений вручную не должно привести к отклонениям, если не имеется ошибок в вычислениях. Расхождения могут возникнуть лишь в точности полученных значений. Коэффициенты ранговой корреляции вычисляются, обычно, до третьего знака после запятой.

Что означают термины

Коэффициент ранговой корреляции Спирмена

Коэффициент ранговой корреляции Кендалла

Инверсия

Проверсия

Символы и формулы

$r = 1 - \dfrac{6 \cdot \sum d_i^2}{n(n^2 - 1)}$	Коэффициент ранговой корреляции Спирмена
$r_s = \dfrac{\pm z}{\sqrt{n-1}}$	Критические значения для проверки значимости коэффициента ранговой корреляции Спирмена
$t = r_s \sqrt{\dfrac{n-2}{1-r_s^2}}$	t-критерий для проверки значимости коэффициента ранговой корреляции Спирмена
$\tau = \dfrac{P - I}{P + I}$	Коэффициент ранговой корреляции Кендалла
P	Число проверсий
I	Число инверсий
$\tau = \dfrac{4P}{n \cdot (n-1)} - 1$	Коэффициент ранговой корреляции Кендалла (вторая формула)
$\tau = 1 - \dfrac{4I}{n(n-1)}$	Коэффициент ранговой корреляции Кендалла (третья формула)
$z = \dfrac{\tau_{выб}}{\sqrt{\dfrac{2 \cdot (2n+5)}{9n \cdot (n-1)}}}$	z-критерий для проверки значимости коэффициента ранговой корреляции Кендалла

Задачи и упражнения

16-1. Родители и подростки. Восемь музыкальных фильмов были проранжированы подростками и их родителями по стилю и ясности (1 – самый высокий ранг). На уровне значимости α = 0,05 есть ли связь между этими данными?

Фильм	1	2	3	4	5	6	7	8
Подростки	4	6	2	8	1	7	3	5
Родители	1	7	5	4	3	8	2	6

16-2. Тренеры и комментаторы. 8 игроков в теннис ранжированы спортивными комментаторами и тренерами (1 – высший ранг). На уровне значимости α = 0,05 есть ли связь между этими данными?

Игроки	A	B	C	D	E	F	G	H
Тренеры	4	6	5	1	7	2	3	8
Комментаторы	7	6	4	3	5	2	1	9

16-3. Кто смотрит телевизор? Исследователь хочет определить, существует ли связь между возрастом человека и тем, сколько часов в день он (или она) смотрит телевизор.

Респондент	A	B	C	D	E
Возраст	18	24	36	40	58
Количество часов	3,9	2,6	2	2,3	1,2

16-4. IQ мальчиков и девочек. Психолог отбирает шесть семей с двумя детьми в каждой, мальчиком и девочкой. Ее цель – сопоставить IQ детей обоего пола, чтобы определить, существует ли зависимость между коэффициентом интеллекта детей одной семьи.

Семья	A	B	C	D	E	F
IQ девочек	107	95	116	109	101	98
IQ мальчиков	107	102	112	104	105	103

Глава 17. Непараметрические критерии. Факторный анализ

Эта глава посвящена непараметрическим методам, которые позволяют проверять наличие влияния на зависимую переменную одного (критерий Краскела-Уоллиса) или двух факторов (критерий Фридмана). Оба критерия основаны на ранжировании данных. Критерий Фридмана позволяет проверять влияние основного фактора, устраняя влияние мешающего фактора.

17-1 Однофакторный непараметрический анализ. Критерий Краскела-Уоллиса

В дисперсионном анализе используется F-критерий, чтобы сравнивать средние трех и более совокупностей. Для критерия ANOVA предполагается, что совокупности нормально распределены и что дисперсии совокупностей равны. Когда эти условия не выполняются, то для сравнения трех и более средних может использоваться непараметрический критерий Краскела–Уоллиса.

Критерий Краскела-Уоллиса – непараметрический тест, который использует ранги трех и более независимых выборок и применяется для проверки гипотезы о том, что выборки получены из генеральных совокупностей, имеющих одинаковый закон распределения.

Данные, которые исследуются, представлены в таблице. Столбцы соответствуют измерениям на каждом уровне фактора. Этот вид представления данных мы уже рассмотрели в главе 14 (рисунок 14-2).

Таблица 17-1. Проверка совпадения законов распределения

УЧИТЕЛЯ	РАНГИ	АДМ. ПЕРСОНАЛ	РАНГИ	ОБСЛ. ПЕРСОНАЛ	РАНГИ
24	1	59	19	34	7
27	3	35	8,5	29	4,5
26	2	29	4,5	35	8,5
50	16	40	12	31	6
48	15	39	10	40	12
40	12	54	17	45	14
		56	18		
Объемы выборок	6		7		6
Суммы рангов	49		89		52
Средние ранги	8,17		12,71		8,67

Для применения критерия требуется соблюдение двух **условий**:

(1) Выборки независимы и получены случайным образом.

(2) Размер каждой выборки должен быть не меньше пяти.

В этом случае исследуемое распределение приближается к хи-квадрат распределению с числом степеней свободы df = k – 1, где k – число уровней исследуемого признака (столбцов таблицы). Для выборок меньшего размера требуются специальные таблицы. (3) При этом, не требуется, чтобы генеральные совокупности имели нормальный закон распределения.

Суть метода состоит в следующем. В критерии Краскела–Уоллиса все выборки перемешиваются и их значения ранжируются. Далее вычисляются средние ранги для каждой выборки и средний ранг по всем данным. Для проверки используется следующая статистика (два варианта формулы):

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{\overline{R}}\right)^2, \text{ или}$$

$$H = \frac{12}{N(N+1)} \cdot \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_k^2}{n_k}\right) - 3(N+1)$$

где R_i - сумма рангов по столбцу,

$\overline{R}_i$ - средний ранг по столбцу,

$\overline{\overline{R}}$ - средний ранг по таблице, $\overline{\overline{R}} = \dfrac{N+1}{2}$,

n_i - количество измерений на уровне i,

N - общее число данных, $N = n_1 + n_2 + \ldots + n_k$.

Если выборки взяты из различных совокупностей, их средние ранги будут сильно различаться, значение H будет велико, что будет означать, что нулевую гипотезу следует отвергнуть. Для двух выборок критерий совпадает с уже известным нам критерием Вилкоксона.

В качестве примера рассматриваем вновь данные о возрасте персонала, приведенные в таблице 14-1.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: распределения генеральных совокупностей совпадают
H_1: распределения отличаются

ШАГ 2. Задаем уровень значимости α=0,05.

ШАГ 3. Определяем критическое значение. Пользуемся таблицей А-4. Число степеней свободы df = (k – 1) = (3 – 1) = 2. Критическое значение равно 5,991. Критическая область правосторонняя.

ШАГ 4. Вычисляем значение статистики по выборке. Сначала дополняем таблицу столбцами, проставляем в них ранги, ранжируя все значения в таблице одновременно. Вычисляем значение статистики по первой формуле:

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{\overline{R}}\right)^2 =$$
$$= \frac{12}{N(N+1)} \cdot \left(6 \cdot (8{,}17 - 10)^2 + 7 \cdot (12{,}71 - 10)^2 + \right.$$
$$\left. + 6 \cdot (8{,}67 - 10)^2 \right) = 2{,}602$$

Таблица 17-2. Таблица данных для двухфакторного анализа

	Фактор А (Обработки)				
Фактор В (Блоки)	1	2	3	…	k
1	x_{11}	x_{12}	x_{13}	…	x_{1k}
2	x_{21}	x_{22}	x_{23}	…	x_{2k}
…	…	…	…	…	…
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nk}

ШАГ 5. Сравниваем полученное значение с критической областью. Значение 2,602 не попадает в критическую область, 2,60 < 5,991.

ШАГ 6. Пишем ответ. У нас нет оснований считать, что выборки получены из генеральных совокупностей, имеющих различные законы распределения. Мы можем считать эти распределения одинаковыми.

Если использовать вторую формулу, мы получили бы тот же ответ, поскольку обе формулы эквивалентны и получены преобразованием одна из другой.

17-2 Двухфакторный непараметрический анализ. Критерий Фридмана

Мы уже рассматривали двухфакторный анализ, когда проверяли гипотезы об отсутствии влияния двух факторов на зависимую переменную, а также гипотезу об отсутствии взаимодействия между двумя факторами. Наш метод был параметрическим, требовалось, чтобы генеральные совокупности имели нормальный закон распределения и дисперсии были равны. Теперь мы рассмотрим критерий, который является непараметрическим и, тем самым, позволяет решать более широкий класс задач.

Предположим, что мы располагаем результатами наблюдений, которые представлены в таблице 17-2.

Таблица 17-3. Переход от результатов наблюдений к рангам

	Фактор А (Обработки)				
Фактор В (Блоки)	1	2	3	…	k
1	r_{11}	r_{12}	r_{13}	…	r_{1k}
2	r_{21}	r_{22}	r_{23}	…	r_{2k}
…	…	…	…	…	…
n	r_{n1}	r_{n2}	r_{n3}	…	r_{nk}

Обсудим терминологию, которую принято использовать для общей модели двухфакторного непараметрического анализа.

На результаты могут оказывать воздействие два фактора - А и В. Фактор А имеет n уровней, фактор В имеет k уровней. Таблица содержит $n \times k$ наблюдений – по одному наблюдению в каждой клетке. Будем считать, что А – основной фактор, а В – мешающий фактор. Уровни основного фактора назовем *обработками,* а уровни мешающего фактора – *блоками*. Влияние основного фактора называют в этом случае *эффектами обработки,* а влияние мешающего фактора – *эффектами блоков.*

Суть критерия Фридмана состоит в переходе от таблицы наблюдений к таблицам рангов. При ранжировании по строкам, нам удается исключить влияние мешающего фактора и проверить влияние основного фактора. Проверяются основная и альтернативная гипотезы:

H_0: влияние фактора А отсутствует
H_1: влияние фактора имеется

Общую модель двухфакторного анализа часто записывают в виде следующего равенства:

$$x_{ij} = \mu + \beta_i + \alpha_j + \varepsilon_{ij}$$

где μ - среднее значение генеральной совокупности,
β_i - влияние фактора В,
α_j - влияние фактора А,
ε_{ij} - влияние случайности.

Это означает, что результаты наблюдений могут быть представлены в виде суммы нескольких составляющих: среднего, величин влияния каждого из факторов и величины, зависящей от случайности эксперимента.

С учетом введенных обозначений, гипотеза об отсутствии влияния фактора А записывается следующим образом:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
H_1: не все α_i равны нулю

Для проверки гипотезы применяется следующий критерий:

$$S = \frac{12n}{k(k+1)} \cdot \sum_{i=1}^{k} \left(\bar{r}_i - \bar{\bar{r}}\right)^2$$

где $\bar{r}_i$ - средние ранги по столбцам,
$\bar{\bar{r}}$ - средний ранг по таблице.

Известно, что средний ранг по таблице находится по формуле:

$$\bar{\bar{r}} = \frac{k+1}{2}$$

Имеется также вторая формула для вычисления критерия:

$$S = \frac{12}{nk(k+1)} \sum r_{ij}^2 - 3n(k-1)$$

где r_{ij} - ранги всех значений в таблице.

Для небольших объемов выборки составлены критических значений для критерия Фридмана (таблица А-12).

После проверки гипотезы об отсутствии влияния фактора А, мы можем провести проверку влияния фактора В, исключив теперь влияние мешающего фактора А. Ранжировать данные в этом случае следует по столбцам. Такая двойная процедура позволит последовательно проверить два блока гипотез: об отсутствии воздействия фактора А, и об отсутствии воздействия фактора В.

Таблица 17-4. Сообразительность и (или) физическая подготовка

ФИЗИЧЕСКИЕ ДАННЫЕ	УМСТВЕННЫЕ СПОСОБНОСТИ			
	МЕНЕЕ 80	80-100	100-120	БОЛЕЕ 120
Очень слабый	34	28	20	19
Слабый	32	31	24	13
Средний	20	18	19	24
Сильный	22	19	15	16
Очень сильный	26	24	14	15

Пример. Задание для новобранцев. Предположим, что военнослужащим срочной службы было предложено выполнение несложного задания на пересеченной местности. По замыслу командиров задание может быть успешно выполнено при наличии как умственных, так и физических данных. *Умственные способности* испытуемых измерялись по шкале IQ, а затем были решено выделить четыре категории: менее 80 баллов, от 80 до 100 баллов, от 100 до 120 баллов и выше 120 баллов. *Физические данные* измерялись по результатам спортивных соревнований, было решено выделить пять уровней: от «очень слабый» до «очень сильный». Двадцать испытуемых, по одному для каждого сочетания уровней прошли испытание и показали результаты, измеренные в минутах. Данные размещены в таблице 17-4. Общее время выполнения задания варьируется от 13 до 34 минут. Очевидно, на результаты испытания могли оказывать влияние и другие факторы, а также случайность эксперимента. Требуется на уровне значимости $\alpha = 0{,}05$ проверить гипотезы об отсутствии влияния умственных способностей и физической подготовки на результаты испытания.

Решение. **Первая часть**. На первом этапе проведем проверку гипотезы об отсутствии влияния умственных способностей на результаты испытания.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: влияние умственных способностей отсутствует

H_1: влияние умственных способностей значимо

ШАГ 2. Задан уровень значимости $\alpha=0{,}05$.

ШАГ 3. Определим критическое значение. Пользуемся таблицей А-12. Количество уровней первого фактора $k = 4$, а второго фактора $n = 5$. Критическое значение равно 7,320. Критическая область правосторонняя.

Таблица 17-5. Проверка влияния умственных способностей

ФИЗИЧЕСКИЕ ДАННЫЕ	УМСТВЕННЫЕ СПОСОБНОСТИ			
	МЕНЕЕ 80	80-100	100-120	БОЛЕЕ 120
Очень слабый	34	28	20	19
Ранги	4	3	2	1
Слабый	32	31	24	13
Ранги	4	3	2	1
Средний	20	18	19	24
Ранги	3	1	2	4
Сильный	22	19	15	16
Ранги	4	3	1	2
Очень сильный	26	24	14	15
Ранги	4	3	1	2
Сумма рангов по столбцу	19	13	8	10
Средний ранг столбца	3,80	2,60	1,60	2,00
Средний ранг таблицы	2,50	2,50	2,50	2,50
Квадрат разности	1,69	0,01	0,81	0,25
Сумма квадратов разностей	2,760			

ШАГ 4. Вычисляем значение статистики. Дополняем таблицу расчетными строками. Ниже каждой строки проставляем ранги значений, ранжируя каждую строку. Затем находим сумму рангов и средний ранг для каждого столбца. Затем находим средний ранг по таблице:

$$\bar{\bar{r}} = \frac{k+1}{2} = \frac{4+1}{2} = 2,5$$

Вычитаем из среднего ранга по столбцу средний ранг по таблице и возводим в квадрат. В самой последней строке таблицы находим сумму квадратов разностей:

$$\sum_{i=1}^{k} \left(\bar{r}_i - \bar{\bar{r}}\right)^2 = 2,760$$

Теперь подставляем полученное значение в формулу критерия:

$$S = \frac{12n}{k(k+1)} \cdot \sum_{i=1}^{k} \left(\bar{r}_i - \bar{\bar{r}}\right)^2 =$$
$$= \frac{12 \cdot 5}{4 \cdot (4+1)} \cdot 2{,}760 = 8{,}280$$

ШАГ 5. Сравниваем полученное значение с критической областью. Значение 8,280 попадает в критическую область, 8,280 > 7,320.

ШАГ 6. Пишем ответ. Проверка показала значимость влияния умственных способностей на результаты испытания.

Решение. Вторая часть. При проверке гипотезы фактор «умственные способности» был основным, а фактор «физические данные» мешающим. При помощи ранжирования по строкам мы устранили влияние мешающего фактора и смогли проверить действие основного. Теперь проведем вторую проверку – проверку влияния физических данных на результаты испытаний. Для этого необходимо «поменять местами» два фактора. Мы это сделаем при помощи поворота относительно диагонали таблицы данных.

ШАГ 1. Формулируем основную и альтернативную гипотезы:

H_0: влияние физических данных не значимо
H_1: влияние физических данных значимо

ШАГ 2. Задан уровень значимости α=0,05.

ШАГ 3. Определим критическое значение. Пользуемся таблицей А-12. Количество уровней первого фактора k = 5, а второго фактора n= 4. Критическое значение равно 8,800. Критическая область правосторонняя.

Таблица 17-6. Проверка влияния физических данных

УМСТВЕННЫЕ СПОСОБНОСТИ	ФИЗИЧЕСКИЕ ДАННЫЕ				
	ОЧЕНЬ СЛАБЫЙ	СЛАБЫЙ	СРЕДНИЙ	СИЛЬНЫЙ	ОЧЕНЬ СИЛЬНЫЙ
менее 80	34	32	20	22	26
Ранги	5	4	1	2	3
80-100	28	31	18	19	24
Ранги	4	5	1	2	3
100-120	20	24	19	15	14
Ранги	4	5	3	2	1
более 120	19	13	24	16	15
Ранги	4	1	5	3	2
Сумма рангов по столбцу	17	15	10	9	9
Средний ранг столбца	4,25	3,75	2,50	2,25	2,25
Средний ранг таблицы	3,00	3,00	3,00	3,00	3,00
Квадрат разности	1,56	0,56	0,25	0,56	0,56
Сумма квадратов разностей	3,50				

ШАГ 4. Вычисляем значение статистики. Дополняем таблицу расчетными строками. Ниже каждой строки проставляем ранги значений, ранжируя каждую строку. Затем находим сумму рангов и средний ранг для каждого столбца. Затем находим средний ранг по таблице:

$$\bar{\bar{r}} = \frac{k+1}{2} = \frac{5+1}{2} = 3{,}0$$

Вычитаем из среднего ранга по столбцу средний ранг по таблице и возводим в квадрат. В самой последней строке таблицы находим сумму квадратов разностей:

$$\sum_{i=1}^{k} \left(\bar{r}_i - \bar{\bar{r}}\right)^2 = 3{,}500$$

Подставляем полученное значение в формулу критерия:

$$S = \frac{12n}{k(k+1)} \cdot \sum_{i=1}^{k} \left(\bar{r}_i - \bar{\bar{r}}\right)^2 =$$
$$= \frac{12 \cdot 4}{5 \cdot (5+1)} \cdot 3{,}500 = 5{,}600$$

ШАГ 5. Сравниваем полученное значение с критической областью. Значение 5,600 не попадает в критическую область, 5,600 < 8,800.

ШАГ 6. Пишем ответ. Проверка показала, что у нас не достаточно оснований отвергнуть основную гипотезу, и мы считаем, что значимое влияние физических данных на результаты испытания отсутствует.

Используем компьютер

Для проведения вычислений по модели однофакторного и двухфакторного анализа удобно использовать статистические пакеты, поскольку при возрастании объема наблюдений довольно сложно без ошибок проводить ранжирование чисел и вычисление критериев. Следует быть осторожными с электронными таблицами, поскольку в EXCEL функция ранжирования для повторяющихся значений выполняется иначе, чем это было определено нами. Это, в частности, может привести к серьезным отклонениям результатов вычислений, полученных в EXCEL и в SPSS.

Что означают термины

Критерий Краскела-Уоллиса

Критерий Фридмана

Модель двухфакторного анализа

Обработки

Основной и мешающий факторы

Блоки

Эффекты обработки

Эффекты блоков

Символы и формулы

$H = \dfrac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{\overline{R}}\right)^2$	Критерий Краскела-Уоллиса
R_i	Сумма рангов по столбцу
$\overline{R}_i$	Средний ранг по столбцу
$\overline{\overline{R}} = \dfrac{N+1}{2}$	Средний ранг по таблице
n_i	Количество измерений на уровне i
$N = n_1 + n_2 + \ldots + n_k$	Общее количество данных
$H = \dfrac{12}{N(N+1)} \cdot \left(\dfrac{R_1^2}{n_1} + \dfrac{R_2^2}{n_2} + \ldots + \dfrac{R_k^2}{n_k}\right) - 3(N+1)$	Вторая формула критерия
$S = \dfrac{12n}{k(k+1)} \cdot \sum_{i=1}^{k} (\overline{r}_i - \overline{\overline{r}})^2$	Критерий Фридмана
$\overline{r}_i$	Средние ранги по столбцам
$\overline{\overline{r}} = \dfrac{k+1}{2}$	Средний ранг по таблице
$S = \dfrac{12}{nk(k+1)} \sum r_{ij}^2 - 3n(k-1)$	Вторая формула для критерия

Задачи и упражнения

17-1. Измерение детского «Я». Измеряется самооценка в трех различных выборках индивидов по возрастным группам. Количество набранных баллов ранжируется от 0 до 50. Существует ли разница в количестве набранных баллов на уровне значимости $\alpha = 0{,}05$?

7-9 лет	5-7 лет	3-5 лет
48	50	47
46	49	45
42	42	46
41	43	30
37	39	32
32	28	41

17-2. Три канала рекламы. Владельцы сети овощных магазинов исследуют эффективность различных каналов рекламы: радио, местного телевидения и газет. Выбраны три небольших города, в каждом из которых используется свой канал рекламы. По результатам акции в течение одной недели в случайно выбранных магазинах были получены данные о продажах. Проверить на уровне значимости $\alpha = 0{,}01$, существует ли разница в эффективности каналов рекламы?

Радио	ТВ	Газета
$832	$1024	$329
648	996	437
562	1011	561
786	853	329
452	471	382
975		495
		262

17-3. Задача про клубнику. Клубнику выращивают на трех различных типах почвы. Урожай (в квартах) на одинаковых участках представлен ниже. Существует ли различие в количестве урожая для трех участков на уровне значимости $\alpha = 0{,}01$?

Почва A	Почва B	Почва C
32	43	50
38	45	56
31	49	58
40	46	54
39	51	52

17-4. Заказы в ателье. Менеджер анализирует количество заказов в трех разных ателье в течение одной рабочей недели. На уровне значимости $\alpha = 0{,}01$ проверить, существует ли значимое различие?

A	B	C
6	5	10
8	1	12
7	10	9
5	3	13
6	6	4

17-5. Режиссер или бюджет. Исследователь хочет проверить, зависит ли качество итальянских фильмов от режиссера и от бюджета. Проведен опрос зрительской аудитории, при котором качество оценивалось по 30 бальной шкале. Результаты представлены в таблице. На уровне значимости $\alpha = 0{,}01$ проверить влияние каждого из указанных факторов.

	до 1 млн	1-3 млн	3-5 млн	от 5 млн
Дебютант	16	25	4	10
Не известный	19	28	18	15
Малоизвестный	9	16	6	12
Известный	4	8	10	9
Очень известный	8	6	21	23
Великий	27	14	9	17

Глава 18. Как провести собственное исследование

Эта глава – заключительная. В ней мы проведем обзор изученных методов с точки зрения их использования для проведения собственного исследования и дадим их классификацию. Первым критерием для выбора методов является тип данных, а вторым – количество исследуемых переменных. В зависимости от этого, необходимо обращаться к различным главам или параграфам настоящей книги.

18-1 Классификация методов

Итак, пора подвести итог. Каждая глава этой книги содержала изложение методов, которые могут быть использованы исследователем в тех или иных случаях. Примеры и задачи, которые были решены, составлены с учебными целями. В реальных исследованиях не известно, какой метод потребуется использовать.

Поэтому мы рассмотрим классификацию изученных методов с точки зрения их применения для различных типов задач. Два важных вопроса, на которые следует ответить при проведении исследования:

Вопрос 1. Какой тип данных рассматривается?

Вопрос 2. Сколько переменных?

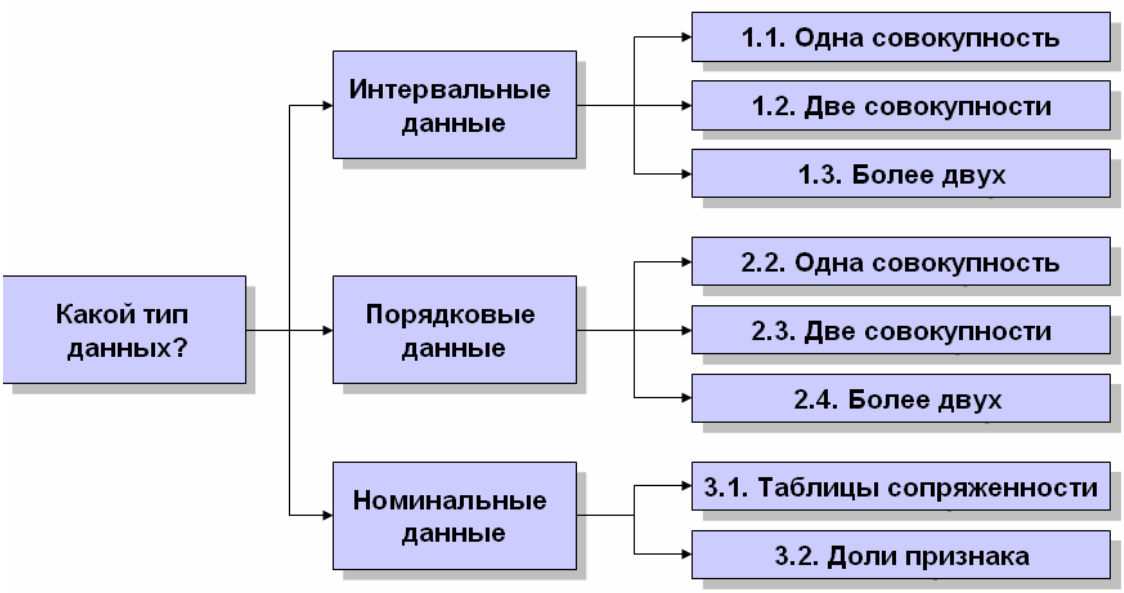


Рисунок 18-1. Тип данных и количество переменных влияют на выбор метода

На рисунке 18-1 показаны восемь различных ситуаций, которые возникают для разных типов данных и количества исследуемых переменных.

Для **одной генеральной совокупности:**

1) построение доверительных интервалов для среднего проводилось в параграфе 9-2;
2) проверка гипотез о среднем изучалась в параграфе 10-2;
3) доверительный интервал для дисперсии строился в параграфе 9-4;
4) проверка гипотез о дисперсии проводилась в 10-4.

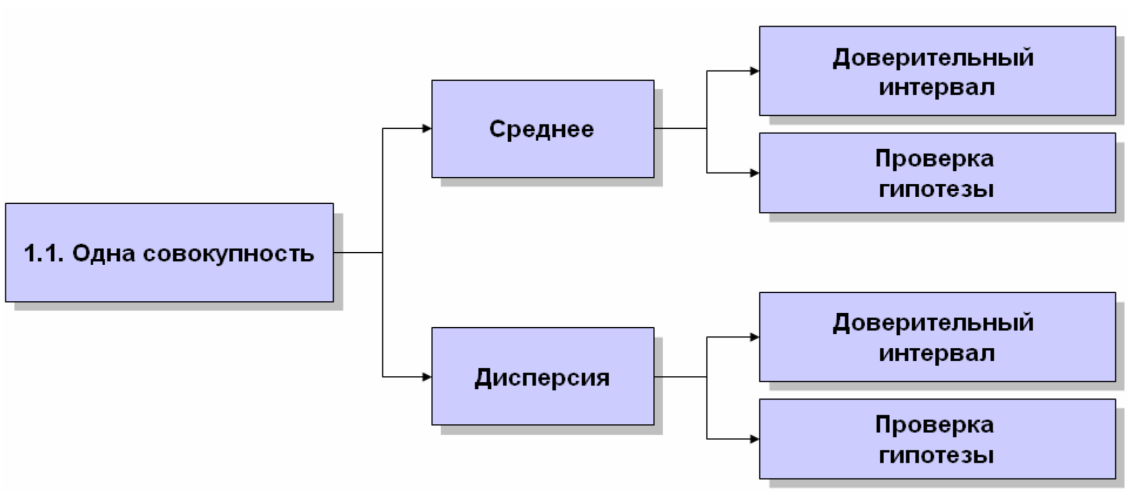


Рисунок 18-2. Интервальные данные, одна совокупность.

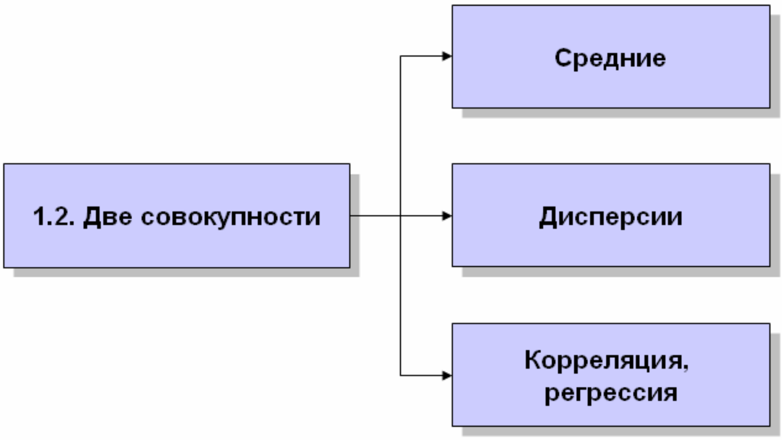


Рисунок 18-3. Интервальные данные, две совокупности.

Для **двух генеральных совокупностей:**

1) сравнение двух средних изучено в параграфе 11-2 и 11-3;
2) сравнение двух дисперсий рассмотрено в параграфе 11-5;
3) корреляция и регрессия изучались в главе 13 (параграфы 13-2 и 13-3)

Для **порядковых данных** проверка гипотезы для одной генеральной совокупности (гипотеза о значении медианы) исследована в параграфе 15-2. Две совокупности рассмотрены в случае независимых и парных выборок в параграфах с 15-2 до 15-4. Случай более двух совокупностей подробно рассмотрен в главе 16.

Для **номинальных данных** таблицы сопряженности исследовались в главе 12. Гипотезы о доле признака одной совокупности проверялись в 10-3. Сравнение двух долей проводилось в 11-4.

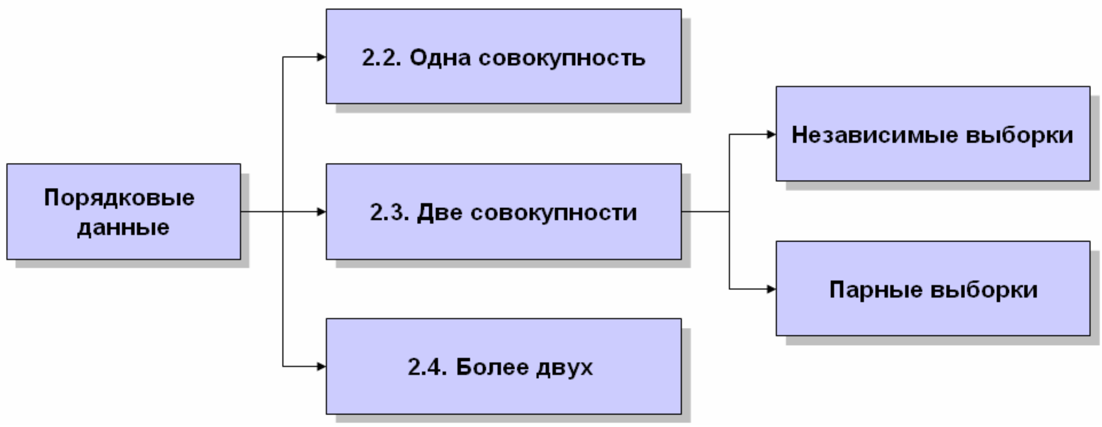


Рисунок 18-4. Порядковые данные.

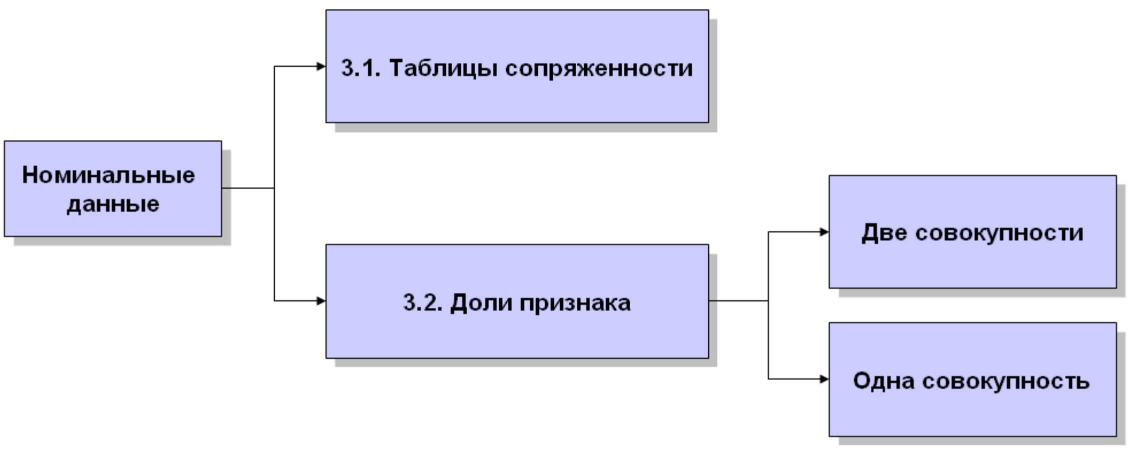


Рисунок 18-5. Номинальные данные.

18-2 Вместо заключения

Спустя некоторое время, совсем скоро, каждый завершивший изучение этого курса сможет понять, что методы, изложенные в нем, это только начало. Этот курс можно было бы назвать, как это делается в западных учебниках и программах, **курсом элементарной статистики**. Мы не сделали этого только лишь потому, чтобы не формировать поверхностного отношения студентов к изучению этого курса, поскольку он является действительным фундаментом для применения статистических методов в социологии и других общественных науках.

А дальше – вторая и другие ступени. Вам потребуется понять, что бывают более сложные шкалы, во многих исследованиях необходимо рассматривать временные ряды, ключевым для социолога является многомерное шкалирование, о котором не было сказано в этом курсе ни слова. Все это впереди.

Приложение А. Таблицы

Приведены все необходимые таблицы, которые используются для решения задач по статистике. Комментарии по использованию таблиц содержатся в соответствующих разделах курса.

ПЕРЕЧЕНЬ ТАБЛИЦ

Таблица А-0. Таблица случайных чисел

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

Таблица А-1. Биномиальное распределение

n = общее количество испытаний
k = количество успехов
p = вероятность успеха в одном испытании

		p													
n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
2	0	980	903	810	640	490	360	250	160	090	040	010	003	000	0
	1	020	095	180	320	420	480	500	480	420	320	180	095	020	1
	2	000	003	010	040	090	160	250	360	490	640	810	903	980	2
3	0	970	857	729	512	343	216	125	064	027	008	001	000	000	0
	1	029	135	243	384	441	432	375	288	189	096	027	007	000	1
	2	000	007	027	096	189	288	375	432	441	384	243	135	029	2
	3	000	000	001	008	027	064	125	216	343	512	729	857	970	3
4	0	961	815	656	410	240	130	063	026	008	002	000	000	000	0
	1	039	171	292	410	412	346	250	154	076	026	004	000	000	1
	2	001	014	049	154	265	346	375	346	265	154	049	014	001	2
	3	000	000	004	026	076	154	250	346	412	410	292	171	039	3
	4	000	000	000	002	008	026	063	130	240	410	656	815	961	4
5	0	951	774	590	328	168	078	031	010	002	000	000	000	000	0
	1	048	204	328	410	360	259	156	077	028	006	000	000	000	1
	2	001	021	073	205	309	346	313	230	132	051	008	001	000	2
	3	000	001	008	051	132	230	313	346	309	205	073	021	001	3
	4	000	000	000	006	028	077	156	259	360	410	328	204	048	4
	5	000	000	000	000	002	010	031	078	168	328	590	774	951	5
6	0	941	735	531	262	118	047	016	004	001	000	000	000	000	0
	1	057	232	354	393	303	187	094	037	010	002	000	000	000	1
	2	001	031	098	246	324	311	234	138	060	015	001	000	000	2
	3	000	002	015	082	185	276	313	276	185	082	015	002	000	3
	4	000	000	001	015	060	138	234	311	324	246	098	031	001	4
	5	000	000	000	002	010	037	094	187	303	393	354	232	057	5
	6	000	000	000	000	001	004	016	047	118	262	531	735	941	6
7	0	932	698	478	210	082	028	008	002	000	000	000	000	000	0
	1	066	257	372	367	247	131	055	017	004	000	000	000	000	1
	2	002	041	124	275	318	261	164	077	025	004	000	000	000	2
	3	000	004	023	115	227	290	273	194	097	029	003	000	000	3
	4	000	000	003	029	097	194	273	290	227	115	023	004	000	4
	5	000	000	000	004	025	077	164	261	318	275	124	041	002	5
	6	000	000	000	000	004	017	055	131	247	367	372	257	066	6
	7	000	000	000	000	000	002	008	028	082	210	478	698	932	7
8	0	923	663	430	168	058	017	004	001	000	000	000	000	000	0
	1	075	279	383	336	198	090	031	008	001	000	000	000	000	1
	2	003	051	149	294	296	209	109	041	010	001	000	000	000	2
	3	000	005	033	147	254	279	219	124	047	009	000	000	000	3
	4	000	000	005	046	136	232	273	232	136	046	005	000	000	4
	5	000	000	000	009	047	124	219	279	254	147	033	005	000	5
	6	000	000	000	001	010	041	109	209	296	294	149	051	003	6
	7	000	000	000	000	001	008	031	090	198	336	383	279	075	7
	8	000	000	000	000	000	001	004	017	058	168	430	663	923	8

продолжение таблицы А-1

		P													
n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
9	0	914	630	387	134	040	010	002	000	000	000	000	000	000	0
	1	083	299	387	302	156	060	018	004	000	000	000	000	000	1
	2	003	063	172	302	267	161	070	021	004	000	000	000	000	2
	3	000	008	045	176	267	251	164	074	021	003	000	000	000	3
	4	000	001	007	066	172	251	246	167	074	017	001	000	000	4
	5	000	000	001	017	074	167	246	251	172	066	007	001	000	5
	6	000	000	000	003	021	074	164	251	267	176	045	008	000	6
	7	000	000	000	000	004	021	070	161	267	302	172	063	003	7
	8	000	000	000	000	000	004	018	060	156	302	387	299	083	8
	9	000	000	000	000	000	000	002	010	040	134	387	630	914	9
10	0	904	599	349	107	028	006	001	000	000	000	000	000	000	0
	1	091	315	387	268	121	040	010	002	000	000	000	000	000	1
	2	004	075	194	302	233	121	044	011	001	000	000	000	000	2
	3	000	010	057	201	267	215	117	042	009	001	000	000	000	3
	4	000	001	011	088	200	251	205	111	037	006	000	000	000	4
	5	000	000	001	026	103	201	246	201	103	026	001	000	000	5
	6	000	000	000	006	037	111	205	251	200	088	011	001	000	6
	7	000	000	000	001	009	042	117	215	267	201	057	010	000	7
	8	000	000	000	000	001	011	044	121	233	302	194	075	004	8
	9	000	000	000	000	000	002	010	040	121	268	387	315	091	9
	10	000	000	000	000	000	000	001	006	028	107	349	599	904	10
11	0	895	569	314	086	020	004	000	000	000	000	000	000	000	0
	1	099	329	384	236	093	027	005	001	000	000	000	000	000	1
	2	005	087	213	295	200	089	027	005	001	000	000	000	000	2
	3	000	014	071	221	257	177	081	023	004	000	000	000	000	3
	4	000	001	016	111	220	236	161	070	017	002	000	000	000	4
	5	000	000	002	039	132	221	226	147	057	010	000	000	000	5
	6	000	000	000	010	057	147	226	221	132	039	002	000	000	6
	7	000	000	000	002	017	070	161	236	220	111	016	001	000	7
	8	000	000	000	000	004	023	081	177	257	221	071	014	000	8
	9	000	000	000	000	001	005	027	089	200	295	213	087	005	9
	10	000	000	000	000	000	001	005	027	093	236	384	329	099	10
	11	000	000	000	000	000	000	000	004	020	086	314	569	895	11
12	0	886	540	282	069	014	002	000	000	000	000	000	000	000	0
	1	107	341	377	206	071	017	003	000	000	000	000	000	000	1
	2	006	099	230	283	168	064	016	002	000	000	000	000	000	2
	3	000	017	085	236	240	142	054	012	001	000	000	000	000	3
	4	000	002	021	133	231	213	121	042	008	001	000	000	000	4
	5	000	000	004	053	158	227	193	101	029	003	000	000	000	5
	6	000	000	000	016	079	177	226	177	079	016	000	000	000	6
	7	000	000	000	003	029	101	193	227	158	053	004	000	000	7
	8	000	000	000	001	008	042	121	213	231	133	021	002	000	8
	9	000	000	000	000	001	012	054	142	240	236	085	017	000	9
	10	000	000	000	000	000	002	016	064	168	283	230	099	006	10
	11	000	000	000	000	000	000	003	017	071	206	377	341	107	11
	12	000	000	000	000	000	000	000	002	014	069	282	540	886	12

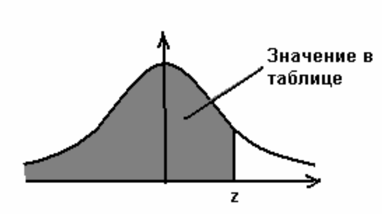
продолжение таблицы А-1

n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
13	0	878	513	254	055	010	001	000	000	000	000	000	000	000	0
	1	115	351	367	179	054	011	002	000	000	000	000	000	000	1
	2	007	111	245	268	139	045	010	001	000	000	000	000	000	2
	3	000	021	100	246	218	111	035	006	001	000	000	000	000	3
	4	000	003	028	154	234	184	087	024	003	000	000	000	000	4
	5	000	000	006	069	180	221	157	066	014	001	000	000	000	5
	6	000	000	001	023	103	197	209	131	044	006	000	000	000	6
	7	000	000	000	006	044	131	209	197	103	023	001	000	000	7
	8	000	000	000	001	014	066	157	221	180	069	006	000	000	8
	9	000	000	000	000	003	024	087	184	234	154	028	003	000	9
	10	000	000	000	000	001	006	035	111	218	246	100	021	000	10
	11	000	000	000	000	000	001	010	045	139	268	245	111	007	11
	12	000	000	000	000	000	000	002	011	054	179	367	351	115	12
	13	000	000	000	000	000	000	000	001	010	055	254	513	878	13
14	0	869	488	229	044	007	001	000	000	000	000	000	000	000	0
	1	123	359	356	154	041	007	001	000	000	000	000	000	000	1
	2	008	123	257	250	113	032	006	001	000	000	000	000	000	2
	3	000	026	114	250	194	085	022	003	000	000	000	000	000	3
	4	000	004	035	172	229	155	061	014	001	000	000	000	000	4
	5	000	000	008	086	196	207	122	041	007	000	000	000	000	5
	6	000	000	001	032	126	207	183	092	023	002	000	000	000	6
	7	000	000	000	009	062	157	209	157	062	009	000	000	000	7
	8	000	000	000	002	023	092	183	207	126	032	001	000	000	8
	9	000	000	000	000	007	041	122	207	196	086	008	000	000	9
	10	000	000	000	000	001	014	061	155	229	172	035	004	000	10
	11	000	000	000	000	000	003	022	085	194	250	114	026	000	11
	12	000	000	000	000	000	001	006	032	113	250	257	123	008	12
	13	000	000	000	000	000	000	001	007	041	154	356	359	123	13
	14	000	000	000	000	000	000	000	001	007	044	229	488	869	14
15	0	860	463	206	035	005	000	000	000	000	000	000	000	000	0
	1	130	366	343	132	031	005	000	000	000	000	000	000	000	1
	2	009	135	267	231	092	022	003	000	000	000	000	000	000	2
	3	000	031	129	250	170	063	014	002	000	000	000	000	000	3
	4	000	005	043	188	219	127	042	007	001	000	000	000	000	4
	5	000	001	010	103	206	186	092	024	003	000	000	000	000	5
	6	000	000	002	043	147	207	153	061	012	001	000	000	000	6
	7	000	000	000	014	081	177	196	118	035	003	000	000	000	7
	8	000	000	000	003	035	118	196	177	081	014	000	000	000	8
	9	000	000	000	001	012	061	153	207	147	043	002	000	000	9
	10	000	000	000	000	003	024	092	186	206	103	010	001	000	10
	11	000	000	000	000	001	007	042	127	219	188	043	005	000	11
	12	000	000	000	000	000	002	014	063	170	250	129	031	000	12
	13	000	000	000	000	000	000	003	022	092	231	267	135	009	13
	14	000	000	000	000	000	000	000	005	031	132	343	366	130	14
	15	000	000	000	000	000	000	000	000	005	035	206	463	860	15

окончание таблицы А-1

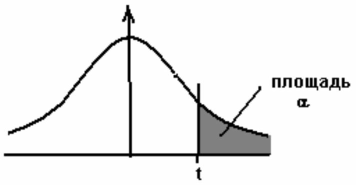
n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
16	0	851	440	185	028	003	000	000	000	000	000	000	000	000	0
	1	138	371	329	113	023	003	000	000	000	000	000	000	000	1
	2	010	146	275	211	073	015	002	000	000	000	000	000	000	2
	3	000	036	142	246	146	047	009	001	000	000	000	000	000	3
	4	000	006	051	200	204	101	028	004	000	000	000	000	000	4
	5	000	001	014	120	210	162	067	014	001	000	000	000	000	5
	6	000	000	003	055	165	198	122	039	006	000	000	000	000	6
	7	000	000	000	020	101	189	175	084	019	001	000	000	000	7
	8	000	000	000	006	049	142	196	142	049	006	000	000	000	8
	9	000	000	000	001	019	084	175	189	101	020	000	000	000	9
	10	000	000	000	000	006	039	122	198	165	055	003	000	000	10
	11	000	000	000	000	001	014	067	162	210	120	014	001	000	11
	12	000	000	000	000	000	004	028	101	204	200	051	006	000	12
	13	000	000	000	000	000	001	009	047	146	246	142	036	000	13
	14	000	000	000	000	000	000	002	015	073	211	275	146	010	14
	15	000	000	000	000	000	000	000	003	023	113	329	371	138	15
	16	000	000	000	000	000	000	000	000	003	028	185	440	851	16
17	0	843	418	167	023	002	000	000	000	000	000	000	000	000	0
	1	145	374	315	096	017	002	000	000	000	000	000	000	000	1
	2	012	158	280	191	058	010	001	000	000	000	000	000	000	2
	3	001	041	156	239	125	034	005	000	000	000	000	000	000	3
	4	000	008	060	209	187	080	018	002	000	000	000	000	000	4
	5	000	001	017	136	208	138	047	008	001	000	000	000	000	5
	6	000	000	004	068	178	184	094	024	003	000	000	000	000	6
	7	000	000	001	027	120	193	148	057	009	000	000	000	000	7
	8	000	000	000	008	064	161	185	107	028	002	000	000	000	8
	9	000	000	000	002	028	107	185	161	064	008	000	000	000	9
	10	000	000	000	000	009	057	148	193	120	027	001	000	000	10
	11	000	000	000	000	003	024	094	184	178	068	004	000	000	11
	12	000	000	000	000	001	008	047	138	208	136	017	001	000	12
	13	000	000	000	000	000	002	018	080	187	209	060	008	000	13
	14	000	000	000	000	000	000	005	034	125	239	156	041	001	14
	15	000	000	000	000	000	000	001	010	058	191	280	158	012	15
	16	000	000	000	000	000	000	000	002	017	096	315	374	145	16
	17	000	000	000	000	000	000	000	000	002	023	167	418	843	17
18	0	835	397	150	018	002	000	000	000	000	000	000	000	000	0
	1	152	376	300	081	013	001	000	000	000	000	000	000	000	1
	2	013	168	284	172	046	007	001	000	000	000	000	000	000	2
	3	001	047	168	230	105	025	003	000	000	000	000	000	000	3
	4	000	009	070	215	168	061	012	001	000	000	000	000	000	4
	5	000	001	022	151	202	115	033	004	000	000	000	000	000	5
	6	000	000	005	082	187	166	071	015	001	000	000	000	000	6
	7	000	000	001	035	138	189	121	037	005	000	000	000	000	7
	8	000	000	000	012	081	173	167	077	015	001	000	000	000	8
	9	000	000	000	003	039	128	185	128	039	003	000	000	000	9
	10	000	000	000	001	015	077	167	173	081	012	000	000	000	10
	11	000	000	000	000	005	037	121	189	138	035	001	000	000	11
	12	000	000	000	000	001	015	071	166	187	082	005	000	000	12
	13	000	000	000	000	000	004	033	115	202	151	022	001	000	13
	14	000	000	000	000	000	001	012	061	168	215	070	009	000	14
	15	000	000	000	000	000	000	003	025	105	230	168	047	001	15
	16	000	000	000	000	000	000	001	007	046	172	284	168	013	16
	17	000	000	000	000	000	000	000	001	013	081	300	376	152	17
	18	000	000	000	000	000	000	000	000	002	018	150	397	835	18

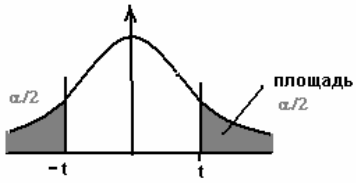
Таблица А-2. Нормальное распределение



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	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	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	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	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
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

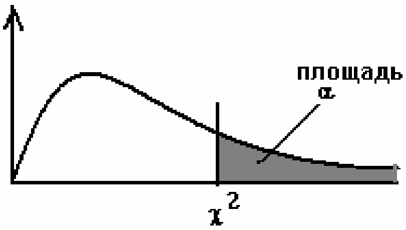
Таблица А-3. Критические точки распределения Стьюдента





Df	Односторонняя область				
	0,005	0,01	0,025	0,05	0,10
	Двусторонняя область				
	0,01	0,02	0,05	0,10	0,20
1	63,656	31,821	12,706	6,314	3,078
2	9,925	6,965	4,303	2,920	1,886
3	5,841	4,541	3,182	2,353	1,638
4	4,604	3,747	2,776	2,132	1,533
5	4,032	3,365	2,571	2,015	1,476
6	3,707	3,143	2,447	1,943	1,440
7	3,499	2,998	2,365	1,895	1,415
8	3,355	2,896	2,306	1,860	1,397
9	3,250	2,821	2,262	1,833	1,383
10	3,169	2,764	2,228	1,812	1,372
11	3,106	2,718	2,201	1,796	1,363
12	3,055	2,681	2,179	1,782	1,356
13	3,012	2,650	2,160	1,771	1,350
14	2,977	2,624	2,145	1,761	1,345
15	2,947	2,602	2,131	1,753	1,341
16	2,921	2,583	2,120	1,746	1,337
17	2,898	2,567	2,110	1,740	1,333
18	2,878	2,552	2,101	1,734	1,330
19	2,861	2,539	2,093	1,729	1,328
20	2,845	2,528	2,086	1,725	1,325
21	2,831	2,518	2,080	1,721	1,323
22	2,819	2,508	2,074	1,717	1,321
23	2,807	2,500	2,069	1,714	1,319
24	2,797	2,492	2,064	1,711	1,318
25	2,787	2,485	2,060	1,708	1,316
26	2,779	2,479	2,056	1,706	1,315
27	2,771	2,473	2,052	1,703	1,314
28	2,763	2,467	2,048	1,701	1,313
29	2,756	2,462	2,045	1,699	1,311
30	2,750	2,457	2,042	1,697	1,310
40	2,704	2,423	2,021	1,684	1,303
60	2,660	2,390	2,000	1,671	1,296
100	2,626	2,364	1,984	1,660	1,290
200	2,601	2,345	1,972	1,653	1,286
300	2,592	2,339	1,968	1,650	1,284
400	2,588	2,336	1,966	1,649	1,284
500	2,586	2,334	1,965	1,648	1,283

Таблица А-4. Критические точки распределения хи-квадрат (χ^2)



df	0,995	0,99	0,95	0,90	0,10	0,05	0,01	0,005
1	0,000	0,000	0,004	0,016	2,706	3,841	6,635	7,879
2	0,010	0,020	0,103	0,211	4,605	5,991	9,210	10,597
3	0,072	0,115	0,352	0,584	6,251	7,815	11,345	12,838
4	0,207	0,297	0,711	1,064	7,779	9,488	13,277	14,860
5	0,412	0,554	1,145	1,610	9,236	11,070	15,086	16,750
6	0,676	0,872	1,635	2,204	10,645	12,592	16,812	18,548
7	0,989	1,239	2,167	2,833	12,017	14,067	18,475	20,278
8	1,344	1,647	2,733	3,490	13,362	15,507	20,090	21,955
9	1,735	2,088	3,325	4,168	14,684	16,919	21,666	23,589
10	2,156	2,558	3,940	4,865	15,987	18,307	23,209	25,188
11	2,603	3,053	4,575	5,578	17,275	19,675	24,725	26,757
12	3,074	3,571	5,226	6,304	18,549	21,026	26,217	28,300
13	3,565	4,107	5,892	7,041	19,812	22,362	27,688	29,819
14	4,075	4,660	6,571	7,790	21,064	23,685	29,141	31,319
15	4,601	5,229	7,261	8,547	22,307	24,996	30,578	32,801
16	5,142	5,812	7,962	9,312	23,542	26,296	32,000	34,267
17	5,697	6,408	8,672	10,085	24,769	27,587	33,409	35,718
18	6,265	7,015	9,390	10,865	25,989	28,869	34,805	37,156
19	6,844	7,633	10,117	11,651	27,204	30,144	36,191	38,582
20	7,434	8,260	10,851	12,443	28,412	31,410	37,566	39,997
21	8,034	8,897	11,591	13,240	29,615	32,671	38,932	41,401
22	8,643	9,542	12,338	14,041	30,813	33,924	40,289	42,796
23	9,260	10,196	13,091	14,848	32,007	35,172	41,638	44,181
24	9,886	10,856	13,848	15,659	33,196	36,415	42,980	45,558
25	10,520	11,524	14,611	16,473	34,382	37,652	44,314	46,928
26	11,160	12,198	15,379	17,292	35,563	38,885	45,642	48,290
27	11,808	12,878	16,151	18,114	36,741	40,113	46,963	49,645
28	12,461	13,565	16,928	18,939	37,916	41,337	48,278	50,994
29	13,121	14,256	17,708	19,768	39,087	42,557	49,588	52,335
30	13,787	14,953	18,493	20,599	40,256	43,773	50,892	53,672
40	20,707	22,164	26,509	29,051	51,805	55,758	63,691	66,766
50	27,991	29,707	34,764	37,689	63,167	67,505	76,154	79,490
60	35,534	37,485	43,188	46,459	74,397	79,082	88,379	91,952
70	43,275	45,442	51,739	55,329	85,527	90,531	100,425	104,215
80	51,172	53,540	60,391	64,278	96,578	101,879	112,329	116,321
90	59,196	61,754	69,126	73,291	107,565	113,145	124,116	128,299
100	67,328	70,065	77,929	82,358	118,498	124,342	135,807	140,170

Таблица А-5. Критические точки распределения Фишера

$\alpha=0{,}05$

		Степени свободы df_1																		
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
Степени свободы df_2	**1**	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
	2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
	3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
	4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
	8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
	11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
	13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
	16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
	21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
	23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
	25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
	26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
	27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
	28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
	29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
	120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25

α=0,01

Степени свободы df_2	Степени свободы df_1																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Таблица А-6. Критические значения для критерия знаков

n	α=0,05	α=0,1
9	1	1
10	1	1
11	1	2
12	2	2
13	2	3
14	2	3
15	3	3
16	3	4
17	4	4
18	4	5
19	4	5
20	5	5
25	7	7
30	9	10

Примечание. Указанное в таблице значение является границей критической области $x \le x_{кр}$

Таблица А-7. Критические значения для знако-рангового критерия

n	α=0,05	α=0,1
9	8	3
10	11	5
11	14	7
12	17	10
13	21	13
14	26	16
15	30	20
16	36	24
17	41	28
18	47	33
19	54	38
20	60	43
25	101	77
30	152	120

Примечание. Указанное в таблице значение является границей критической области $x \leq x_{кр}$

Таблица А-8. Критические значения для критерия Манна-Уитни

$\alpha = 0{,}05$

k1/k2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	56	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	96
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

$\alpha = 0{,}01$

k1/k2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																			0	0
3									0	0	0	1	1	1	2	2	2	2	3	3
4						0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5					0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6				0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7				0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8				1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9			0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10			0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11			0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12			1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13			1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14			1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15			2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16			2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17			2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18			2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19		0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20		0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

Таблица А-9. Критические значения для критерия Вилкоксона

α = 0,01

k1/k2	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
6	24	25	27	28	29	30	32	33	34	36	37	39	40	41	43	44	45	47	48	50
7	25	34	35	37	39	40	42	44	45	47	49	51	52	54	56	58	59	61	63	64
8	27	35	45	47	49	51	53	56	58	60	62	64	66	68	70	72	74	76	78	81
9	28	37	47	59	61	63	66	68	71	73	76	78	81	83	85	88	90	93	95	98
10	29	39	49	61	74	77	79	82	85	88	91	93	96	99	102	105	108	110	113	116
11	30	40	51	63	77	91	94	97	100	103	107	110	113	116	119	123	126	129	132	136
12	32	42	53	66	79	94	109	113	116	120	124	127	131	134	138	142	145	149	153	156
13	33	44	56	68	82	97	113	130	134	138	142	146	150	154	158	162	166	170	174	178
14	34	45	58	71	85	100	116	134	152	156	161	165	170	174	178	183	187	192	196	200
15	36	47	60	73	88	103	120	138	156	176	181	186	190	195	200	205	210	214	219	224
16	37	49	62	76	91	107	124	142	161	181	202	207	212	218	223	228	233	238	244	249
17	39	51	64	78	93	110	127	146	165	186	207	230	235	241	246	252	258	263	269	275
18	40	52	66	81	96	113	131	150	170	190	212	235	259	265	271	277	283	289	295	301
19	41	54	68	83	99	116	134	154	174	195	218	241	265	291	297	303	310	316	323	329
20	43	56	70	85	102	119	138	158	178	200	223	246	271	297	324	331	337	344	351	358
21	44	58	72	88	105	123	142	162	183	205	228	252	277	303	331	359	366	373	381	388
22	45	59	74	90	108	126	145	166	187	210	233	258	283	310	337	366	396	403	411	419
23	47	61	76	93	110	129	149	170	192	214	238	263	289	316	344	373	403	434	443	451
24	48	63	78	95	113	132	153	174	196	219	244	269	295	323	351	381	411	443	475	484
25	50	64	81	98	116	136	156	178	200	224	249	275	301	329	358	388	419	451	484	517

α = 0,05

k1/k2	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
6	28	30	31	33	35	37	38	40	42	44	46	47	49	51	53	55	57	58	60	62
7	30	39	41	43	45	47	49	52	54	56	58	61	63	65	67	69	72	74	76	78
8	31	41	51	54	56	59	62	64	67	69	72	75	77	80	83	85	88	90	93	96
9	33	43	54	66	69	72	75	78	81	84	87	90	93	96	99	102	105	108	111	114
10	35	45	56	69	82	86	89	92	96	99	103	106	110	113	117	120	123	127	130	134
11	37	47	59	72	86	100	104	108	112	116	120	123	127	131	135	139	143	147	151	155
12	38	49	62	75	89	104	120	125	129	133	138	142	146	150	155	159	163	168	172	176
13	40	52	64	78	92	108	125	142	147	152	156	161	166	171	175	180	185	189	194	199
14	42	54	67	81	96	112	129	147	166	171	176	182	187	192	197	202	207	212	218	223
15	44	56	69	84	99	116	133	152	171	192	197	203	208	214	220	225	231	236	242	248
16	46	58	72	87	103	120	138	156	176	197	219	225	231	237	243	249	255	261	267	273
17	47	61	75	90	106	123	142	161	182	203	225	249	255	262	268	274	281	287	294	300
18	49	63	77	93	110	127	146	166	187	208	231	255	280	287	294	301	307	314	321	328
19	51	65	80	96	113	131	150	171	192	214	237	262	287	313	320	328	335	342	350	357
20	53	67	83	99	117	135	155	175	197	220	243	268	294	320	348	356	364	371	379	387
21	55	69	85	102	120	139	159	180	202	225	249	274	301	328	356	385	393	401	410	418
22	57	72	88	105	123	143	163	185	207	231	255	281	307	335	364	393	424	432	441	450
23	58	74	90	108	127	147	168	189	212	236	261	287	314	342	371	401	432	465	474	483
24	60	76	93	111	130	151	172	194	218	242	267	294	321	350	379	410	441	474	507	517
25	62	78	96	114	134	155	176	199	223	248	273	300	328	357	387	418	450	483	517	552

Таблица А-10. Критические значения для коэффициента ранговой корреляции Спирмена

n	α = 0,10	α = 0,05	α = 0,01
5	0,900	-	-
6	0,829	0,886	-
7	0,714	0,786	-
8	0,643	0,738	0,881
9	0,600	0,683	0,833
10	0,564	0,648	0,794
11	0,523	0,623	0,818
12	0,497	0,591	0,780
13	0,475	0,566	0,745
14	0,457	0,545	0,716
15	0,441	0,525	0,689
16	0,425	0,507	0,666
17	0,412	0,490	0,645
18	0,399	0,476	0,625
19	0,388	0,462	0,608
20	0,377	0,450	0,591
21	0,368	0,438	0,576
22	0,359	0,428	0,562
23	0,351	0,418	0,549
24	0,343	0,409	0,537
25	0,336	0,400	0,526
26	0,329	0,392	0,515
27	0,323	0,385	0,505
28	0,317	0,377	0,496
29	0,311	0,370	0,487
30	0,305	0,364	0,478

Таблица А-11 Критические значения для коэффициента ранговой корреляции Кендалла

n	α = 0,10	α = 0,05	α = 0,01
4	0,667	1,000	
5	0,600	0,800	1,000
6	0,467	0,600	0,867
7	0,429	0,619	0,810
8	0,429	0,500	0,714
9	0,389	0,444	0,611
10	0,333	0,422	0,600
11	0,309	0,418	0,564
12	0,303	0,394	0,515
13	0,282	0,359	0,487
14	0,275	0,341	0,473
15	0,257	0,333	0,448
16	0,250	0,317	0,433
17	0,235	0,309	0,426
18	0,229	0,294	0,412
19	0,216	0,287	0,392
20	0,211	0,274	0,379
21	0,210	0,267	0,371
22	0,203	0,255	0,359
23	0,194	0,249	0,352
24	0,188	0,246	0,341
25	0,187	0,240	0,333
26	0,182	0,231	0,323
27	0,179	0,225	0,316
28	0,175	0,222	0,312
29	0,172	0,222	0,305
30	0,168	0,214	0,301
31	0,166	0,211	0,295
32	0,161	0,206	0,290
33	0,159	0,205	0,284
34	0,155	0,201	0,280
35	0,153	0,197	0,277
36	0,152	0,194	0,273
37	0,150	0,189	0,267
38	0,147	0,189	0,263
39	0,144	0,185	0,258
40	0,144	0,182	0,256

Таблица А-12 Критические значения для критерия Фридмана

k	n	α = 0,10	α = 0,05	α = 0,01
3	2	4,000		
3	3	4,667	6,000	6,000
3	4	4,500	6,500	8,000
3	5	5,200	6,400	8,400
3	6	5,333	6,333	9,000
3	7	4,571	6,000	8,857
3	8	4,750	6,250	9,000
3	9	4,667	6,222	8,667
3	10	5,000	6,200	8,600
3	11	4,909	6,545	8,909
3	12	4,667	6,167	8,667
3	13	4,769	6,000	9,385
3	14	5,143	6,143	9,000
3	15	4,933	6,400	8,933
3	16	4,875	6,125	9,125
3	17	4,588	6,118	8,941
3	18	4,778	6,333	9,000
3	19	5,053	6,000	8,842
3	20	4,900	6,100	9,100
3	21	4,667	6,000	8,857
3	22	4,727	5,818	9,091
3	23	4,522	5,826	9,391
3	24	4,750	6,083	9,083
3	25	4,880	6,080	9,333
4	2	5,400	6,000	
4	3	6,600	7,000	8,200
4	4	6,000	7,500	9,300
4	5	6,360	7,320	9,960
4	6	6,400	7,400	10,200
4	7	6,257	7,629	10,371
4	8	6,300	7,650	10,500
4	9	6,200	7,667	10,467
4	10	6,240	7,680	10,680
5	2	6,800	7,600	8,000
5	3	7,467	8,533	10,133
5	4	7,600	8,800	11,000
5	5	7,680	8,960	11,680
5	6	7,600	9,067	11,867
5	7	7,057	9,114	12,114
5	8	7,700	9,200	12,300
6	2	8,286	8,857	9,714
6	3	8,714	9,857	11,762
6	4	8,857	10,143	12,714
6	5	9,000	10,371	13,229
6	6	9,048	10,517	13,619
6	7	9,200	10,476	14,100
6	8	9,000	10,790	13,860
7	7	7,710	9,550	13,810
7	8	7,850	9,780	13,690

Приложение В. Ответы к задачам

9-1. (302,2; 317,8)
9-2. (37 940; 42 060)
9-3.а. (22,7; 23,3)
9-3.б. (22,68; 23,32)
9-4. (12,4%; 27,6%)
9-5. (23,8; 29,8)
9-6. (9,9; 14,1)
9-7. (25,1; 30,9)
9-8. больше 28
9-9. (49,4%; 65,2%)
9-10. (34,3%; 45,7%)
9-11. (54,5%; 65,5%)
9-12. больше 1418
9-13. больше 8454

В этой версии учебника задачи к другим главам не приводятся.

Приложение С. Библиография

Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М. ЮНИТИ-ДАНА, 2001.

Вентцель Е.С. Теория вероятностей. – М. Физматгиз (Любое издание).

Гмурман В.Е. Теория вероятностей и математическая статистика. – М. Высшая школа (Любое издание).

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М. Высшая школа (Любое издание).

Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. – М. Наука (Любое издание).

Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики. Учебник. – М. Инфра-М, 1999.

Калинина В.Н., Панкин В.Ф. Математическая статистика. – М. Высшая школа, 1998.

Малхотра Нэреш. Маркетинговые исследования. – М. Вильямс, 2003.

Рабочая книга социолога. Под ред. Осипова Г.В. – М. Едиториал УРСС, 2003.

Сигел Эндрю. Практическая бизнес-статистика. – М. Вильямс, 2004.

Теория статистики. Учебник. Под ред. Шмойловой Р.А. – М. Финансы и статистика, 2002.

Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. Под ред Фигурнова В.Э. - М. Инфра-М, 2003.

Сулицкий В.Н. Методы статистического анализа в управлении. – М. Дело, 2002.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

СОДЕРЖАНИЕ

СВЕДЕНИЯ ОБ АВТОРЕ

Иванов Олег Валентинович, заведующий кафедрой социальной информатики социологического факультета МГУ им.М.В.Ломоносова, кандидат физико-математических наук.

На социологическом факультете МГУ в течение ряда лет руководит преподаванием цикла математических дисциплин, включающего основы математического анализа, теорию вероятностей, математическую статистику, информатику, теорию измерений и анализ данных.