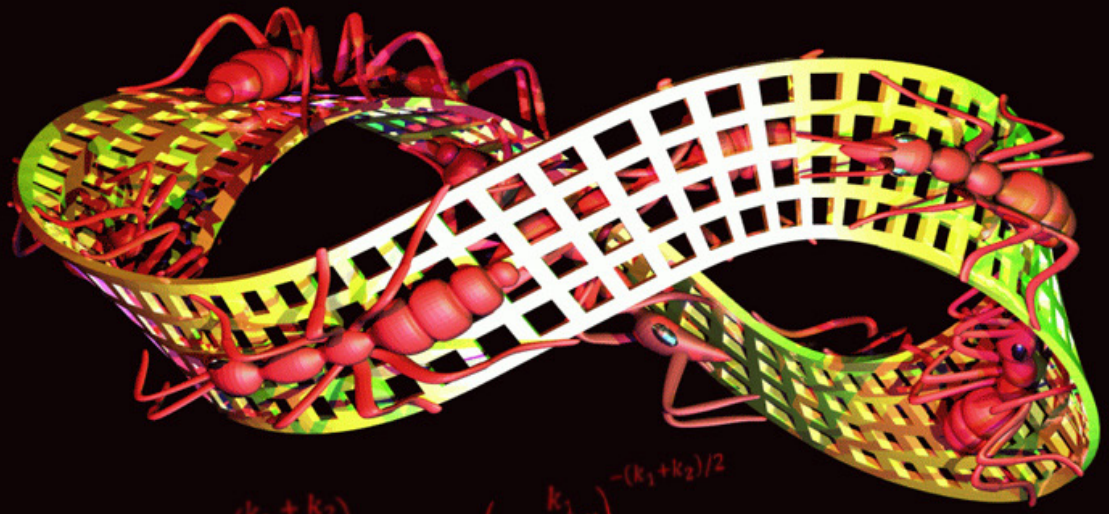
ОСНОВЫ ПРАКТИЧЕСКОЙ БИОСТАТИСТИКИ



$$P(x) = \left(\frac{k_1}{k_2}\right)^{k_1/2} \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} x^{(k_1-2)/2} \left(1+\frac{k_1}{k_2}x\right)^{-(k_1+k_2)/2}$$

$$P(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\left(\frac{k}{2}-1\right)} exp\left(-\frac{x}{2}\right)$$

$$P(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left[1+\frac{t^2}{k}\right]^{-\frac{k+1}{2}}$$

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right)$$

Титульный экран

Аннотация

Содержание

Используйте Acrobat Reader версии 9 и выше. Для входа на сайт производителя для бесплатного скачивания щелкните по данному тексту

Методические указания

Глоссарий предметный

Глоссарий общенаучный

Справочный раздел

✓ Критические значения *U*-критерия Манна-Уитни	Таблицы М-У
✓ Критические значения статистики Колмогорова-Смирнова	Таблицы К-С
✓ Соответствие наименований некоторых функций ✓ Пакет "Анализ данных" Microsoft Excel ✓ Статистические функции электронных таблиц	Функции ЭТ
✓ Виды ошибок при задании формул в Excel и Calc	Ошибки ЭТ
✓ Список дополнительной (рекомендуемой) литературы ✓ Интернет порталы по статистическим исследованиям ✓ Программное обеспечение по статистике и анализу данных	Допматериалы
✓ Задания для самостоятельной работы	Задания
✓ Вопросы для тестирования	Тесты
✓ Исходные данные приведенных примеров	Исходные данные

Содержание

ПРИЛОЖЕНИЯ, их содержание	текст приложения
ПРИЛОЖЕНИЕ 1. Типовые распределения П1.1. Нормальное распределение П1.2. Распределение хи-квадрат П1.3. Распределение Стьюдента П1.4. Распределение Фишера	Приложение 1
ПРИЛОЖЕНИЕ 2. Предварительная обработка данных (отсев грубых погрешностей) П2.1. Правило "ящик с усами" П2.2. Правило Томпсона (критерий Рошера)	Приложение 2
ПРИЛОЖЕНИЕ 3. Полигон и гистограмма частостей, эмпирическая функция распределения	Приложение 3
ПРИЛОЖЕНИЕ 4. Расчет параметров линейного уравнения регрессии методом наименьших квадратов	Приложение 4
ПРИЛОЖЕНИЕ 5. Проверка гипотезы о независимости двух выборок	Приложение 5
ПРИЛОЖЕНИЕ 6. Проверка выборки на нормальность (метод моментов)	Приложение 6
ПРИЛОЖЕНИЕ 7. Статистические оценки для интервального ранжированного частотного ряда	Приложение 7
ПРИЛОЖЕНИЕ 8. Отчет о полевой практике (пример оформления)	Приложение 8

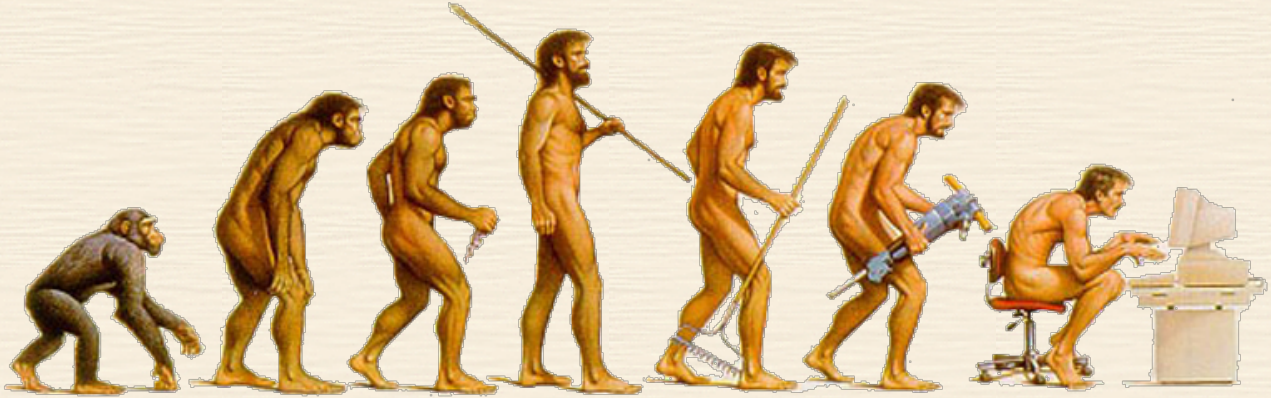
Каталог примеров

номер примера	содержание примера, ссылка в тексте	текст примера
1.1.	Проверка соответствия выборки нормальному закону распределения. Используется критерий χ^2 сравнения непрерывных экспериментальных данных с теоретическим (нормальным) распределением	Нормальность распредел. по χ^2
1.2.	Исследование цветовых предпочтений лабораторной крысы. Сравнение дискретных экспериментальных данных с теоретическим (равномерным) распределением. Используется критерий χ^2	Сравнение распределений по χ^2
1.3.	Моногибридное скрещивание, анализ гибридов второго поколения. Доказывается, что количества мух – особей с рецессивным и с доминантным признаками, равно отношению 3:1. Сравнение дискретных экспериментальных данных с заданным распределением. Используется критерий χ^2	Сравнение распределений по χ^2
1.4.	Ди- и полигибридное скрещивание, анализ гибридов второго поколения. Доказывается, что среди мух (особей как с парами альтернативных рецессивных, так и доминантных признаков) устанавливающиеся отношения расщепления фенотипических классов равны 9:3:3:1. Сравнение дискретных экспериментальных данных с заданным распределением. Используется критерий χ^2	Сравнение распределений по χ^2

1.5.	Тестирование остаточных знаний группы студентов проведением контрольных работ с разрывом месяц и оцениванием по пятибалльной системе. Анализируется вопрос изменения знаний студентов (по частотам отметок). Используется критерий χ^2 соответствия (сравнение с "прошлыми" событиями)	Соответствие по критерию χ^2
1.6	Анализируется вопрос изменения распределения по кариотипам *Anopheles* за три года. Проверка наличия взаимосвязи между двумя качественными переменными (критерий χ^2 независимости)	Взаимосвязь по критерию χ^2
1.7	Анализ сопряженности двух качественных признаков для таблицы 2×2. Проверка наличия взаимосвязи между двумя качественными переменными (критерий χ^2 независимости)	Независимость табл. 2×2 по χ^2
1.8.	Тестирование индекса массы двух групп особей, исследуется вопрос дисперсной однородности показателей индекса между группами. Сравнение двух групп непрерывных экспериментальных данных с использованием критерия Фишера	Однородность по Фишеру
2.1.	Сравнение среднего размера листьев у двух ярусов кроны акации. Используется t-критерий Стьюдента непрерывных данных для несвязных и неравных по численности равнодисперсных выборок	Стьюдент равнодисперсных

2.2.	Изучение воздействия (эффективности) препарата на группу особей. Используется *t*-критерий Стьюдента непрерывных данных для парных выборок	Стьюдент парных
2.3.	Сравнение среднего размера листьев у двух ярусов кроны акации. Используется *t*-критерий Стьюдента (критерий Беренса-Фишера) непрерывных данных для несвязных и неравных по численности разнодисперсных выборок	Стьюдент разнодисперсных
2.4.	Сравнение уровня какого-либо признака двух выборок с непрерывными данными. Используется *U*-критерий Манна-Уитни	Критерий Манна-Уитни
2.5.	Сравниваются функции распределения листьев по размерам у двух ярусов кроны акации. Используется критерий Смирнова (Колмогорова-Смирнова)	Критерий КС
П2.1.	Отсев грубых погрешностей непрерывных данных. Используется правило "ящик с усами"	Отсев "ящик"
П2.2.	Отсев грубых погрешностей непрерывных данных. Используется правило Томпсона (критерий Рошера)	Отсев Томпсон
П3.	Построение эмпирической функции распределения, полигона и гистограммы частостей непрерывных данных	Распределение
П4.	Расчет параметров линейного уравнения регрессии непрерывных данных методом наименьших квадратов	Регрессия

П5.	Проверка гипотезы о независимости двух выборок. Проверка независимости для выборок с произвольным числом классов (разрядов, признаков). Используется критерий χ^2 независимости	Гипотеза независимости
П6.	Проверка соответствия выборки нормальному закону распределения. Используется оценка параметров по моментам распределения – асимметрии и эксцессу	Нормальность распределения
П7.	Вычисление статистических оценок для интервального ранжированного частотного ряда	Статоценки
П8.	Отчет о полевой практике (пример оформления)	О т ч е т



К мудрости путь – по ухабам ошибок
– иди же и носа не вешай:
ушибы, ушибы и снова ушибы,
Но реже, и реже, и реже.

Поль Хайн, "Путь к мудрости"

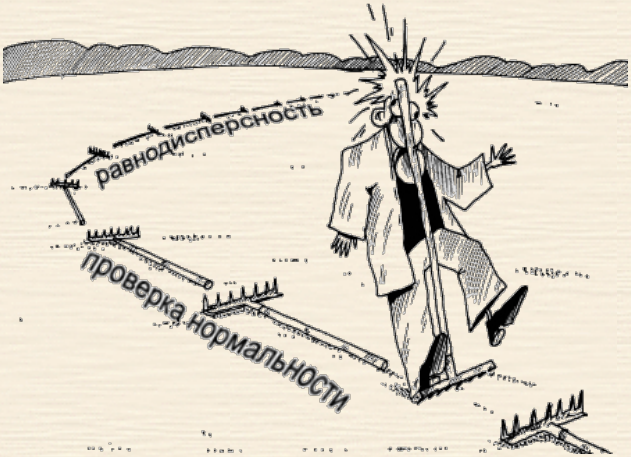
Предисловие

При разработке настоящего учебного пособия решались две главные задачи: дать студентам краткий обзор базовых понятий и методов статистики (основные принятые определения и формулы); а также отразить содержание курса "Основы биологической статистики[1]". Особое внимание уделено поясняющим примерам по темам, традиционно вызывающим затруднение при освоении материала ("Даже маленькая практика стоит большой теории", – мерфология, закон Буккера). Кроме того, автор надеется, что пособие явится стартовым комплексом для углубленного изучения вопроса.

В дополнительных материалах указаны некоторые интернет порталы по статистическим исследованиям, адреса сайтов с программным обеспечением по статистике и анализу данных. Там же приводится список дополнительной (рекомендуемой) литературы. А бóльшую часть вопросов снимут два глоссария.

Автор надеется, что на рассмотренных простых примерах типовых задач анализа данных студент сознательно или пусть даже неосознанно познакомится с практическими приемами статистических вычислений, поскольку "в любой науке столько истины, сколько в ней математики" (Иммануил Кант), а целью математики, в конечном итоге, является осмысление действительности.

[1] За основу предлагаемого взят и переработан материал учебного пособия Бондарчук С.С. Основы практической биостатистики / С.С. Бондарчук, И.Г. Годованная, В.П. Перевозкин. Томск: Издательство ТГПУ. – 132 с.



Автор считает, что решение предлагаемых простых задач обработки данных инструментарием электронных таблиц позволит учащимся "прочувствовать" числовой материал и методы, а освоение курса обеспечит в дальнейшем обучении легкость восприятия и теории статистики, и квалифицированное использование специализированного программного обеспечения. "*Adde parvum parvo magnus acertus erit*" ("Добавляя малое к малому, получишь большую кучу", – Овидий).

Теоретический материал пособия иллюстрируется подобранными подробными решениями примеров. Величайший Гений времен и народов сэр Исаак Ньютон говорил, что *"при изучении наук примеры полезнее правил"*. Умение решать задачи означает не только знание теории, но и способность использовать имеющееся знание для получения нового.

Кроме нескольких примеров решения задач по популяционной генетике одно из приложений пособия содержит образец отчета лабораторной работы по этой теме, что, помимо познавательного смысла, иллюстрируют способы наглядного и корректного оформления.

Старинная мудрость гласит: "Чем больше учишь, тем меньше знаешь". Надеюсь, что вдумчивый студент, изучив данную тему, поймет, что теперь он не знает больше, чем не знал раньше.

Автор

Статистика – в высшей мере логичный и точный метод, позволяющий весьма уклончиво формулировать полуправду.
(Из постулатов NASA)

Статистику (немец. *Statistik*, от латинского *Status* – состояние) можно рассматривать как науку *о методах изучения массовых явлений.* Некоторые процессы, наблюдаемые в *массовом* количестве, обнаруживают определенные *закономерности*, которые, однако, невозможно заметить в отдельном случае или же при небольшом числе наблюдений.

1. Статистические методы обработки результатов

В данном разделе приводятся начальные сведения и простые методы решения следующих задач биологической статистики.

1. Выявление различий на уровне исследуемого признака для двух выборок (*t*–критерий Стьюдента и *U*–критерий Манна-Уитни).
2. Оценка сдвига значений исследуемого признака для двух замеров на одной и той же выборке (*t* критерий Стьюдента и *U* – критерий Манна-Уитни).

3. Выявление различий в распределении признака при сопоставлении эмпирического распределения с теоретическим (χ^2-критерий Пирсона).
4. Выявление различий в распределении признака при сопоставлении двух эмпирических распределений (χ^2–критерий Пирсона; F -критерий Фишера).

В настоящем пособии представлены, естественно, не все методы биологической статистики. Более подробно с ними можно познакомиться в дополнительной рекомендуемой литературе*.

В разделах курса и в приложениях приведено достаточное количество примеров для практического применения излагаемых методов в среде Excel (Microsoft Office) или Calc (Open Office).

Рядом с наименованиями функций для русифицированных пакетов в фигурных скобках приводятся их английские аналоги.

Вспомогательные материалы по электронным таблицам приведены в разделах:

статистические функции электронных таблиц	Функции ЭТ
виды ошибок при задании формул в Excel и Calc	Ошибки ЭТ

* Обширная библиография и библиотека вопроса приведены на сайте http://www.biometrica.tomsk.ru (главный редактор В.П. Леонов)

Ученый, ты объясняешь нам науку,
но кто объяснит нам твое объяснение?
лорд Байрон

1.1. Общие определения

Статистической гипотезой называется всякое непротиворечивое множество утверждений $\{H_0, H_1, \dots, H_{k-1}\}$ относительно свойств распределения случайной величины.

Любое из утверждений H_i называется альтернативой гипотезы H_0. Простейшей гипотезой является двухальтернативная $\{H_0, H_1\}$. В этом случае альтернативу H_0 называют нулевой гипотезой, а H_1 – конкурирующей гипотезой.

При проверке гипотез можно допустить ошибки двух родов.

Ошибка первого рода состоит в том, что будет отклонена гипотеза H_0, если она верна ("пропуск цели").

Вероятность совершить ошибку первого рода обозначается α и называется *уровнем значимости*.

Наиболее часто на практике принимают, что $\alpha = 0{,}05$ или $\alpha = 0{,}01$.

Ошибка второго рода заключается в том, что гипотеза H_0 принимается, если она неверна ("ложное срабатывание").

Вероятность ошибки этого рода обычно обозначается β.

Критерием согласия называется критерий проверки гипотезы о предполагаемом законе распределения.

Одной из типичных задач статистического анализа, решаемой после определения основных (выборочных) характеристик и анализа каждой конкретной совокупности, является совместный анализ двух или нескольких выборок, определение их сходства или различия. Обычно для этого проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве средних.

Критерии (тесты сравнения) статистических совокупностей делятся на две группы: параметрические и непараметрические. Особенностью параметрических критериев является предположение о том, что распределение признака в генеральной совокупности подчиняется некоторому известному, например, нормальному закону (см. приложение 1). Нормальность эмпирического распределения выборки желательно или даже обязательно проверять до применения любого параметрического теста с помощью какого-либо метода.

Если вид распределения или функция распределения выборки известны или заданы, то в этом случае задача оценки различий двух групп независимых наблюдений решается с использованием *параметрических критериев* статистики.

Из параметрических наиболее часто употребляемыми являются критерий Стьюдента *t*, если сравнение выборок ведется по средним значениям ($\overline{x}_1$ и $\overline{x}_2$), либо критерий Фишера *F*, если сравнение выборок ведется по их дисперсиям.

Использование параметрических критериев статистики без предварительной проверки вида распределения может привести к значимым ошибкам в ходе проверки рабочей гипотезы.

Существует большое количество опытных данных, которые не показывают нормальности распределения, поэтому применение параметрических критериев для их анализа не может быть обоснованным.

При несоответствии данных какому-либо закону распределения в практике исследований используются *непараметрические критерии статистики*, такие, например, как критерий знаков, двухвыборочный критерий Вилкоксона, критерии Манна-Уитни, Смирнова, выбор которых хотя и не требует большого объема выборки и знания вида распределения, но, тем не менее, зависит от ряда условий. Непараметрические критерии статистики свободны от допущения о законе распределения выборок и базируются на предположении о независимости наблюдений.

Практически ценными явились робастные методы, которые применимы в широком диапазоне условий. Робастные, непараметрические и свободные от распределения процедуры традиционно относят к одному классу, хотя есть и альтернативные мнения.

Под робастностью обычно понимается слабая чувствительность к отклонениям от стандартных условий (например, эмпирическое распределение может отличаться от теоретического нормального); методы, применимые в диапазоне реальных условий, принято называть робастными.

Непараметрические критерии не требуют предварительных предположений относительно вида исходного распределения и являются более робастными, чем их параметрические аналоги. Их называют также критериями значимости, независимыми от типа распределения. Естественно, непараметрические критерии применимы и для случая нормального распределения. Однако непараметрические критерии в большинстве случаев являются менее мощными, чем их параметрические аналоги. Если существуют предпосылки использования параметрических критериев, но используются непараметрические, увеличивается вероятность ошибки II рода.

В группу параметрических критериев математической статистики помимо методов вычисления описательных статистик должны включаться схемы проверки "на нормальность" распределения и идентификации принадлежности двух выборок одной совокупности. Эти методы основываются на предположении о том, что распределение выборок подчиняется нормальному (гауссовому) закону распределения. Среди параметрических критериев статистики рассматриваются, как правило, критерии Стьюдента и Фишера.

Что касается полученных в ходе исследования данных, то их необходимо представить в наиболее удобном для анализа виде с учетом характера изменения (или вариации) исследуемого признака.

Вариация может быть качественной, когда отличия носят именно качественный характер (например, цвет глаз у мушки *Drosophila*). В этом случае обычно пользуются ранее принятыми обозначениями, например – черная, красная и т.д. окраска. В остальных случаях вариация носит количественный характер и зависит от того, какая степень точности принимается исследователем (например, вес можно измерять килограммах, граммах, миллиграммах и т.д.). В последнем случае чаще всего необходимо проводить предварительную обработку данных (результатов измерений или наблюдений), состоящую в отсеивании грубых погрешностей измерения или погрешностей, неизбежно имеющих место при подготовке числового материала исследования. Два достаточно простых подобных метода приведены в приложении 2 .

При работе с данными часто используется термин

варианта (x_i) – значение или мера признака для той или иной единицы совокупности, i – порядковый номер варианты.

Ряд вариант в совокупности обычно обозначается как $x_1, x_2 x_3, \ldots, x_n$. Перед анализом совокупности бывает полезно сгруппировать отдельные варианты для последующей математической обработки и вычисления статистических показателей, характеризующих изучаемую совокупность.

Группировка по качественной вариации наиболее простая. Например, если мушки отличаются цветом глаз, то их распределение может быть выражено в количестве особей каждой окраски и их процентном количестве.

Группировка данных при количественной дискретной вариации может базироваться на классах, охватывающих все полученные количественные данные от минимальных до максимальных.

Например, если оценивается плодовитость 50 самок рыси, то группировку по классам в этом случае лучше проводить по значениям отдельных вариант: минимально котят в помете – один, максимально – пять. Отсюда 5 классов: с 1 котенком, с 2 и т.д. (табл. 1.1). После распределения всех вариант по классам получаются ряды, в которых показано, как часто встречаются варианты каждого класса и как варьируют признаки от минимального до максимального – это вариационные ряды.

Перечень вариант и соответствующих им частот или частостей называется статистическим распределением выборки или статистическим рядом.

Записывается статистическое распределение в виде таблицы. Первая строка содержит варианты, а вторая – их частоты n_i (или частости p_i^*).

Таблица 1.1 – Распределение самок рыси по количеству котят

количество котят	1	2	3	4	5
частота (у скольких самок встречается)	2	4	16	24	4
частость p_i^*	0,04	0,08	0,32	0,48	0,08

В данном примере количество классов соответствует количеству значений вариант. В случае, когда число значений вариант велико или признак является непрерывным (т.е. когда варианта может принять любое значение в некотором интервале), составляют *интервальный статистический ряд.* В первую строку таблицы статистического распределения вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2)$, $[x_{M-1}, x_M)$ которые берут обычно одинаковыми по длине: $h = x_1 - x_0 = x_2 - x_1 = \ldots$. Для определения величины интервала h можно использовать формулу Sturgess'a:

$$h = \frac{x_{max} - x_{min}}{1 + \log_2 n},$$

где $x_{max} - x_{min}$ – разность между наибольшим и наименьшим значениями признака, количество интервалов $M = 1 + \log_2 n$.

За начало первого интервала рекомендуется брать величину $x_{нач} = x_{min} - h/2$. Во второй строчке статистического ряда подсчитывают количество наблюдений n_i *(i=1,…,M)*, попавших в каждый интервал.

Для наглядности и качественной оценки ряды распределения представляют также и в виде эмпирических кривых. Используются следующие виды эмпирических кривых: полигон, гистограмма и кумулятивные кривые (ступенчатая и ломанная). Полигон и гистограмма соответствуют изображению дифференциальной функции распределения, а кумулятивные кривые – интегральной.

Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны отношению n_i/h – плотность частоты (p_i^*/h или $n_i/(n \cdot h)$ – плотности частости).

Очевидно, что площадь гистограммы частот равна объему выборки, а площадь гистограммы частостей равна единице.

Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2), …, (x_M, n_M) *полигоном частостей* – ломаную с координатами (x_1, p_1^*), (x_2, p_2^*), …, (x_M, p_M^*).

Варианты (x_i) откладываются на оси абсцисс, а частоты и, соответственно, частости – на оси ординат. Как правило, для случайных величин дискретного типа употребляются полигон и ступенчатая кумулятивная кривая, а для непрерывных случайных величин – гистограмма и ломанная кумулятивная кривая.

Одним из способов обработки вариационного ряда является построение эмпирической функции распределения.

Эмпирической (статистической) функцией распределения называется функция $F_n^*(x)$, определяющая для каждого значения x частость события $\{X < x\}$: $F_n^*(x) = p_{X<x}^*$.

Для нахождения значений эмпирической функции удобно $F_n^*(x)$ записать в виде $F_n^*(x) = n_x/n$, где n – объем выборки, n_x – число наблюдений, меньших x $(x \in R)$.

При увеличении числа n наблюдений (опытов) относительная частота события $\{X < x\}$ приближается к вероятности этого события (теорема Бернулли). Эмпирическая функция распределения $F_n^*(x)$ является *оценкой* вероятности события $\{X < x\}$, т.е. кой теоретической функции распределения $F(x)$ случайной величины X.

В приложении 3 содержится пример построения упомянутых выше зависимостей.

Наиболее часто для описания статистической связи признаков используется линейное уравнение регрессии. Внимание к линейной форме связи объясняется четкой интерпретацией параметров линейного уравнения регрессии, ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют (путем логарифмирования или замены переменных) в линейную форму.

Параметры линейной регрессии могут быть найдены разными методами, в приложении 4 представлен пример использования наиболее распространенного метода наименьших квадратов.

При описании данных вариационные ряды могут различаться:

- ✓ По тому значению признака, вокруг которого концентрируется большинство вариант. Это значение признака отражает как бы уровень развития признака в данной совокупности или, иначе, центральную тенденцию вариационного ряда (или наиболее типичное для ряда).
- ✓ По степени вариации вариант вокруг уровня, по степени отклонения от центральной тенденции вариационного ряда.

В соответствии с этим различают две группы основных статистических показателей для характеристики совокупности:

показатели, характеризующие	
центральные тенденции ряда	степень отклонения от центральной тенденции ряда
Мода, Медиана, Среднее арифметическое	Вариационный размах, Среднее квадратичное отклонение, Дисперсия, Коэффициент вариации.

Вычисление перечисленных зависимостей в случае, если выборка задана в виде интервального ранжированного частотного ряда, приведено приложении 7 .

вопросы и задачи для самостоятельной работы

1.2. Критерий согласия χ^2 (критерий Пирсона)

Karl (Charles) Pearson

Критерий χ^2 (читается "хи-квадрат") применяется для сравнения распределений объектов двух совокупностей на основе измерений по шкале наименований в двух *независимых* выборках.

Критерий χ^2 отвечает на вопрос, с одинаковой ли частотой встречаются разные значения признака в эмпирическом и теоретическом распределениях или в двух и более эмпирических распределениях. Это один из наиболее часто применяемых критериев при анализе данных как для количественных дискретных вариаций, так и для непрерывных. Преимущество метода состоит в том, что он позволяет сопоставлять распределения признаков, представленных в любой шкале, начиная от шкалы наименований. В самом простом случае это оценка двух альтернативных (дихотомических) показателей, когда можно обоснованно применять критерий χ^2: "да-нет", "живой-мертвый", "решил задачу – не решил задачу" и т.д.

В случае, если исследователя интересует проблема совпадения (или несовпадения) его данных с результатами других исследователей, то с помощью метода χ^2 можно сопоставить два эмпирических распределения – скажем, ваше и приведенное в других работах.

Аналогично можно сопоставлять распределения выборок из трех и более альтернатив. Например, если в выборке из 50 человек 30 выбрали ответ (а), 16 человек – ответ (б) и четверо – ответ (в), то с помощью метода χ^2 можно проверить, отличается ли это распределение от равномерного распределения или от распределения ответов в другой выборке, где ответ (а) выбрали 12 человек, ответ (б) – 25 человек, ответ (в) – 13 человек.

В тех случаях, если признак измеряется количественно (в килограммах, секундах, миллиметрах и т.д.), необходимо представить всю совокупность значений признака по нескольким классам и затем с помощью метода χ^2 сопоставить частоты встречаемости признака по классам. В остальном принципиальная схема применения метода не меняется.

При сопоставлении эмпирического распределения с теоретическим определяется степень расхождения между эмпирическими и теоретическими частотами. При сопоставлении двух эмпирических распределений определяется степень расхождения между эмпирическими частотами и теоретическими частотами, которые наблюдались бы в случае совпадения двух этих эмпирических распределений.

Чем больше расхождение между двумя сопоставляемыми распределениями, тем больше эмпирическое значение χ^2.

Для нахождения χ^2 обычно используется следующая формула:

$$\chi^2 = \sum_{i=1}^{M} \frac{\left(x_{i\,\text{экс}} - x_{i\,\text{теор}}\right)^2}{x_{i\,\text{теор}}}$$

где $x_{i\,\text{теор}}$ – теоретически ожидаемое число или показатель для данного класса (диапазона, интервала, группы);

$x_{i\,\text{экс}}$ – фактически наблюдаемое; M – количество классов (интервалов).

Ограничения критерия следующие.

1. Объем выборки должен быть достаточно большим: $n > 30$. При $n < 30$ критерий χ^2 дает весьма приближенные значения. Точность критерия повышается при больших значениях n.
2. Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5. Это означает, что если число классов M задано заранее и не может быть изменено, то применять метод χ^2, не накопив определенного минимального числа наблюдений, нельзя. Если, например, проверяются предположения о том, что частота заболеваний гриппом неравномерно распределяются по 7 дням недели, то потребуется исследование 5·7=35 случаев для анализа. Таким образом, если количество классов M задано заранее (M=7), как в данном случае, минимальное число наблюдений (n_{min}) определяется по формуле: $n_{min} = 5 \cdot M = 35$.

3. Выбранные классы должны включать всё распределение, то есть охватывать весь диапазон вариативности признаков. При этом группировка на классы должна быть одинаковой во всех сопоставляемых распределениях.

4. Необходимо (скажем, желательно) вносить так называемую "поправку на непрерывность" Yates'a при сопоставлении распределений признаков, которые принимают всего *два значения* (таблиц сопряженности 2×2). Уточнение обусловлено тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен.

$$\chi^2 = \sum \frac{\left(\left|x_{i\,\text{экс}} - x_{i\,\text{ожид}}\right| - 1/2\right)^2}{x_{i\,\text{ожид}}}.$$

Здесь $x_{i\,экс}$ – фактически наблюдаемое, а $x_{i\,ожид}$ –ожидаемое (расчетное или теоретическое) число или показатель для данной группы.

5. Классы должны быть неперекрещивающимися: если наблюдение отнесено к одному классу, то оно уже не может быть отнесено ни к какому другому классу. И, очевидно, что сумма наблюдений по классам всегда должна быть равна общему количеству наблюдений.

1.2.1. Критерий согласия χ^2 для непрерывной вариации

Во многих практических задачах точный закон распределения неизвестен, поэтому выдвигается гипотеза о соответствии имеющегося в распоряжении исследователя эмпирического закона, построенного по наблюдениям, которая требует статистической проверки.

Пусть X – исследуемая случайная величина. Требуется проверить гипотезу H_0 о том, что данная случайная величина подчиняется закону распределения $F(x)$. Для этого необходимо произвести выборку из n независимых наблюдений и по ней построить эмпирический закон распределения $F_0(x)$.

Сравниваются экспериментальная и теоретическая функции распределений. Общий алгоритм проверки "нулевой" гипотезы следующий.

1. Построить гистограмму равновероятностным способом.

2. По виду гистограммы выдвинуть гипотезу H_0: $f(x)=f_0(x)$, H_1: $f(x)\neq f_0(x)$.

Здесь $f_0(x)$ – плотность вероятности гипотетического закона распределения: нормального, равномерного или какого-либо другого.

3. Вычислить значение критерия по формуле

$$\chi^2 = \sum_{i=1}^{M} \frac{\left(n_i - n_i^{\text{теор}}\right)^2}{n_i^{\text{теор}}},$$

где n_i – число данных в i-том интервале;

$n_i^{\text{теор}}$ – число данных, которое должно содержаться в -том интервале согласно выбранному закону распределения, т.е. при условии, что гипотеза H_0 верна.

Для нормального закона распределения

$$n_i^{\text{теор}} = n\left[\Phi(x_i^{\text{кон}}) - \Phi(x_i^{\text{нач}})\right],$$

$$\Phi(x) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right) dx$$

Здесь $x_i^{\text{нач}}$, $x_i^{\text{кон}}$ – начало и конец i-того интервала.

Замечание. После вычисления всех вероятностей $n_i^{\text{теор}}$ полезно проверить, выполняется ли контрольное соотношение

$$\sum n_i^{\text{теор}} \approx n = \sum n_i$$

4. Из вспомогательных данных (значений функции MS Excel или соответствующих таблиц) выбирается "критическое" значение, $\chi^{2\ \text{крит}}_{\alpha,\ df}$, где α – заданный уровень значимости (α = 0,05 или α = 0,01), а df – число степеней свободы, определяемое по формуле df = M-1-s, s – число параметров, от которых зависит выбранный гипотезой H_0 закон распределения. Значение s для равномерного и нормального закона равно 2.

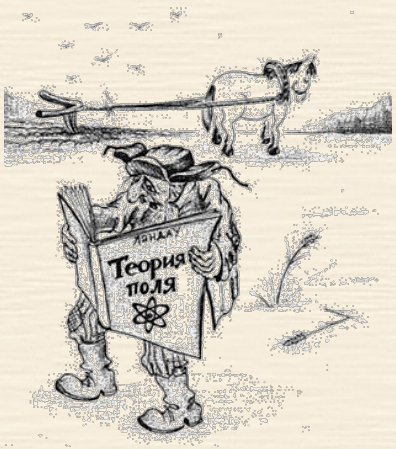
5. Если $\chi^2 > \chi^{2\ \text{крит}}_{\alpha,\ df}$, то гипотеза H_0 отклоняется. В противном случае оснований ее отклонить нет.

Пример 1.1. Проверка соответствия выборки (задана таблицей) нормальному закону распределения при уровне значимости α = 0,05.

9,0	47,5	63,0	17,5	48,0	64,5	21,0	50,0	65,0
26,5	51,0	67,5	27,5	53,5	68,5	31,0	55,0	70,0
32,5	56,0	72,5	34,0	56,0	77,5	36,0	56,5	81,0
36,5	57,5	82,5	39,0	58,0	90,0	40,0	59,0	96,0
41,0	59,0	101,5	42,5	60,0	117,5	43,0	61,0	127,5
45,0	61,5	130,0	46,0	62,0				

Сопоставление пример 1.1 данной выборки и соответствующих значений нормального распределения показывает, что $\chi^2_{крит} \approx 9{,}49$ и $\chi^2_{экс} \approx 11{,}37$ на доверительном уровне $\alpha=0{,}05$.

Поскольку $\chi^2_{экс} > \chi^2_{крит}$, то нулевая гипотеза (соответствия выборки нормальному закону распределения) отклоняется, т.е. выборка не подчиняется нормальному закону распределения.



Учительница: – Вовочка, назови пять африканских животных.
Вовочка: – Запросто! Три слона и две обезьяны!
Анекдот

1.2.2. Критерий согласия χ^2 для дискретной вариации

В данном разделе на примере сравнения экспериментальной и теоретической функций распределения "взгляда" лабораторной крысы рассматривается общий алгоритм проверки "нулевой" для качественной дискретной вариации.

Пример 1.2. (часть 1) Распределение "взгляда" лабораторной крысы*

Чтобы определить цветовые предпочтения у лабораторной крысы, перед ней в клетке были повешены 4 цветных круга (красный, синий, зеленый и желтый). К голове крысы были прикреплены сенсоры фиксации глаз. Исследователем за полчаса эксперимента были получены результаты, представленные в табл. 1.2.

Таблица 1.2 – Распределение "взгляда" лабораторной крысы между 4-мя цветными кругами

Цвет	Красный $n_1^{\text{экс}}$	Синий $n_2^{\text{экс}}$	Зеленый $n_3^{\text{экс}}$	Желтый $n_4^{\text{экс}}$	Всего "взглядов"
Количество "взглядов"	14	5	8	5	32

Полученные эмпирические частоты необходимо сопоставить с теоретическими. Если крыса не отдает предпочтения ни одному цвету, то данное распределение показателя направленности "взгляда" не будет отличаться от равномерного распределения: на всех цветовых кругах задержка взгляда будет примерно одинаковой частоты.

Обсчет данного примера **пример 1.2** показывает, что $\chi^2_{\text{крит}} \approx 7{,}815$ и $\chi^2_{\text{экс}} \approx 6{,}750$ на доверительном уровне $\alpha=0{,}05$.

* Исходные данные примера и ряда других задач для самостоятельной работы взяты из учебно-методического пособия Д. Е. Гавриков. Введение в биологическую статистику / Д. Е. Гавриков. Иркутск: Изд-во Иркутск. гос. пед. ун-та, 2002. –72 с.

Поскольку $\chi^2_{\text{экс}} > \chi^2_{\text{крит}}$, то нулевая гипотеза (соответствия выборки равномерному закону распределения) отклоняется, т.е. выборка не подчиняется равномерному закону распределения.

Вывод: нулевая гипотеза принимается. Распределение "взгляда" лабораторной крысы между цветными кругами не отличается от равномерного распределения.

Пример 1.2. (часть 2) Допустим, что исследование (пример 1.2, часть 1) не остановилось и было получено следующее распределение предпочтений: красный – 15 "взглядов", синий – 6 "взглядов", зеленый – 9 "взглядов", желтый – 6 "взглядов".

Очевидно, что красный цвет остался предпочтительным. У исследователя есть следующий путь, чтобы доказать это статистически: суммировать количества "взглядов" в первом опыте по каждому цвету: количество "взглядов" на красный цвет в опыте 1 + количество "взглядов" на красный цвет в опыте 2, и сопоставить полученное распределение с равномерным. Поскольку количество наблюдений возросло, есть вероятность, что различия окажутся достоверными.

Создается новая таблица эмпирических частот (табл. 1.3), в которой суммированы все "взгляды" лабораторной крысы.

Таблица 1.3 – Распределение взгляда лабораторной крысы между 4-мя цветными кругами

Цвет	Красный	Синий	Зеленый	Желтый	Всего "взглядов"
Количество "взглядов"	29	11	17	11	68

Формулируется нулевая гипотеза: распределение проявлений цветовых предпочтений крысы не отличается от равномерного распределения. Все расчеты производятся аналогично предыдущему случаю (пример 1.2, часть 1).

Таким образом, имеется:

$n_{\text{теор}}=68/4=17$

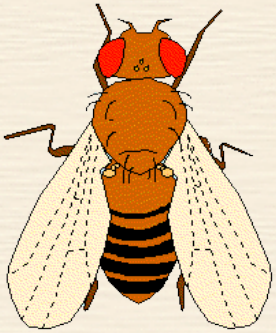
$df=3$

$$\chi^2_{\text{кр}} = \begin{cases} 7{,}815\ (P \le 0{,}05) \\ 11{,}345\ (P \le 0{,}01) \end{cases} \qquad \chi^2_{\text{эмп}}=12{,}706$$

$$\chi^2_{\text{эмп}} > \chi^2_{\text{кр}}$$

Вывод: нулевая гипотеза отклоняется. Распределение "взгляда" лабораторной крысы между цветными кругами отличается от равномерного распределения ($\alpha < 0{,}01$).

На этом примере можно убедиться, что увеличение числа наблюдений повышает достоверность результата, если, конечно, в новых наблюдениях воспроизводится прежняя тенденция различий.



Пример 1.3. Моногибридное скрещивание. Анализ гибридов второго поколения. Из эксперимента получено: количество мух (e^+-) равно 100, (ее)=47. Доказать, что отношение количества мух – особей с рецессивным и с доминантным признаками, равно 3:1.

Алгоритм анализа пример 1.3 заключается в построении ряда теоретических частот $n_{i\,\text{теор}}$ для данного количества экспериментальных значений, вычисление $\chi^2_{\text{эмп}}$ и $\chi^2_{\text{кр}}$ с их последующим сравнением.

Поскольку $\chi^2_{\text{эмп}} = 3{,}812 < \chi^2_{\text{кр}} = 3{,}841$, то формулируется вывод: нулевая гипотеза принимается по уровню значимости α=0,05. Отношение количества особей рассмотренных фенотипических классов равно 3:1.

Пример 1.4. Дигибридное скрещивание. Анализ гибридов второго поколения.

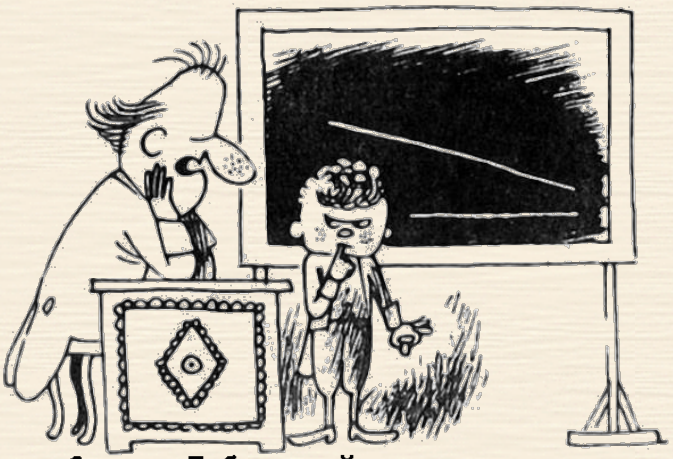
Доказать, что у мух, у которых учитываются две пары альтернативных признаков, в гибридах второго поколения наблюдается расщепление по фенотипическим классам 9 : 3 : 3 : 1. Отношение, полученное экспериментально, следующее:

135 : 51 : 54 : 18.

Алгоритм анализа пример 1.4 заключается в построении ряда теоретических частот $n_{i\,\text{теор}}$ для данного количества экспериментальных значений, вычисление $\chi^2_{\text{эмп}}$ и $\chi^2_{\text{кр}}$ с их последующим сравнением.

Поскольку $\chi^2_{\text{эмп}} = 1{,}721 < \chi^2_{\text{кр}} = 7{,}815$, то формулируется вывод: нулевая гипотеза принимается по уровню значимости $\alpha = 0{,}05$. Отношение количества особей рассмотренных фенотипических классов равно

135 : 51 : 54 : 18.



Студенты, помните: все, сказанное вами на экзамене,
может быть использовано против вас!
Анекдот

1.2.3. Критерий χ^2 соответствия (сравнение с "прошлыми" событиями)

Аналогия ("сейчас и раньше") используется в тех случаях, когда нужно определить, является ли набор частот *нынешнего эксперимента типичным* по отношению к *предыдущему эксперименту* (опыту), набор частот которого используется в качестве *опорных ("теоретических") величин*.

Тест χ^2 соответствия используется для проверки гипотезы о том, что комбинация наблюдаемых частот, характеризующих одну качественную переменную, строится на данных из некоторой генеральной совокупности уже известных значений (опорных величин). Или: "Наблюдаемые сейчас фактические данные аналогичны прошлым данным"? При этом подразумевается, что и те и другие относятся к одной и той же генеральной совокупности, но определялись в разное время.

Ожидаемое значение частоты x_i^T для каждой категории рассчитывается как произведение его доли (определенной по "старой" выборке x_i^C), на размер "новой" выборки x_i^n.

$$x_i^T = \frac{x_i^C}{\sum_{j=1}^{M} x_i^C} \sum_{j=1}^{M} x_i^n$$

где i=1,2,...,M,

M – количество категорий.

На основании имеющихся знаний о наблюдаемой и ожидаемой ("теоретической") частотах анализируемого события обычным образом определяется собственно показатель χ^2, который затем сравнивается с критическим. Соответствующее число степеней свободы определяется как количество категорий (классов) минус единица $df = M\text{-}1$.

Если оказывается справедливым неравенство $\chi^2_{\text{эмп}} > \chi^2_{\text{кр}}$, то с заданной вероятностью (или уровнем значимости) можно утверждать, что наблюдаемые частоты (последний опыт) значимо отличаются от тех, которые ожидаются исходя из известных нам опорных значений частот. Следовательно, обоснованно можно делать вывод о том, что наблюдаемые выборочные проценты значимо отличаются от заданных опорных значений.

Если имеем соотношение, $\chi^2_{\text{эмп}} < \chi^2_{\text{кр}}$, то наблюдаемые значения незначительно отличаются от опорных показателей и, следовательно, фактические результаты не имеют значимых отличий от опорных значений.

Анализ критерия соответствия процентов (частот) удобно выполнять, придерживаясь следующей схемы.

1. Выдвигаются следующие гипотезы:

а) нулевая: частоты "нынешнего" опыта равны набору известных опорных величин "прошлого" опыта;

б) альтернативная: частоты нынешнего опыта не равны набору данных прошлого опыта.

2. Вычисляются ожидаемые частоты. При этом предполагается, что набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности и ожидается наличие не менее пяти объектов в любой категории.

Пример 1.5. Для группы студентов проводилось тестирование остаточных знаний проведением контрольных работ с разрывом в месяц с оцениванием по пятибалльной системе. Результаты контрольных сведены в таблице справа; количество участников тестирования было различное.

Необходимо ответить на вопрос – изменились или нет знания студентов (по частотам отметок).

Результаты контрольных работ

отметка	контрольная работа	
	№1	№2
2	8	10
3	32	27
4	11	7
5	5	7

Если составить таблицу сопоставления "нынешнего" опыта (контрольной №2) с ожидаемыми (табл. 1.4), то из нее видно, что по количественному показателю все отметки отличаются от опорных значений, причем это различие далеко неравноценно; возникает вопрос: значима ли эта разница?

Таблица 1.4 – Имеющийся и ожидаемый результаты

отметка	контрольная работа №2	ожидаемый результат
2	10	7,29
3	27	29,14
4	7	10,02
5	7	4,55

Ответ на поставленный вопрос дает сравнение расчетного и критического значений χ^2. Поскольку пример 1.5 в данном случае $\chi^2_{эмп} < \chi^2_{кр}$ (3,39<7,81), то можно сделать вывод: разница в распределении отметок незначительна, статистически они одинаковы.

Отметим следующее. По значению $\chi^2 \approx 3{,}39$ можно оценить вероятность ошибки сделанного заключения. В ячейку B16 введем формулу вычисления этой вероятности =ХИ2РАСП(E9;3) {=CHIDIST(E9;3)}. Полученное число показывает: гипотеза о том, что результаты контрольной работы №2 отличаются от таковых контрольной №1, высказывается с "риском" допустить ошибку 34%. И, напротив, с вероятностью 66% можно говорить о том, что различие между этими отметками несущественное.

Компьютер – это устройство для упорядочивания, ускорения и автоматизации человеческих ошибок
Неизвестный исследователь

1.2.4. Проверка наличия взаимосвязи между двумя качественными переменными (критерий χ^2 независимости)

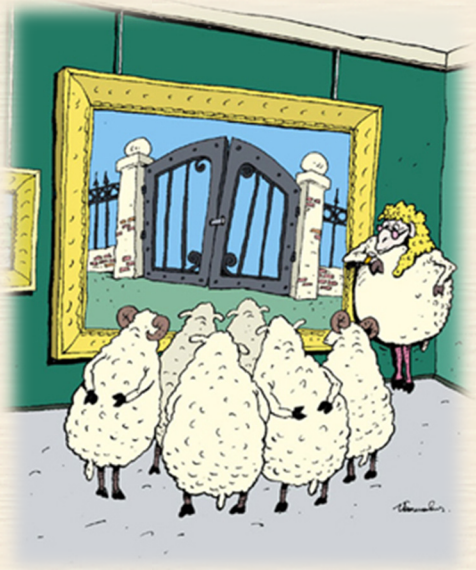
Возможны ситуации, когда имеются две качественные переменные, т.е. набор экспериментальных данных представляет собой *двумерные качественные* данные. После изучения каждой из них *отдельно* с помощью анализа частот (или процентов) может возникнуть вопрос о наличии *связи* между ними.

Считается, что две качественные переменные являются *независимыми*, если знание значения одной переменной не помогает предсказать значение другой.

Сказанное можно пояснить следующим примером. Пусть надежность работы компьютеров в классе составляет в среднем 93%. Однако когда в классе находится г-жа $\mathcal{Z}$, надежность функционирования снижается до 71%. В этом случае знание значения одной переменной (имя конкретного человека) позволяет прогнозировать значение другой переменной (надежности работы). Сбой работы компьютеров более вероятен во время присутствия этой г-жи $\mathcal{Z}$ и менее вероятен, когда работает кто-то другой. Следовательно, эти две переменные *не являются независимыми*.

Использование критерия χ^2 позволяет решить вопрос о том, являются ли рассматриваемые качественные совокупности зависимыми или же независимыми друг от друга.

В этом случае применяется так называемый *критерий χ^2 независимости,* который устанавливает наличие (или отсутствие) связи между двумя качественными переменными. Для такого анализа используется таблица частот, которые можно было бы ожидать в том случае, *если переменные оказались бы независимыми*. В общем случае критерий χ^2 независимости принято применять следующим образом.



1. Составляется таблица исходные данные в виде списка частот Z_{ij} всех комбинаций категорий двух качественных переменных (см. табл. 1.5). Выдвигаются гипотезы:

а) две переменные статистически неразличимы (нулевая гипотеза);	*а*) две переменные не зависят одна от другой (нулевая гипотеза);
б) две переменные статистически различны;	*б*) две переменные связаны, взаимозависимы

Таблица 1.5 – Переменные (выборки) и их категории (экспериментальные частоты)

переменные i=1,2	категория j=1,2,…,M			
	признак 1	признак 2	* * *	признак M
переменная (выборка) 1	Z_{11}	Z_{12}	* * *	Z_{1M}
переменная выборка 2	Z_{21}	Z_{22}	* * *	Z_{2M}

2. Составляется таблица (см. табл. 1.6) ожидаемых ("теоретических") частот.

Таблица 1.6 – Переменные (выборки) и их категории ("теоретические" частоты)

переменные i=1,2	категория j=1,2,…,M			
	признак 1	признак 2	* * *	признак M
переменная (выборка) 1	T_{11}	T_{12}	* * *	T_{1M}
переменная выборка 2	T_{21}	T_{22}	* * *	T_{2M}

где $T_{ij} = \dfrac{(Z_{1j} + Z_{2j})}{\sum_{p=1}^{M}(Z_{1p} + Z_{2p})} \sum_{p=1}^{M} Z_{ip}$.

Последнюю формулу можно (для вычислений иногда даже удобнее) представить в виде

$$T_{ij} = \frac{\sum_{q=1}^{2} Z_{qj} \times \sum_{p=1}^{M} Z_{ip}}{\sum_{q=1}^{2} \sum_{p=1}^{M} Z_{qp}},$$

или, если выражаться словами:

$$T_{ij} = \frac{(\text{сумма значений } i - \text{той строки}) \times (\text{сумма значений } j - \text{того столбца})}{(\text{сумма всех значений})}.$$

3. Расчетное значение χ^2 определяется по формуле

$$\chi^2_{\text{расч}} = \sum_i \sum_j \frac{(Z_{ij} - T_{ij})^2}{T_{ij}}.$$

Это же значение можно вычислить "сразу", без подсчета ожидаемых частот, по формуле

$$\chi^2_{\text{расч}} = \frac{1}{n_1 n_2} \sum_{j=1}^{M} \frac{(n_1 Z_{2j} - n_2 Z_{1j})^2}{Z_{1j} + Z_{2j}}, \qquad n_1 = \sum_j Z_{1j}, \quad n_2 = \sum_j Z_{2j}.$$

4. Критическое значение $\chi^2_{\text{кр}}$ (M категорий и две переменных) определяется для df = (количество категорий – 1) (количество переменных – 1) = (M-1)(2-1) = M-1 степеней свободы.

Далее проводится сравнение расчетного значения $\chi^2_{\text{расч}}$ с критическим $\chi^2_{\text{кр}}$ (по "обычным" уровням значимости 0,05 или 0,01).

При $\chi^2_{\text{расч}} > \chi^2_{\text{кр}}$ гипотеза об отсутствии связи между признаками и параметрами отвергается, при $\chi^2_{\text{расч}} < \chi^2_{\text{кр}}$ – подтверждается.

Пример 1.6. В Тегульдетском районе Томской области в 1999 и 2002 годах В. П. Перевозкин выполнил сбор материала и получил распределение сочетаний хромосомных инверсий комаров *Anopheles messeae* (данные в таблице ниже).

Распределение кариотипов *Anopheles* в различные годы

год выборки	кариотипы				
	XL11 2R11	XL22 2R11	XL11 2R01	XL00 2R00	XL11 2R00
1999	38	26	16	62	69
2002	37	18	14	78	93

Вопрос: изменилось или нет распределение по сочетанию инверсий *Anopheles messeae* за три года? Иными словами – выборки 1999 и 2002 года статистически одинаковы или различны?

Сравниваются пример 1.6 расчетное (эмпирическое) и критическое значений χ^2. Поскольку в данном случае $\chi^2_{\text{расч}} < \chi^2_{\text{кр}}$ (5,14<9,49), то можно сделать вывод: гипотеза о неизменности распределения *Anopheles* подтверждается.

Пример 1.7. На сайте В. Леонова* приводится интерактивный пример анализа сопряженности двух качественных признаков 2×2. А именно, рассматривается вопрос: "Можно ли утверждать, исходя из данных конкретной выборки, что два исследуемых дискретных качественных признака независимы друг от друга в генеральной совокупности?" Иными словами, определяется то, что *между этими признаками отсутствует взаимосвязь*. Если эта гипотеза будет отвергнута, то с высокой долей вероятности можно утверждать, что такая зависимость существует. Реальным содержанием примера является исследование наличия взаимосвязи между приемом контрацептивных таблеток матерями, и желтухой у детей получающих грудное вскармливание:

Данные исследования

Прием матерью таблеток	**b1**. Есть желтуха	**b2**. Нет желтухи	Всего
a1. Принимала таблетки	33	24	57
a2. Не принимала таблетки	14	45	59
Всего	47	69	116

В этом примере у 33 матерей принимавших таблетки дети болели желтухой, а у 24 матерей также принимавших таблетки дети не болели желтухой. Далее, у 14 матерей, которые не принимали таблетки, дети болели желтухой и у 45 матерей, не принимавших таблетки, дети не болели желтухой.

* http://www.biometrica.tomsk.ru/freq1.htm

Из сравнения пример 1.7 $\chi^2_{расч}$ с $\chi^2_{кр}$ для α = 0,005 видно, что вычисленное χ^2 превосходит критическое. Другими словами – выборки статистически различны и поэтому гипотеза о независимости между заболеванием желтухой и приемом контрацептивных таблеток отвергается при уровне значимости α < 0,005, т.е. *какая-то зависимость* между заболеванием и приемом таблеток *существует*.

В примере отмечается, что полученному значению $\chi^2_{расч}$=14,04 при числе степеней свободы равном df = 1 отвечает достигнутый уровень значимости α = 0,000179.

Пример проверки гипотезы о независимости двух выборок *X* и *Y* с произвольным числом классов приведен в приложении 5

вопросы и задачи для самостоятельной работы

Если вам непонятно какое-то слово в техническом тексте,
не обращайте на него внимания.
Текст полностью сохраняет смысл и без него.
Мерфология, Закон Купера

Sir Ronald
Aylmer Fischer

1.3. *F* – критерий Фишера

Критерий Фишера позволяет сравнивать величины выборочных дисперсий двух независимых выборок. Для вычисления $F_{эмп}$ нужно найти отношение дисперсий двух выборок X_1 и X_2 такое, когда большая по величине дисперсия находится в числителе, а меньшая – в знаменателе. Формула вычисления критерия Фишера следующая:

$$F_{эмп} = \frac{\sigma^2_{X_1}}{\sigma^2_{X_2}} \geq 1 \quad \left(\sigma^2_{X_1} \geq \sigma^2_{X_2}\right).$$

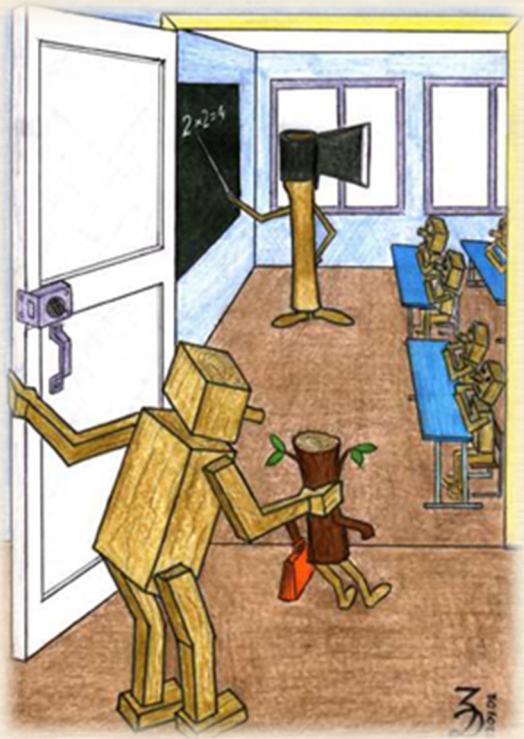
Здесь $\sigma^2_{X_1}$, $\sigma^2_{X_2}$ – дисперсии первой и второй выборки соответственно.

Если $F_{эмп} > F_{кр}$, то дисперсии считаются различными.

Пример 1.8. Для двух групп особей проводилось тестирование индекса массы. Полученные значения величин средних достоверно не различались, однако исследователя интересует вопрос – есть ли различия в степени дисперсной однородности показателей индекса между группами.

В результате анализа данных пример 1.8 установлено, что $(F_{кр}) = 2{,}896 < 3.963 = (F_{эмп})$, поэтому в терминах статистических гипотез можно утверждать, что H_0 (гипотеза о сходстве) может быть отвергнута на уровне (ошибки) 5%, и есть основание в этом случае принять гипотезу H_1.

Можно утверждать, что по степени однородности показателя (индекса массы) между выборками рассматриваемых двух групп имеется различие.



вопросы и задачи для самостоятельной работы

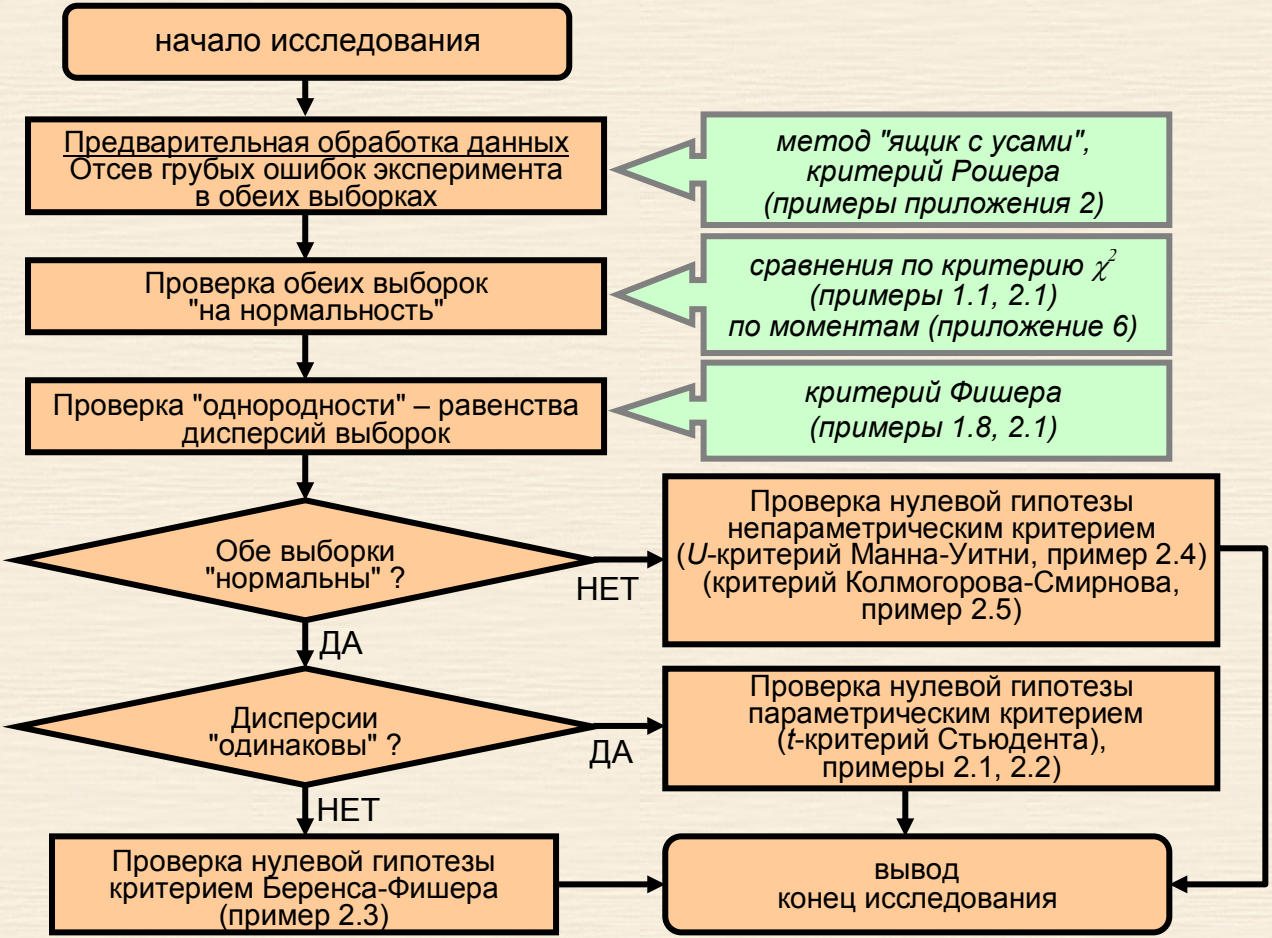
Если мой сосед бьет жену каждый день, а я никогда,
то с точки зрения статистики мы оба бьем своих жен через день.
Бернард Шоу

2. Сравнительный анализ двух непрерывных выборок

Одной из важнейших задач статистических исследований является совместный анализ нескольких выборок, а важнейшим вопросом, возникающем при анализе двух выборок, является вопрос о наличии различий между этими выборками. Обычно для этого проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве средних. Диаграмма алгоритма сравнения средних значений двух выборок представлена на следующей странице.

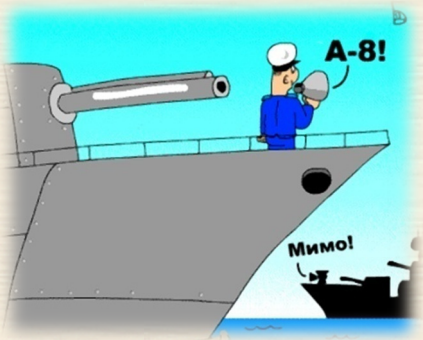
Если вид распределения или функция распределения выборки известны или заданы, то в этом случае задача оценки различий двух групп независимых наблюдений может решаться с использованием *параметрических критериев* статистики – например критерия Стьюдента *t*, если сравнение выборок ведется по средним значениям, или критерия Фишера *F*, если сравнение выборок ведется по их дисперсиям.

Использование параметрических критериев статистики без предварительной проверки вида распределения может привести к ошибкам проверки рабочей гипотезы.



Алгоритм сравнения средних значений двух выборок

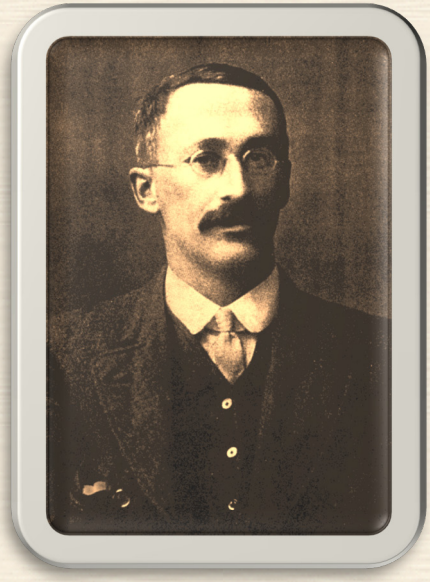
Если вид распределения или функция распределения выборки неизвестен, то используются *непараметрические критерии статистики,* такие, как критерий знаков, двухвыборочный критерий Вилкоксона, критерий Манна-Уитни, Спирмена, Смирнова, применение которых хотя и не требует большого числа членов совокупности и знания вида распределения, но все же зависит от некоторых условий. Еще раз: непараметрические критерии статистики свободны от допущения о законе распределения выборок и базируются на предположении о независимости наблюдений.



Нельзя заранее правильно определить,
какую сторону бутерброда мазать маслом.
Мерфология, Закон своенравия природы

2.1 Параметрический *t*-критерий Стьюдента

Традиционный метод проверки однородности (критерий Стьюдента) позволяет найти вероятность того, что оба средних значения в выборках относятся к одной и той же совокупности.



William Sealy Gosset

Классические условия применимости критерия Стьюдента. Согласно математической теории статистики должны быть выполнены два условия применимости критерия Стьюдента, основанного на использовании статистики *t:*

а. результаты наблюдений имеют нормальные распределения с математическими ожиданиями μ_1 и μ_2 и дисперсиями σ_1^2 и σ_2^2 в первой и во второй выборках соответственно;

б. дисперсии результатов наблюдений в первой и второй выборках совпадают $\boldsymbol{\sigma_1^2 = \sigma_2^2}$.

Если условия а) и б) выполнены, то нормальные распределения выборок отличаются только математическими ожиданиями и гипотеза H_0 сводятся к гипотезе $H_0 : \mu_1 = \mu_2$, а альтернативная к $H_1 : \mu_1 \neq \mu_2$.

Если хотя бы одно из условий а) и б) не выполнено, то нет оснований считать, что статистика t имеет распределение Стьюдента, поэтому применение традиционного метода, строго говоря, не обосновано. Специальным случаем является выполнение только условия а) – нормальности распределения, когда можно использовать "модификацию" распределения Стьюдента – критерий Беренса-Фишера.

При использовании критерия Стьюдента можно выделить два случая. Первый, когда его применяют для проверки гипотезы о равенстве генеральных средних двух *независимых, несвязанных* выборок (так называемый *двухвыборочный t-критерий*). В этом случае обычно анализируются контрольная и экспериментальная (опытная) группы (выборки) разных объемов.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотезы равенства средних, используется так называемый *парный t-критерий*. Выборки при этом называют *зависимыми, связанными.*

Лекарство – это вещество, которое, будучи введено в крысу, дает научный результат или статью.
Мерфология, правило Матца относительно лекарств

2.1.1 Критерий Стьюдента для независимых выборок равных дисперсий (*гомоскедастический тест)*

Алгоритм применения критерия Стьюдента для независимых выборок x_i и y_i равных дисперсий (*гомоскедастический тест)* базируется на вычислениях следующих параметров

параметр	*выборка* $\boldsymbol{x_i}$ *(объемом* $\boldsymbol{n_x}$*)*	*выборка* $\boldsymbol{y_i}$ *(объемом* $\boldsymbol{n_y}$*)*
выборочное средне-арифметическое	$\overline{x} = \frac{1}{n_x}\sum_i x_i$	$\overline{y} = \frac{1}{n_y}\sum_i y_i$
выборочная дисперсия	$\sigma_x^2 = \frac{1}{n_x - 1}\sum_i (x_i - \overline{x})^2$	$\sigma_y^2 = \frac{1}{n_y - 1}\sum_i (y_i - \overline{y})^2$
статистика Стьюдента t, на основе которой и принимается решение по гипотезе H_0:	$t = \frac{\overline{x} - \overline{y}}{\sqrt{(n_x - 1)\sigma_x^2 + (n_y - 1)\sigma_y^2}} \sqrt{\frac{n_x n_y}{n_x + n_y} df}$	

По заданному уровню значимости α и числу степеней свободы $df = (n_x + n_y - 1)$ из распределения Стьюдента находят критическое значение $t_{\text{крит}}$.

Если $|t| > t_{крит}$, то гипотезу однородности средних (отсутствия статистически значимого различия) отклоняют, если же $|t| < t_{крит}$, то принимают.

При односторонних альтернативных гипотезах вместо $|t| > t_{крит}$ проверяют условие $t > t_{крит}$; данная постановка задачи не имеет принципиальных отличий от описанной выше.

Ниже приводится пример использования *t*-критерия Стьюдента для анализа несвязных и неравных по численности выборок.

Пример 2.1. Сравнивается – одинаков ли средний размер листьев у двух ярусов кроны акации. Данные эксперимента приведены в таблице.

Результаты эксперимента

ярус 1 выборка x_i

11,4	11,9	11,5	11,6	12,0	11,5	11,1
11,3	12,4	12,1	12,6	12,1	12,5	12,2
14,1	14,8	8,2	10,1	10,7	10,4	10,7
13,7	13,9	13,2	13,8	16,3	16,2	14,7
14,3	14,8	14,8	15,2	15,6	15,5	14,7

ярус 2 выборка y_i

14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4
11,8	11,4	16,3	16,1	11,4	11,9	15,8	12,1
12,5	12,2	17,0	16,6	12,3	17,3	13,2	13,9
13,0	14,4	14,1	13,9	13,8	13,5	15,6	15,5
15,3	15,1	15,1	15,0	15,1	15,8		

Алгоритм решения предусматривает следующие этапы (см. представленную выше схему исследования . Для сокращения объема примера будем считать, что для выборок уже предварительно проведен отсев грубых ошибок эксперимента каким-либо методом (см. Приложение 2).

Общая структура данного исследования предполагает

- ✓ расчет основных характеристик (объем, среднее, дисперсия, отклонение) выборок x_i и y_i;
- ✓ проверку применимости критерия Стьюдента, а именно
  - a) проверку равенства (однородности) дисперсий выборок;
  - b) проверку нормальности распределений для обеих выборок;

✓ при выполнении условия b) выполняется анализ нулевой гипотезы о равенстве выборок по критерию Стьдента. Если дисперсии статистически одинаковы (выполнилось условие a)), то используется форма критерия для равных дисперсий (гомоскедастический тест)

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(n_x - 1)\sigma_x^2 + (n_y - 1)\sigma_y^2}} \sqrt{\frac{n_x n_y}{n_x + n_y} df}.$$

Проверка нормальности распределений проводится сравнением по критерию $\chi^2_{\text{экс}} = \sum_{i=1}^{M} \frac{(n_i - n_i^{\text{теор}})^2}{n_i^{\text{теор}}}$ частот экспериментальных данных n_i и теоретических $n_i^{\text{теор}}$ (нормального закона распределения) для выборки такого же объема.

Для нормального закона распределения

$$n_i^{\text{теор}} = n\left[\Phi(x_i^{\text{кон}}) - \Phi(x_i^{\text{нач}})\right], \quad \Phi(x) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x - \overline{x})^2}{2\sigma^2}\right) dx$$

где $x_i^{\text{нач}}$, $x_i^{\text{кон}}$ – начало и конец i-того частотного интервала.

В результате анализа пример 2.1 было определено, что дисперсия обеих выборок уровне 5% одинакова и обе выборки подчиняются нормальному закону распределения. Критерий Стьюдента (гомоскедастический тест) показал, что различия между выборками являются статистически значимыми.

Джон Лайтерут, он же архиепископ Ушер Ирландский,
подсчитал со всем усердием в Кембридже, в 1654 году,
что создатель сотворил человека из глины точно
в 9 часов утра 23 октября 4004г. до рождения Христа.
Справка

2.1.2 Парный критерий Стьюдента

В случае связанных выборок (с равным числом n измерений в каждой) можно использовать следующую формулу вычисления t-критерия Стьюдента:

$$t = |\overline{d}| \sqrt{\frac{n(n-1)}{\sum_i d_i^2 - n\overline{d}^2}}$$

где $d_i = x_i - y_i$ разности между соответствующими значениями пар переменных, а $\overline{d}$ – среднее этих разностей.

Число степеней свободы df определяется по формуле df=n-1 (в данном случае n есть число *пар* данных).

Если $t < t_{крит}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная.

Пример 2.2. Изучался уровень воздействия препарата на группу 10 особей. Вопрос: какова эффективность препарата? С целью проверки эффективности до начала эксперимента и после проводился идентифицирующий тест. Данные эксперимента приведены в таблице.

Результаты эксперимента

выборка X	14	20	15	11	16	13	16	19	15	9
выборка Y	18	19	22	17	24	21	25	26	24	15

Выполненная в приложении 6 проверка данных показала, что обе выборки соответствуют нормальному распределению и имеют статистически одинаковую дисперсию по уровню значимости α=0,05. В связи с этим (выполнены условия применимости *t*-критерия) и с учетом того, что выборки относятся к одним и тем же объектам, эффективность применения препарата можно анализировать парным критерием Стьюдента

$$t = |\overline{d}| \sqrt{\frac{n(n-1)}{\sum_i d_i^2 - n\overline{d}^2}},$$

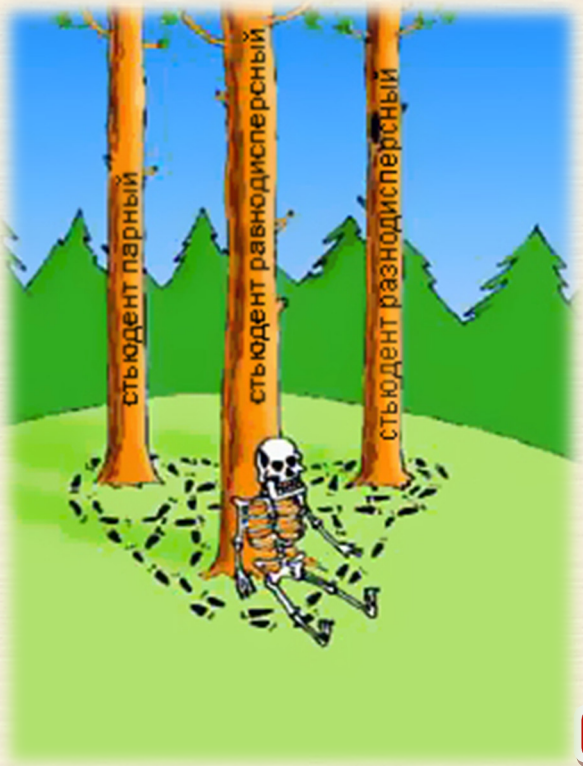
где $d_i = x_i - y_i$ разности между соответствующими значениями пар переменных, а $\overline{d}$ – среднее этих разностей.

Из сравнения (пример 2.2) экспериментального и критического значений *t*-критерия следует возможность принятия альтернативной гипотезы о достоверных различиях средних арифметических, т.е. делается вывод об эффективности воздействия препарата.

Терпеть не могу логики. Она всегда банальна и нередко убедительна.
Оскар Уайльд

2.1.3 Критерий Стьюдента для независимых выборок неравных дисперсий (*гетероскедастический тест)*

Проверка гипотезы о генеральных средних двух групп *с нормальным распределением и неравными дисперсиями* в математической статистике называется проблемой Беренса-Фишера и имеет в настоящее время только приближенные решения. Можно показать, что чем больше различаются между собой дисперсии и объемы выборок, тем сильнее отличается расчетное значение *t*-критерия от классического "стьюдентовского". При этом различную величину имеет и сам *t*-критерий, и число степеней свободы. И, тем не менее, когда дисперсии неизвестны и



их равенство не предполагается ($\sigma_x^2 \neq \sigma_y^2$), можно использовать так называемую статистику критерия Беренса-Фишера

$$t = \frac{|\overline{x} - \overline{y}|}{\sqrt{\Omega_x + \Omega_y}}, \text{где } \Omega = \frac{\sigma^2}{n}.$$

Известно, однако, что это распределение близко к распределению Стьюдента с числом степеней свободы, равным

$$df = \frac{\left(\Omega_x + \Omega_y\right)^2}{\frac{\Omega_x^2}{n_x - 1} + \frac{\Omega_y^2}{n_y - 1}}.$$

Строго говоря, описанные выше критерии применимы только к выборкам, извлеченным из *нормальной* генеральной совокупности. Вместе с тем специальные исследования показали, что *t*-критерий является (особенно при больших объемах выборок) достаточно устойчивым по отношению к отклонениям исследуемых генеральных совокупностей от нормальных. А это значит, что он может применяться и к выборкам из негауссовских генеральных совокупностей с той лишь оговоркой, что истинные значения уровня значимости и мощности критерия в этом случае будут незначительно отличаться от "стандартных".

Пример 2.3. Сравнивается (одинаков или нет?) средний размер листьев у двух ярусов кроны акации. Данные эксперимента приведены в таблице.

Результаты эксперимента

ЯРУС 1 (выборка X)							
11,4	11,9	11,5	11,6	12,0	11,5	11,1	11,3
12,4	12,1	12,6	12,1	12,5	12,2	14,1	14,8
8,2	10,1	10,7	10,4				

ЯРУС 2 (выборка Y)							
14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4
10,8	11,4	16,3	16,1	11,4	11,9	15,8	12,1
12,5	12,2	17,0	16,6	12,3	17,3	13,2	13,9
13,0	14,4	14,1	13,9	12,0	13,5	12,0	15,5
10,0	10,0	10,0	10,0	10,0	10,0		

Алгоритм решения предусматривает следующие этапы (см. представленную в разд. 2 схему исследования):

- ✓ расчет основных характеристик (объем, среднее, дисперсия, отклонение) выборок X$\{x_i\}$ и Y$\{y_i\}$;

- ✓ проверку применимости критерия Стьюдента, а именно
  - a) проверку равенства (однородности) дисперсий выборок;
  - b) проверку нормальности распределений для обеих выборок;
- ✓ при выполнении условия b) выполняется анализ нулевой гипотезы о равенстве выборок по критерию Стьюдента. Если дисперсии статистически различны (не выполнилось условие a)), то используется форма критерия для неравных дисперсий (гетероскедастический тест)

$$t = \frac{|\overline{x} - \overline{y}|}{\sqrt{\Omega_x + \Omega_y}}, \text{где } \Omega = \frac{\sigma^2}{n}.$$

Проверка нормальности распределений проводится сравнением по критерию $\chi^2_{\text{экс}} = \sum_{i=1}^{M} \frac{(n_i - n_i^{\text{теор}})^2}{n_i^{\text{теор}}}$ частот экспериментальных данных n_i и теоретических $n_i^{\text{теор}}$ (нормального закона распределения) для выборки такого же объема.

В результате анализа пример 2.3 было определено, что дисперсия обеих выборок уровне 5% неодинакова и обе выборки подчиняются нормальному закону распределения. Критерий Стьюдента (гетероскедастический тест) показал, что различия между выборками являются статистически значимыми.

вопросы и задачи для самостоятельной работы

Люди делятся на три категории:
умеющие считать и не умеющие считать.
Мерфология, Закон Уинкорна

2.2. Непараметрические критерии

2.2.1 Непараметрический *U*-критерий Манна-Уитни

Критерий предназначен для оценки различий между двумя выборками (скажем *X* и *Y*) по уровню какого-либо количественно измеренного признака, и что более важно, критерий Манна-Уитни позволяет оценивать выборки при неизвестном законе распределении признака (общепринятая интерпретация – проверка равенства медиан). Тест *U* позволяет проверить, существует ли достоверная разница между двумя независимыми выборками после того, как сгруппированные данные этих выборок ранжированы и вычислены суммы рангов для каждой выборки.

Для корректной работы теста требуется выполнение следующих условий:

- распределения *X* и *Y* непрерывны (или являются дискретными распределениями, хорошо аппроксимирующими непрерывное распределение);
- распределения *X* и *Y* имеют одинаковую форму, единственным возможным отличием является их расположение (т.е. медиана);

- число элементов в каждой выборке не менее 5 ($n_x \geq 3, n_y \geq 5$);
- выборки независимы;
- шкала измерений должна быть порядковой, интервальной или относительной (т.е. тест не применяется к номинальным переменным).

Данным методом определяется насколько перекрещиваются (совпадают) значения между двумя выборками (1-я выборка – ряд, в котором, по предварительной оценке значения выше, а 2-я выборка – ряд, где они предположительно ниже).

Чем меньше перекрещивающихся значений (чем меньше U), тем более вероятно, что различия достоверны: т.е.

если $U < U_{кр}$, то нулевая гипотеза отвергается.

В упрощенном изложении U-тест выглядит следующим образом. Обе выборки объединяются в один массив, с сохранением информации о принадлежности каждого элемента данных конкретной выборке. В новом созданном массиве элементы заменяются их рангами (порядковыми номерами) по правилам ранжирования. Статистика критерия определяется по формуле:

$$U = (n_x \cdot n_y) + \frac{n_*(n_* + 1)}{2} - T_*,$$

где n_x, n_y – количество вариант в выборках X и Y;

T_* – большая из двух ранговых сумм;

n_* – количество вариант в группе с большей суммой рангов.

Распределение U очень достаточно сложно аппроксимировать – оно дискретное, а его интегральная функция не разлагается ни в ряд, ни в цепные дроби. В стандартных электронных таблицах функция, возвращающая параметры U-распределения, отсутствует, поэтому приходится пользоваться таблицей таблица $U_{кр}$

Пример 2.4. Дана таблица экспериментальных данных. Проверить нулевую гипотезу о равенстве признака в этих выборках на уровне значимости $\alpha = 0{,}05$.

X	14	20	15	11	16	13	16	19	15	9		
Y	18	19	22	17	24	21	25	26	24	15	18	22

Для анализа используется критерий Манна-Уитни, позволяющий оценить выборки при неизвестном законе распределении признака (общепринятая интерпретация – проверка равенства медиан). Тест *U* позволяет проверить, существует ли достоверная разница между двумя независимыми выборками *после того*, как сгруппированные данные этих выборок ранжированы и вычислены суммы рангов для каждой выборки.

При построении решения (пример 2.4) алгоритм предусматривает построение массивов ранговых значений объединенных данных и вычисление сумм рангов каждой выборки.

Из сравнения величин критериев $(U < U_{кр})$ следует возможность принятия альтернативной гипотезы о достоверных различиях средних арифметических.

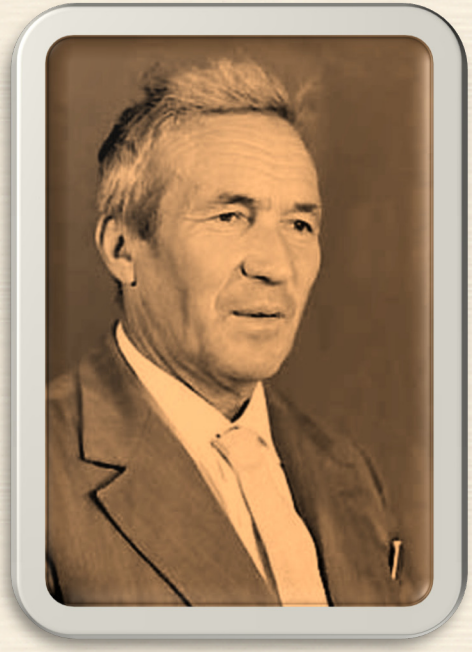
Статистический анализ показывает, что вероятность крупного выигрыша в лотерею всегда одинакова и не зависит от того, купили вы лотерейный билет или нет.
Народная мудрость

2.2.2 Двухвыборочный критерий однородности Колмогорова-Смирнова

Критерий Манна – Уитни позволяют обнаруживать лишь различия в центральных тенденциях распределений двух случайных величин. Если важно обнаружить любые расхождения в форме распределений, то можно использовать, в частности, двухвыборочный критерий однородности Смирнова (Колмогорова-Смирнова). С помощью этого критерия проверяется гипотеза H_0: $F_1(x) = F_2(y)$ о том, что функции распределения $F_1(x)$ и $F_2(y)$ случайных величин идентичны против альтернативной гипотезы H_1: $F_1(x) \neq F_2(y)$ о том, что они различны.

Данный непараметрический критерий базируется на распределении Колмогорова

$$P(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, \quad x > 0.$$



Андрей Николаевич Колмогоров

Статистика критерия Смирнова $D_{m,n}$ определяется как максимум модуля разности между эмпирической функцией $F_1(x)$, построенной по выборке $x_1, x_2, \dots, x_n$, и эмпирической функцией $F_2(x)$, построенной по выборке $y_1, y_2, \dots, y_m$

$$D_{m,n} = \max_x \left| F_1(x) - F_2(x) \right|.$$

При справедливости гипотезы H_0 статистика $\lambda = D_{m,n}\sqrt{\frac{mn}{m+n}} < \lambda_{\text{крит}}$ имеет асимптотическое распределение Колмогорова, а $\lambda_{\text{крит}}$ определяется из $P\{\lambda > \lambda_\alpha\} = \alpha$, где α – уровень значимости.

Пример 2.5. Требуется определить, одинаковы ли функции распределения листьев по размерам у двух ярусов (частот n по выборкам X и Y) кроны акации. Данные эксперимента приведены в таблице справа.

Результаты эксперимента

диапазон размеров		эксперимент	
		n(X)	n(Y)
8,2	9,8	1	0
9,8	11,4	5	7
11,4	13,0	12	10
13,0	14,6	1	10
14,6	16,2	1	4
16,2	17,8	0	7

Алгоритм решения предусматривает следующую последовательность операций.

Из сокращенной (см. ниже) или полной (таблица $\lambda_{крит}$) таблиц критических значений статистики выбирается величина $\lambda_{крит}$, соответствующая заданному уровню значимости.

$P\{\lambda > \lambda_\alpha\} = \alpha$			
α	0,1	0,05	0,01
λ_α	1,22	1,36	1,63

Критические значения статистики Колмогорова-Смирнова (сокращенная таблица)

Далее подсчитываются суммарные частоты и частности по выборкам X и Y; строятся интегральные функции распределений. Затем определяется максимум модуля разности между эмпирическими функциями распределений и выявляется максимальное значение статистики D, через которое вычисляется значение $\lambda_{экс}$.

пример 2.5

Из сравнения $\lambda_{экс} > \lambda_{крит}$ делается заключение: на заданном уровне значимости нулевая гипотеза отвергается, т.е. функции распределения листьев по размерам у двух ярусов различны.

вопросы и задачи для самостоятельной работы

Пожелания и замечания направляете автору вот он по адресу bhf@tspu.edu.ru