Если ничто другое не помогает,
прочтите, наконец, инструкцию!
Мерфология, Аксиома Кана

Глоссарий предметный

Альтернативная (конкурирующая) гипотеза

Аппроксимация

Асимметрия

Беренса-Фишера критерий

Варианта

Вариации коэффициент

Вариации коэффициента ошибка

Вероятность P(A) события A

Вероятность статистическая

Выборочная совокупность, выборка, варианта

Генеральная совокупность

Гипотеза статистическая

Гистограмма

Группировка

Sturgess'a правило

на начало

Джини средняя разность

Дисперсионный анализ

Дисперсия выборки

Дихотомические данные

Доверительная вероятность

Доверительный интервал

Доверительный уровень

Достоверность

Закон распределения случайной величины

Закон распределения вероятностей дискретной случайной величины

Значимости уровень

Интерквартильный размах

Испытание

Yates'a поправка

Качественная переменная

Качественные данные

Квантиль

Квартиль

Класс

Кластерный анализ

Количественные данные

Количественный

Конкорданции коэффициент

Контрольная группа

Корреляции коэффициент

Критерий

на начало

Критерий статистический

Лимиты

Манна-Уитни *U* – критерий

Медиана

Межквартильный размах, интерквартильный размах

Мера точности

Многофакторный анализ

Мода, модальный класс

Модель

Мощность статистического критерия

Наблюдаемое значение оценки показателя

Наблюдение

Надежности уровень

Независимость

Непрерывные данные

Нормальной случайной величины дифференциальная функция распределения

Нулевая гипотеза

Однородность двух независимых выборок

Однофакторный дисперсионный анализ, *ANOVA* метод

Отношение шансов

Оценка

Ошибки наблюдения

Ошибки I и II рода

на начало

Планирование эксперимента

Показатель точности опыта

Полигон частот
Полигон частостей

Порядковые данные

Размах выборки
Размах вариации

Ранжирование

Ранжирования правила

Распределения функция эмпирическая (статистическая)

Регрессия

Регрессия линейная

Репрезентативность (представительность) выборки

Рошера критерий

Случайная величина

Смирнова (Колмогорова-Смирнова) двухвыборочный критерий однородности

Событие

Среднее значение выборочное, среднеарифметическое выборочное

Среднее отклонение выборочное

Стандартная ошибка по выборке

Стандартного отклонения по выборке ошибка

Стандартное отклонение по выборке

Статистика описательная

Степень(и) свободы

на начало

Стьюдента критерий для независимых выборок для разных дисперсий (гетероскедастический тест)

Стьюдента критерий парный

Стьюдента критерий для независимых выборок для равных дисперсий (гомоскедастический тест)

Стьюдента критерий равенства среднего значения

Sturgess'a правило

Тест статистический

Томпсона правило, Рошера критерий

Фишера *F*-критерий, (критерий Фишера-Снедекора)

Функция распределения случайной величины интегральная

χ^2 критерий Пирсона

Частот (частостей) гистограмма

Частота
Частость

Шансы

Шкалирование

Эксцесс

"Ящик с усами" правило

на начало

Альтернативная (конкурирующая) гипотеза

Гипотеза, которая является логическим отрицанием нулевой гипотезы.

Аппроксимация

Аппроксимация (приближение) – математический метод, состоящий в замене одних математических объектов другими, в том или ином смысле близкими к исходным, но более простыми. Аппроксимация позволяет исследовать числовые характеристики и качественные свойства объекта сведением задачи к изучению этих более простых объектов. Примером аппроксимации является, в частности, линейная регрессия, способом ее построения – метод наименьших квадратов.

Асимметрия (skewness)

Асимметрия характеризует форму статистического распределения. Если коэффициент асимметрии больше нуля, асимметрия правосторонняя (положительная), форма кривой распределения скошена вправо относительно кривой плотности нормального распределения. Если коэффициент асимметрии меньше нуля, то асимметрия левосторонняя (отрицательная), форма кривой распределения скошена влево относительно кривой плотности нормального распределения.

на начало

Коэффициент асимметрии выборочной совокупности вычисляется по формуле:

$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma} \right)^3,$$

где n – численность выборки,

x_i, i=1,2...,n – значения вариант выборки,

$\overline{x}$ – выборочное среднее значение,

σ – стандартное отклонение.

Дисперсия асимметрии определяется формулой

$$D_A = \frac{6(n-1)}{(n+1)(n+3)}.$$

Асимметрия находит применение, в частности, при исследовании формы распределения выборки.

Функция MS Excel: СКОС(данные); Calc (Open Office): SKEW(данные).

Беренса-Фишера критерий

См. *Стьюдента критерий для независимых выборок для разных дисперсий*.

Варианта
variate

(лат.: vario – разнообразить, менять, видоизменять, быть непостоянным, меняться, колебаться).

Значение изменяющегося признака, показателя, переменной.

на начало

Вариации коэффициент

Коэффициент вариации представляет собой характеристику рассеяния случайной величины. Он показывает, какой процент составляет стандартное отклонение от среднего значения. Коэффициент вариации используется для установления степени выравненности совокупности по тому или иному признаку.

Коэффициент вариации вычисляется по формуле:

$$V_\sigma = \frac{\sigma}{\overline{x}} \cdot 100\%,$$

где σ – стандартное отклонение,

$\overline{x}$ – выборочное среднее значение.

Вариации коэффициента ошибка

См. замечание к *стандартная ошибка*

Ошибка коэффициента вариации вычисляется по формуле:

$$m_\Theta = \frac{\Theta}{\sqrt{2n}}.$$

Вероятность P(A) события A

Вероятность буквально означает "приемлемость в качестве истины".

Вероятность P(A) события A есть численная мера степени объективной возможности появления события A; отношение числа исходов, благоприятствующих наступлению этого события, к общему числу равновозможных исходов (классическое определение вероятности).

на начало

По классическому определению, $P(A)$ – вероятность P события A равна отношению числа исходов – m, "благоприятствующих" данному событию, к общему числу "равновозможных" исходов – n, то есть $P(A) = m/n$. Например, для реализации события A – выпадения трех очков при бросании игральной кости число благоприятных (т.е. выпадение тройки) "исходов" равно m=1; общее число возможных событий (выпадение какого-либо числа очков) равно n=6; а вероятность – P того, что выпадет тройка (событие A) равна: $P(A) = 1/6$ или $P(A) \approx 16{,}7\%$.

Вероятность достоверного события равна единице (или к 100%). Вероятность невозможного события равна нулю.

Строго говоря, в действительности нет событий ни достоверных, ни невозможных, но есть события лишь случайные, вероятные. Вероятность случайного события есть положительное число, заключенное между нулем и единицей: $0<P(A)<1{,}0$. Для оценки вероятности пользуются относительной частотой события, вычисляемой по фактически проведенным наблюдениям.

Вероятность статистическая

Относительная частота (частость) $P(A)$ появления события A в n произведенных испытаниях.

на начало

Выборочная совокупность, выборка, варианта

Выборочной совокупностью (выборкой) называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Более строго: *выборка* – это последовательность независимых одинаково распределенных случайных величин, распределение каждой из которых совпадает с распределением генеральной случайной величины.

Число объектов (наблюдений) в совокупности называется ее *объемом.*

Конкретные значения выборки, полученные в результате наблюдений (испытаний), называют *реализацией выборки, обычно* обозначают строчными буквами $x_1, x_2 x_3, \ldots, x_n$ и называют *вариантами.*

Генеральная совокупность parent population, general population

(лат.: generalis – общий, всеобщий).

Генеральной совокупностью называют совокупность объектов, из которых производится выборка.

Это совокупность всех подлежащих изучению объектов или мыслимых результатов наблюдений, которые могут быть получены в данных (неизменных) условиях. Различают конечные, содержащие конечное число элементов, и бесконечные, содержащие бесконечное число элементов, генеральные совокупности.

на начало

Генеральная совокупность может содержать конечное число объектов. Однако если это число достаточно велико, то иногда, в целях упрощения вычислений или для облегчения теоретических выводов, допускают, что генеральная совокупность состоит из бесчисленного множества объектов. Такое допущение оправдывается тем, что увеличение объёма генеральной совокупности практически не сказывается на результатах обработки данных выборки.

Генеральная совокупность характеризуется генеральным распределением с генеральными параметрами, например генеральным математическим ожиданием и генеральной дисперсией. Генеральные параметры могут быть оценены по выборочным данным.

Гипотеза статистическая

Различного рода предположения относительно характера или параметров распределения случайной переменной, которые можно проверить, опираясь на результаты наблюдений в случайной выборке.

Пусть дана выборка $X = (X_1, \dots, X_n)$ из неизвестного совместного распределения $\mathbb{R}^X$. Тогда любое утверждение, касающееся природы $\mathbb{R}^X$ называется статистической гипотезой.

на начало

Гипотезы различают по виду предположений, содержащихся в них:

- Статистическая гипотеза, однозначно определяющая распределение $\mathbb{R}^X$, то есть H: $\{\mathbb{R}^X = \mathbb{R}_0\}$ где $\mathbb{R}_0$ какой-то конкретный закон, называется простой.
- Статистическая гипотеза, утверждающая принадлежность распределения $\mathbb{R}^X$ к некоторому семейству распределений, то есть вида H: $\{\mathbb{R}^X \in \mathcal{R}\}$, где $\mathcal{R}$ – семейство распределений, называется сложной.

На практике обычно требуется проверить какую-то конкретную и как правило простую гипотезу H_0. Такую гипотезу принято называть *нулевой*. При этом параллельно рассматривается противоречащая ей гипотеза H_1, называемая *конкурирующей* или *альтернативной*.

Выдвинутая гипотеза нуждается в проверке, которая осуществляется статистическими методами, поэтому гипотезу называют статистической. Для проверки гипотезы используют критерии, позволяющие принять или опровергнуть гипотезу. Статистической гипотезой называется любое предположение о виде неизвестного распределения или о параметрах известного распределения.

на начало

Этапы проверки статистических гипотез

1. Формулировка основной гипотезы H_0 и конкурирующей гипотезы H_1. Гипотезы должны быть чётко формализованы в математических терминах.
2. Задание вероятности α, называемой уровнем значимости и отвечающей ошибкам первого рода, на котором в дальнейшем и будет сделан вывод о правдивости гипотезы.
3. Расчёт статистики φ критерия такой, что:
   - её величина зависит от исходной выборки $X = (X_1, \dots, X_n)$: $\varphi = (X_1, \dots, X_n)$;
   - по её значению можно делать выводы об истинности гипотезы H_0;
   - сама статистика φ должна подчиняться какому-то известному закону распределения, т.к. сама φ является случайной в силу случайности X.
4. Построение критической области. Из области значений φ выделяется подмножество $\mathbb{C}$ таких значений, по которым можно судить о существенных расхождениях с предположением. Его размер выбирается таким образом, чтобы выполнялось равенство $P(\varphi \in \mathbb{C}) = \alpha$. Это множество $\mathbb{C}$ и называется критической областью.
5. Вывод об истинности гипотезы. Наблюдаемые значения выборки подставляются в статистику φ и по попаданию (или непопаданию) в критическую область $\mathbb{C}$ выносится решение об отвержении (или принятии) выдвинутой гипотезы H_0.

на начало

Выделяют три вида критических областей:

- Двусторонняя критическая область определяется двумя интервалами $(-\infty, x_{\alpha/2}) \cup (x_{1-\alpha/2}, +\infty)$ где $x_{\alpha/2}, x_{1-\alpha/2}$ находят из условий $P(\varphi < x_{\alpha/2}) = \alpha/2$, $P(\varphi < x_{1-\alpha/2}) = 1 - \alpha/2$.

- Левосторонняя критическая область определяется интервалом $(-\infty, x_{\alpha})$, где x_{α} находят из условия $P(\varphi < x_{\alpha}) = \alpha$.

- Правосторонняя критическая область определяется интервалом $(x_{\alpha}, +\infty)$, где x_{α} находят из условия $P(\varphi > x_{\alpha}) = \alpha$.

Гистограмма

Гистограмма представляет собой дискретный или интервальный вариационный ряд (ряд распределения), полученный в результате группировки исходной эмпирической выборки, измеренной в порядковой или количественной шкале, по особым образом подобранным классовым интервалам.

Данный вариационный ряд служит основой для многих статистических алгоритмов, таких, как глазомерный метод проверки нормальности распределения, установление типа распределения, и других.

на начало

Группировка, Sturgess'a правило

Имеется два пути практической группировки: задавшись границами классовых интервалов (классов) или задавшись их количеством.

В "Пакете анализа" электронных таблиц Microsoft Excel имеются удобные средства группировки для случая, когда пользователем заданы границы классовых интервалов (в терминологии Microsoft Excel – карманов), либо они устанавливаются автоматически по специальному встроенному в "Пакет анализа" алгоритму. Поэтому выполнить группировку не представляет труда. Во втором случае для вариационного ряда число классов равно числу градаций переменной, выбранном исследователем. При этом число классов дискретного вариационного ряда обычно равно числу градаций вариант выборки, измеренной в порядковой шкале. Для интервального вариационного ряда число классов задается на основе какого-либо правила. Возможность задать число классов Microsoft Excel не предоставляет.

Критерием правильности выбора количества классов считается верная передача типа распределения эмпирических частот данной выборочной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении можно затушевать реальную картину распределения частот случайными отклонениями.

на начало

Выделяются несколько способов вычисления числа классов для выборок умеренной численности. Весьма распространенным является правило Sturgess'a, выражаемое формулой

$$M = 1 + \log_2 n$$
$$\textit{или} \quad M \approx 1 + 1{,}44 \cdot \ln n$$
$$\textit{или} \quad M \approx 1 + 3{,}32 \cdot \lg n,$$

где M – число классов, n – численность выборочной совокупности.

После решения вопроса о числе классов производится вычисление границ классовых интервалов и разнесение вариант исходной количественной выборки по классовым интервалам.

Джини средняя разность

Средняя разность Джини характеризует разброс значений вариант эмпирической выборки друг относительно друга и не зависит от какого-либо центрального значения, например, от среднего значения или медианы. Вычисление выборочной средней разности Джини производится по формуле

$$g = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |x_i - x_j|,$$

где n – численность выборки,

x_i, i=1,2...,n – значения вариант выборки.

на начало

Дисперсионный анализ
analysis of variance

(лат.: dispergo, -persi, -persum – рассыпать, рассеивать, разбрасывать).

Статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента, а также для последующего планирования экспериментов.

Первоначально дисперсионный анализ был предложен в 1925 году английским математиком Р. Фишером (Fisher R.A.).

Дисперсионный анализ может быть сделан с помощью статистических программ для персональной ЭВМ. Из многочисленных статистических программ можно рекомендовать хорошо известные программы:

- Statistica (URL: http://www.statsoftinc.com/textbook/stathome.html) или
- SPSS (URL: http://www.spssscience.com/spss11).

Дисперсия выборки, стандартное отклонение по выборке

Основным статистическим показателем, характеризующим разброс *выборки*, является дисперсия, вычисляемая по формуле:

$$\sigma^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$$

где n – численность выборки,

x_i – значения вариант выборки,

$\overline{x}$ – *выборочное среднее значение*.

на начало

В иностранной литературе дисперсию иногда называют вариансой.

Представленная формула вычисляет так называемую несмещенную (исправленную) выборочную оценку дисперсии. Формула смещенной оценки отличается от показанной формулы делителем не (n-1), а n. Предполагается, что формула смещенной оценки должна использоваться, если известна вся генеральная совокупность, что на самом деле, конечно, встречается редко.

Стандартным отклонением σ (средним квадратическим отклонением, средним квадратичным отклонением) называют корень квадратный из дисперсии $\sigma = \sqrt{\sigma^2}$.

Стандартное отклонение – это мера того, насколько широко разбросаны точки данных относительно их среднего.

Функция MS Excel: для σ^2: ДИСП(данные); для σ: СТАНДОТКЛОН(данные).
Функции Calc (Open Office): VAR(данные), STDEV(данные).

Дихотомические данные (dichotomous data) – признаки, которые могут иметь только два значения (присутствует–отсутствует, да–нет, жив–умер).

на начало

Доверительная вероятность, доверительный уровень

Доверительная вероятность (доверительный уровень) требуется для вычисления ряда выборочных статистических показателей, и, в отличие от ряда других параметров, является не вычисляемой по выборке, а задаваемой исследователем величиной.

Доверительная вероятность выбирается из следующей стандартной линейки (в основном, следуя классификации Плохинского):

- Нулевой порог 0,90 применяется для работы с пониженной ответственностью, при первом ознакомлении с явлением.
- Первый порог 0,95 применяется в большинстве исследований (например, биологические исследования).
- Второй порог 0,99 для работ с повышенной ответственностью (например, медицинские исследования).
- Третий порог 0,999 применяется для работ с высокой ответственностью (например, исследования эффективности лекарств).

Доверительный уровень может быть выражен в долях, например, 0,95, либо в процентах – 95%.

на начало

Доверительный интервал

Доверительный интервал среднего значения вычисляется на заданном *доверительном уровне*, выражаемом в долях или процентах. Доверительный интервал, вычисленный на доверительном уровне, например, 95% (или, то же самое, 0,95), означает, что 95% вариант выборочной совокупности попадают в данный интервал.

Доверительный интервал для среднего представляет интервал значений вокруг оценки, где с данным уровнем доверия, находится "истинное" (неизвестное) среднее выборки. Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p=0,95 равны 19 и 27 соответственно, то можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее выборки. Если устанавливается больший уровень доверия, то интервал становится шире и возрастает вероятность, с которой он "накрывает" неизвестное среднее (и наоборот). Хорошо известно, например, что чем "неопределенней" прогноз погоды (т.е. шире доверительный интервал), тем вероятнее, что он будет верным. Очевидно, что ширина доверительного интервала зависит от объема или размера выборки, а также от разброса (изменчивости) данных.

Увеличение размера выборки делает оценку среднего более надежной. Увеличение разброса наблюдаемых значений уменьшает надежность оценки.

на начало

Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Если это предположение не выполнено, то оценка может оказаться плохой (особенно для малых выборок). При увеличении объема выборки ($n \geq 100$) качество оценки улучшается и без предположения нормальности выборки.

Если распределение малой эмпирической выборки отличается от нормального, следует, в общем случае, пользоваться непараметрическим доверительным интервалом.

Вычисление двустороннего доверительного интервала в случае, если эмпирическая выборка распределена нормально, производится по формуле:

$$I_m = \left(\overline{x} - t_\alpha \frac{\sigma}{\sqrt{n}};\ \overline{x} + t_\alpha \frac{\sigma}{\sqrt{n}}\right) \quad или \quad I_m = (\overline{x} - t_\alpha m;\ \overline{x} + t_\alpha m)$$

σ – стандартное отклонение,

t_α – значение обратной функции – распределения Стьюдента с параметрами (n-1) и (1+α)/2.

α – уровень значимости, выраженный в долях,

m – стандартная ошибка.

Вычисление доверительного интервала для выборочной совокупности производится также, как и уровня надежности.

на начало

Для генеральной совокупности доверительный интервал вычисляется по формуле

$$I_m = \left(\overline{x} - U_\alpha \frac{\sigma}{\sqrt{n}};\ \overline{x} + U_\alpha \frac{\sigma}{\sqrt{n}}\right),$$

где U_α определяет ту часть *стандартной нормальной кривой*, которая равняется (1-α).

Данную величину в MS Excel можно определить через функцию СТЬЮДРАСПОБР(α;100000). Кроме того, в MS Excel имеется функция ДОВЕРИТ(α; σ; n) {функция CONFIDENCE(α;σ;n) в Calc (Open Office)}, определяющая искомое отклонение $U_\alpha \frac{\sigma}{\sqrt{n}}$.

Доверительный уровень

См. доверительная вероятность, значимости уровень, надежности уровень.

Достоверность (validity)

– характеристика, показывающая в какой мере результат измерения соответствует истинной величине. Достоверность исследования определяется тем, в какой мере полученные результаты справедливы в отношении данной выборки (internal validity).

на начало

Закон распределения случайной величины

Всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Закон распределения вероятностей дискретной случайной величины
distribution of discrete random variable

Законом распределения дискретной случайной величины называют перечень всех возможных ее значений и их вероятностей.
Обычно закон распределения задается в виде таблицы, одна графа которой содержит все возможные значения случайной величины, а вторая – соответствующие им вероятности: $X: x_1, x_2, \dots, x_n$ $P: p_1, p_2, \dots, p_n$. Представление закона распределения возможно также в виде формулы, посредством параметров или в виде графика.

Значимости уровень

Уровень значимости – одна из характеристик качества критерия статистической проверки гипотез. Пусть выдвинута гипотеза H_0 (основная, или "нулевая"). Всякое статистическое решение, принимаемое на основе ограниченного ряда наблюдений, неизбежно сопровождается *вероятностью* ошибочного заключения.

на начало

С вероятностью α (альфа) гипотеза H_0 может оказаться отвергнутой, в то время как на самом деле она является справедливой (*ошибка первого рода*), или, наоборот, с вероятностью β (бэта) может быть принята гипотеза H_0 в то время, как на самом деле она является ошибочной (*ошибка второго рода*). В частности, при фиксированном объеме выборки обычно задаются величиной альфа вероятности ошибочного отвержения проверяемой гипотезы H_0. Эту вероятность ошибочного отклонения "нулевой" гипотезы принято называть уровнем значимости.

На практике часто пользуются следующими стандартными значениями уровня значимости α: 0,1 ; 0,05 ; 0,025 ; 0,01 ; 0,005 ; 0,001. Особенно распространенной является величина уровня значимости α= 0,05. Она означает, что в среднем в пяти случаях из ста ошибочно отвергают высказанную гипотезу при пользовании данным статистическим критерием.
Задаваемая исследователем величина.

Интерквартильный размах

См. *межквартильный размах*.

Испытание

Наблюдение того или иного явления (события) при осуществлении определенного комплекса условий (наблюдение того же явления в других условиях считается другим испытанием).

на начало

Yates'a поправка

Поправка Yates'a (Йетса) применяется для уточнения критерия χ^2 в случае таблицы 2×2 (то есть при 1 степени свободы). Уточнение обусловлено тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен.

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

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

Качественная переменная qualitative/categorical variable

Качественная (или категориальная) переменная – это переменная, которая не может иметь количественных числовых значений. К качественным переменным относят номинативную переменную и порядковую переменную.

Качественные данные (nominal data)

– это такие признаки, которые нельзя выразить количественно. Например: диагноз, группа крови, страна проживания. Частным случаем качественных данных являются дихотомические данные.

на начало

Квантиль
quantile

(лат.: quantum – насколько возможно, quantillus – насколько возможно малый).

Одна из числовых характеристик распределения вероятностей $F(X)$ случайной величины X. Квантили дают представление о расположении и рассеянии данного распределения. Квантиль X_Q порядка Q ($0<Q<1$) определяется как корень уравнения $F(X_Q) = Q$, т.е. как значение X_Q (по аргументу Q) функции $F^{-1}(Q)$, обратной к $F(X)$. Вероятность Q (порядок квантили) может быть задана дробью или в процентах.

В общем, название квантили определяется её порядком. Так, $X_{30\%}$ или то же, что $X_{0,3}$, называют 30-ти процентной квантилью, $X_{5\%}$ – 5-ти процентной квантилью и т.д. Для некоторых квантилей, кроме общих названий, могут использоваться и специальные названия. Например, $X_{50\%}$ (50-ти процентная квантиль), или $X_{0,5}$, или $X_{1/2}$ называется медианой распределения $F(X)$; $X_{1/4}$, $X_{1/2}$, $X_{3/4}$ – квартили (первая, вторая, третья квартили); $X_{0,1}$, $X_{0,2}$,..., $X_{0,9}$ – дециль (первая, вторая,..., 9-я децили); $X_{0,01}$, $X_{0,02}$,..., $X_{0,99}$ – процентили распределения (первая, вторая,..., 99-я процентили).

Таким образом, квантиль X_Q – это значение переменной X, ниже которого лежит $100Q$ процентов распределения. Или иначе, квантиль X_Q есть такое значение случайной переменной X, для которого вероятность того, что это значение X меньше, чем X_Q равна Q: $P\{X<X_Q\} = Q$.

на начало

Квантили существуют у каждого распределения вероятностей. Для широко используемых в статистических исследованиях распределений они однозначно определены и представлены в таблицах. Они используются при решении многих статистических задач, например при проверке статистических гипотез и при оценивании (параметров генеральных совокупностей) по выборочным данным.

Существительное квантиль женского рода, ударение на вторую гласную.

Квартиль

Квартили (а также медиана) обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности.

Функция MS Excel: КВАРТИЛЬ(данные; часть);

Функция Calc (Open Office): QUARTILE(данные; часть).

Если часть равна	То КВАРТИЛЬ {QUARTILE} возвращает
0	Минимальное значение
1	Первую квартиль (25-ую персентиль)
2	Значение медианы (50-ую персентиль)
3	Третью квартиль (75-ую персентиль)
4	Максимальное значение

на начало

Класс
class, set

(лат.: classis – разряд, класс, группа, смена). Совокупность однородных предметов, обладающих каким-то определенным качеством, свойством или отношением.

Кластерный анализ
cluster analysis

Статистический метод выделения групп (классов, блоков) схожих между собой элементов в их множестве. Метод ориентирован как на количественные, так и качественные признаки. Имеются методы неиерархического кластерного анализа и иерархического кластерного анализа. В последнем случае результаты представляются в виде дендрограммы или дерева семейства. Кластерный анализ используется совместно с другими методами многомерного анализа для описания структуры сложных совокупностей данных.

Количественные данные
(numerical, or interval data)

– это такие величины, которым присущ естественный порядок расположения последовательных значений независимо от их места на шкале. Примеры: масса тела, концентрация глюкозы в крови.

на начало

Количественный
quantitative

Прилагательное к количество. Объективный относительный признак, свойство, которое характеризует степень пространственно-временной определенности объектов и выражается через число, меру, величину.

Примеры: количественные изменения, количественные отношения, количественные методы, количественный состав.

Конкорданции коэффициент
coefficient of concordance

(лат.: со – *приставка* с, вместе + efficiens – побудительная причина; concordia – согласие, гармония).

Коэффициент согласованности. Характеристика связи между несколькими признаками, измеряемыми в порядковой шкале.

Коэффициент конкорданции чаще всего используется как характеристика согласованности мнений нескольких экспертов о значимости тех или иных признаков.

Контрольная группа
control group

Группа, данные которой служат эталоном для сравнения.

на начало

Корреляции коэффициент

Коэффициент корреляции используется для определения наличия взаимосвязи между двумя свойствами. Например, можно установить зависимость между номером яруса кроны дерева и средней длиной листа.

Зависимость подразумевает влияние, связь – любые согласованные изменения, которые могут объясняться различного рода причинами. Корреляционные связи не могут рассматриваться как свидетельство причинно-следственной связи, они указывают лишь на то, что изменениям одного признака, как правило, сопутствуют определенные изменения другого. Но находится ли причина изменений в одном из признаков или она оказывается за пределами исследуемой пары признаков, остается неизвестным.

Корреляционная связь может быть положительной ("прямой") и отрицательной ("обратной"). При положительной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким – низкие значения другого. При отрицательной корреляции соотношения обратные.

При положительной корреляции коэффициент корреляции имеет положительный знак, например r = +0,3, при отрицательном корреляции – отрицательный знак, например r =-0,3.

Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. Максимально возможное абсолютное значение коэффициента корреляции r =1; минимальное r =0.

на начало

Справа в таблице приведена наиболее распространенная классификация корреляционных связей.

сильная (тесная)	$r \geq 0{,}70$
средняя	$0{,}5 \leq r < 0{,}7$
умеренная	$0{,}3 \leq r < 0{,}5$
слабая	$0{,}2 \leq r < 0{,}3$
очень слабая	$r < 0{,}2$

Функция MS Excel: КОРРЕЛ(данные1; данные2);
Функция Calc (Open Office): CORREL(данные1; данные2).

Критерий
criterion

(греч.: κρίτήριον – средство для решения, признак по которому можно судить верно).

1. Правило проверки гипотезы, приводящее с определенной вероятностью ошибки к некоторому заключению. Признак, который при оценке функционирующих объектов рассматривается как наиболее существенный.
2. Прогнозируемое тестом поведение, или независимая от тестирования оценка поведения.

на начало

Критерий статистический, тест статистический

Правило, по которому нулевая гипотеза отвергается или принимается.

Лимиты

Лимиты (пределы) – значения крайних классов.

Манна-Уитни *U* – критерий, Mann-Whitney *U* test

Критерий предназначен для оценки различий между двумя выборками по уровню какого-либо количественно измеренного признака, и что более важно, критерий Манна-Уитни позволяет оценивать выборки при ненормальном распределении вариант. Кроме того, критерий позволяет выявлять различия между малыми выборками, когда $n_1, n_2 \geq 3$ или $n_1=2, n_2>5$.

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

Чем меньше перекрещивающихся значений, тем более вероятно, что различия достоверны.

Чем меньше $U_{эмп}$, тем более вероятно, что различия достоверны.

на начало

Нулевая гипотеза формулируется следующим образом: уровень признака в выборке 2 не ниже уровня признака в выборке 1.

Статистика критерия подсчитывается как

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

где n_1,n_2 – количество вариант в первой и второй выборках;

T_* – большая из двух ранговых сумм;

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

Критические значения U определяются по соответствующим таблицам (см. Приложение 6). Если $U > U_{кр}$ нулевая гипотеза принимается. Если $U < U_{кр}$, то нулевая гипотеза отвергается. Чем меньше значения U, тем достоверность различий выше.

Медиана

Медиана – это число, которое является серединой множества чисел выборки, то есть половина чисел имеют значения большие, чем медиана, а половина чисел имеют значения меньшие, чем медиана. Для вычисления медианы количественной выборки численностью n сначала строится интервальный вариационный ряд, т.е. исходная выборка упорядочивается по возрастанию.

на начало

Для нечетного n медианой является варианта полученного интервального вариационного ряда, имеющая порядковый номер $(n+1)/2$. Для четного n медиана равна среднему значению двух средних вариант. Некоторые исследователи предпочитают медиану среднему значению, считая ее более точной оценкой меры положения выборки.

Функция MS Excel: МЕДИАНА(данные);
Calc (Open Office): MEDIAN(данные).

Межквартильный размах, интерквартильный размах

Квартили, а также медиана, обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности. Вычисление данных показателей производится по правилам, принятым для вычисления медианы.

Межквартильный размах выборки (интерквартильный размах) характеризует степень разброса данных в абсолютных числах. Выборочный межквартильный размах – это разность между верхней и нижней квартилями выборки, иначе 75% и 25% процентилями выборки. Вычисление межквартильного размаха упорядоченной по возрастанию количественной выборки производится по формуле:

$$R_\mu = \mu_{3/4} - \mu_{1/4},$$

где $\mu_{3/4}$ – значение верхней квартили выборки,

$\mu_{1/4}$ – значение нижней квартили выборки.

на начало

Межквартильный размах является более репрезентативной оценкой разброса значений выборки по сравнению с точечной оценкой стандартного отклонения. Точечная оценка *стандартного отклонения* для нормально распределенной совокупности может быть получена из межквартильного размаха как

$$\sigma = \frac{R_\mu}{2\psi(0{,}75)} \approx 0{,}741301 R_\mu ,$$

где ψ – функция, обратная функции стандартного нормального распределения.

Межквартильный размах находит применения в качестве основы одного из методов выявления аномальных наблюдений (выбросов).

Величина $0{,}5R_\mu$ также используется как характеристика рассеяния и называется семи-интеркварительной широтой.

Вычисление R_μ
функциями MS Excel: КВАРТИЛЬ(данные;3) - КВАРТИЛЬ(данные;1)
Функции Calc (Open Office): QUARTILE(данные;3) - QUARTILE(данные;1)

Мера точности
measure of exactness / accuracy / precision, degree of exactness / accuracy / precision]

Характеристика рассеяния значений случайной величины. Термин раздела математической статистики – теории ошибок.

Мера точности определяется как $h = \frac{1}{\sqrt{2}\sigma}$; здесь σ – среднее квадратичное отклонение ошибки $\delta_i = x_i - \mu$, то есть разности между математическим ожиданием μ и значениями x_i наблюдаемой случайной величины X.

на начало

Многофакторный анализ (multivariable analysis) – это совокупность статистических методов, которые одновременно рассматривают влияние многих переменных на какой–либо один фактор. Если после устранения влияния этих переменных действие фактора сохраняется, его воздействие считается независимым. Кроме того, эти методы применяются для выделения из большого числа признаков малого подмножества, которое вносит независимый и существенный вклад в конечный результат (исход), что используется для ранжирования переменных по силе их воздействия на исход и для других целей.

Мода, модальный класс – Мода, Модальный класс – класс обладающий наибольшей частотой. Значение, которое принимает данный класс, называют модой.

Функция MS Excel: МОДА(данные);
Функция Calc (Open Office): MODE(данные).

Модель – Преднамеренно упрощенная схема (имитация) некоторой части реальной действительности, с помощью которой исследователь получает рекомендации к решению реальных проблем;

на начало

система элементов, воспроизводящих определенные стороны, связи, функции объекта исследования.

Модель математическая – модель, при описании которой используется язык математики; формализованное описание с помощью математического аппарата взаимосвязей между элементами изучаемой системы.

Мощность статистического критерия
power of a criterion/test

Мощность критерия есть вероятность того, что нулевая гипотеза будет отвергнута, если верна конкурирующая гипотеза. Другими словами мощность критерия – это вероятность P, дополняющая до единицы вероятность ошибки второго рода β : $P = 1 - \beta$.
Чем больше м.к., тем вероятность ошибки 2-го рода меньше.

Наблюдаемое значение оценки показателя
observed score

Значение оценки показателя, полученное при однократном применении процедуры оценивания.

Единственное значение оценки является оценкой истинного значения уровня показателя. Из теории вероятностей, из простой логики и из практики известно, что чем больше выборка, по которой составляется представление об исследуемом объекте, тем больше вероятность соответствия истине

на начало

получаемой в исследовании оценки. Отсюда ясно, что единственное наблюдаемое значение оценки показателя является наиболее простой, и в то же время наименее надежной, (наименее вероятной) оценкой уровня истинного значения показателя объекта.

Очевидно, что по такому единственному наблюдаемому значению невозможно составить представление о другом не менее важном общем сущностном показателе кроме уровня – вариативности. Для того, чтобы составить представление о вариативности, необходимо как минимум два наблюдаемых значения показателя. Тогда размах между наблюдаемыми значениями может служить простейшей оценкой вариативности. При большем числе наблюдений возможны более надежные оценки вариативности по сравнению с размахом.

Наблюдение
observation

Статистическое наблюдение – этап научного исследования, процесс организованного получения данных о вероятностных сущностях и явлениях. Наблюдение включает в себя измерение вероятностных сущностей и явлений. В теории вероятностей и математической статистике наблюдением также называют единичный результат процесса получения данных. Наблюдение в науке является разновидностью эксперимента.

на начало

В теории планирования эксперимента различают активный и пассивный эксперимент. Наблюдение без активного вмешательства в процесс исследования называется пассивным экспериментом.

Различают следующие способы наблюдения: текущее, периодическое, единовременное, сплошное наблюдение, наблюдение основного массива (наиболее значимой доли совокупности), выборочное, непосредственное, документальный способ наблюдения, способ опроса и т.д.

Наблюдение является этапом исследования. Результаты этого этапа определяются эффективностью осуществления предшествующих этапов исследования. Этими предшествующими этапами могут быть:

(1) Постановка цели исследования.

(2) Анализ существующих знаний, соответствующих избранной цели исследования.

(3) Выбор методологии исследования.

(4) Формулирование исходных гипотез.

(5) Планирование исследования.

(6) Создание теоретической модели и ее изучение.

(7) Планирование эксперимента.

на начало

Допустим, все эти этапы проведены эффективно, следовательно без ошибок. Следующим этапом может быть проведение эксперимента. Наблюдение – разновидность научного эксперимента, называемая в теории планирования эксперимента пассивным экспериментом. Статистическое наблюдение включает в себя следующие ступени:

(1) Определение объекта наблюдения, той сущности или явления, характеристики которого интересуют исследователя.

(2) Определение характеристик объекта подлежащих измерению: либо наблюдаемых характеристик определяемых понятиями, либо конструктов.

(3) Определение показателей, соответствующих выбранным характеристикам объекта исследования.

(4) Определение переменных, которыми обозначаются выбранные показатели.

(5) Выбор или создание шкалы для измерения значений оценок переменных.

(6) Определение (стандартизация) условий измерения.

(7) Определение способа наблюдения (см. выше), правил получения данных и создание программы наблюдения.

(8) Проведение измерений, получение данных и оформление отчета о наблюдениях.

Этим собственно наблюдение завершается. Исследователь переходит к очередному этапу исследования – статистическому анализу данных.

на начало

Очевидно, что на любой ступени (1) – (8) могут быть совершены ошибки наблюдений – неправильные действия, повлекшие за собой непредусмотренный исследованием результат – несоответствие результата наблюдения истине.

Надежности уровень

Уровень надежности показывает возможное отклонение среднего по выборке от среднего для генеральной совокупности при заданном уровне, выражаемом в долях или процентах. Уровень надежности равняется 100(1-α)%; другими словами α=0,05 означает 95% уровень надежности. Вычисление уровня надежности в случае, если эмпирическая выборка распределена нормально, производится по формуле:

$$\mathbb{N} = t_\alpha \frac{\sigma}{\sqrt{n}},$$

где σ – *стандартное отклонение*,

t_α – значение обратной функции – распределения Стьюдента с параметрами (n-1) и (1+α)/2.

α – уровень значимости, выраженный в долях.

Вычисление значения уровня надежности.

Через функции MS Excel: =СТАНДОТКЛОН(данные) / КОРЕНЬ(n) * СТЬЮДРАСПОБР(α;n -1);

посредством Calc (Open Office): =STDEV(данные) / SQRT(n) * TINV(α;n -1).

на начало

Независимость
independence

Независимость – это свойство какой-либо сущности и/или явления, характеризующее практически достоверное отсутствие отношений зависимости, связи, взаимодействия этих сущностей и/или явлений.

Независимый – обладающий свойством независимости.

Непрерывные данные
(continuous data)

– это количественные данные, которые могут принимать любое значение на непрерывной шкале. Примеры: масса тела, артериальное давление, парциальное давление кислорода в артериальной крови.

Нормальной случайной величины дифференциальная функция распределения

В работах Гаусса и Лежандра утверждалось положение о нормальном законе распределения ошибок наблюдений. Нормальный закон распределения (или распределение Гаусса) задается следующей дифференциальной функцией

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

где σ – *стандартное отклонение*,

$\overline{x}$ – выборочное среднее значение.

на начало

Нулевая гипотеза

Согласно нулевой гипотезе H_0 предполагается, что между рассматриваемыми показателями достоверного различия нет, т.е., в частности, обе анализируемые группы составляют однородный материал, одну совокупность (различий "ноль").

Однородность двух независимых выборок

Понятие "однородности", то есть "отсутствия различия", может быть формализовано в терминах вероятностной модели различными способами.

Наивысшая степень однородности достигается, если обе выборки взяты из одной и той же генеральной совокупности, т. е. справедлива нулевая гипотеза H_0: $F(x) = G(x)$ для всей рассматриваемой области x.

Отсутствие однородности означает, что верна альтернативная гипотеза H_1: $F(x) \neq G(x)$ хотя бы при одном значении аргумента x_*. Если гипотеза H_0 принята, то выборки можно объединить в одну, если нет – нельзя.

В некоторых случаях целесообразно проверять не совпадение функций распределения, а совпадение некоторых характеристик случайных величин X и Y – математических ожиданий, медиан, дисперсий, коэффициентов вариации и др. Например, однородность математических ожиданий означает, что справедлива гипотеза H_0^M: $M(X) = M(Y)$.

на начало

Среди множества статистических критериев, параметрических и непараметрических, используемых для проверки однородности, выделяют три группы критериев: для проверки гипотез о средних (о математических ожиданиях и медианах), для проверки гипотез о характеристиках рассеяния (о дисперсиях и размахах), для проверки гипотез о законах распределения.

Однофакторный дисперсионный анализ, *ANOVA* метод

При однофакторном дисперсионном анализе (дисперсионном анализе по одному признаку, *ANOVA* (**An**alysis **o**f **Va**riance – анализ отклонений, вариаций)) предполагается, что результаты наблюдений для разных уровней представляют собой выборки из нормально распределенных генеральных совокупностей.

Эти совокупности имеют свои средние и дисперсии, которые полагаются одинаковыми и не зависят от уровней. Задачей анализа является проверка нулевой гипотезы о равенстве средних рассматриваемых совокупностей. Вычисление критерия производится по формуле

$$T = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k} n_i(\overline{x_i}-\overline{x})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2},$$

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} - \textit{общее среднее},$$

где $N = \sum_{i=1}^{k} n_i$ – *численность*,

n_i – численность i-й выборки,

$\overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ – общее среднее для i-й выборки,

k – число столбцов (выборок).

на начало

Сумма, стоящая в числителе формулы вычисления критерия, служит приближенной мерой вариации между анализируемыми выборками, а двойная сумма, стоящая в знаменателе, служит мерой вариации внутри выборок.

Статистика критерия имеет F-распределение с параметрами k и N.

Одна из частых ошибок – применение t-критерия Стьюдента для проверки гипотез о равенстве трех и более групповых средних. В этом случае целесообразно применять представленный выше метод.

Отношение шансов (odds ratio) – определяется как отношение шансов события в одной группе к шансам события в другой группе, или как отношение шансов того, что событие произойдет к шансам того, что событие не произойдет. В исследованиях случай – контроль отношение шансов используется для оценки относительного риска.

Оценка
1. estimation, evaluation; 2. scoring; 3. assess, assessment

Вероятностное высказывание, суждение о свойствах объекта. Соответствующий процесс оценивания и его результат.

на начало

Получение оценки включает в себя:

- ✓ выбор показателей, характеризующих свойства объекта,
- ✓ выбор (разработка) методов и методик определения количественных значений этих показателей,
- ✓ установление базовых значений этих показателей,
- ✓ расчет реальных значений показателей,
- ✓ сравнение расчетных значений с базовыми.

Ошибки наблюдения
errors of observation,
observational errors

Ошибки статистического наблюдения – ошибки исследования, совершенные в процессе организованного получения данных о вероятностных сущностях и явлениях.

Наблюдение является этапом исследования. Результаты этого этапа определяются эффективностью осуществления предшествующих этапов исследования. Этими предшествующими этапами могут быть:

(1) Постановка цели исследования.
(2) Анализ существующих знаний, соответствующих избранной цели исследования.
(3) Выбор методологии исследования.
(4) Формулирование исходных гипотез.
(5) Планирование исследования.
(6) Создание теоретической модели и ее изучение.
(7) Планирование эксперимента.

на начало

Допустим, все эти этапы проведены эффективно, следовательно, без ошибок. Следующим этапом может быть проведение эксперимента. Наблюдение – разновидность научного эксперимента, называемая в теории планирования эксперимента пассивным экспериментом.

Статистическое наблюдение включает в себя следующие ступени:

(a) Определение объекта наблюдения.
(b) Определение характеристик объекта подлежащих измерению: либо наблюдаемых характеристик определяемых понятиями, либо конструктов.
(c) Определение показателей, соответствующих выбранным характеристикам объекта исследования.
(d) Определение переменных, которыми обозначаются выбранные показатели.
(e) Выбор или создание шкалы для измерения значений оценок переменных.
(f) Определение (стандартизация) условий измерения.
(g) Определение способа наблюдения (см. выше), правил получения данных и создание программы наблюдения.
(h) Проведение измерений, получение данных и оформление отчета о наблюдениях.

Этим собственно наблюдение завершается. Исследователь переходит к очередному этапу исследования – статистическому анализу данных.

Очевидно, что на любой ступени (a) – (h) могут быть совершены ошибки наблюдений – неправильные действия, повлекшие за собой непредусмотренный исследованием следствие – несоответствие результата наблюдения истине.

на начало

В соответствии с пунктами (a) – (h) можно назвать наиболее типичные ошибки наблюдения: неверное определение объекта исследования, характеристик объекта, показателей, переменных, ошибки в шкалировании или использовании готовых шкал, неправильное определение или несоблюдение правильно определенных условий наблюдения, ошибки измерений, ошибки оформления данных. Приведенный список возможных ошибок не ограничивается перечисленными. Очевидно, что при осуществлении исследования по программе, отличающейся от приведенной, иными могут быть и ошибки наблюдения.

Ошибки I и II рода

Ошибкой I рода называют неприятие верной нулевой гипотезы, ошибкой II рода – принятие ложной нулевой гипотезы.

Планирование эксперимента experimental design

Математическая теория экстремальных экспериментов, позволяющая выбирать оптимальную стратегию исследования при неполном знании процесса.

Это раздел математической статистики, изучающий рациональную организацию измерений случайных переменных.

на начало

Поскольку все переменные, описывающие поведение живых систем – случайные переменные, планирование эксперимента является инструментом научных исследований, адекватным сущности живых объектов. Применение теории планирования эксперимента существенно (в десятки, сотни раз) повышает эффективность научных исследований.

Показатель точности опыта

Показатель точности опыта (иначе – показатель точности определения среднего значения) выражает величину ошибки среднего значения в процентах от самого среднего. Точность считается удовлетворительной, если величина показателя не превышает 5%, а при бóльших значениях рекомендуется увеличить число наблюдений. Иногда величину показателя точности можно уменьшить, если повысить точность измерений параметров объектов опыта.

Показатель точности опыта вычисляется по формуле:

$$\Re = \frac{m}{\overline{x}} \cdot 100\%,$$

где m – стандартная ошибка,

$\overline{x}$ – выборочное среднее значение

Обычно показатель точности определения среднего значения – это именно то, что имеют в виду исследователи в медико-биологических науках, указывая в публикациях после величины среднего через запятую (иногда называя это достоверностью). Хотя это определение в данном случае не совсем верно, но оно используется традиционно.

на начало

Для того чтобы другим было понятно, что именно имеет в виду исследователь, в работе следует расшифровывать абсолютно все используемые математические обозначения и аббревиатуры, не полагаясь на то, что данные показатели общеупотребительны. Естественно, что сказанное относится и ко всем применяемым в исследовании статистическим показателям.

Полигон частот,
Полигон частостей

Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2), …, (x_M, n_M);

полигоном частостей – с координатами (x_1, p_1^*), (x_2, p_2^*), …, (x_M, p_M^*).

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

Порядковые данные
(ordinal data)

– это такие показатели, которые могут быть расположены в естественном порядке (ранжированы), например, от малого до большого или от хорошего до плохого.

на начало

Размер интервала между такими категориями не может быть выражен количественно (например: стадии болезни; оценки – "высокий, средний, низкий" или "отсутствует, слабый, умеренный, тяжелый").

Размах выборки
Размах вариации

Размах выборки (размах вариации, амплитуда ряда) характеризует степень разброса данных в абсолютных числах. Выборочный размах – это разность между максимумом и минимумом вариант выборки.

Вычисление размаха количественной выборки производится по формуле:

$$R = x_{max} - x_{min}$$

где x_{max} – значение максимальной варианты,

x_{min} – значение минимальной варианты выборки.

Функции MS Excel: МИН(данные); МАКС(данные).

Функция Calc (Open Office): =MIN(данные) и =MAX(данные).

Ранжирование

Операция расположения значений случайной величины (признака) по неубыванию называется ранжированием статистических данных. Полученная таким образом последовательность x_1, x_2,..., x_n значений случайной величины X (где $x_1 \le x_2 ... \le x_n$ где x_1= min$\{x_i\}$, ..., x_n=max$\{x_i\}$ называется вариационным рядом.

на начало

Ранжирования правила

1. Меньшему значению начисляется меньший ранг. Наибольшему значению начисляется ранг, соответствующий количеству ранжируемых значений (если *n=10,* то наибольшее значение получит ранг 10). Исключения оговорены в пункте 2.

2. Если несколько значений равны, им начисляется ранг, представляющий собой среднее значение из тех рангов, которые они получили бы, если бы не были равны. Например, если в упорядоченной
по возрастанию выборке три следующие за минимальным значением величины равны, то каждой из них присваивается ранг 3

$$\frac{2+3+4}{3} = \frac{9}{3} = 3.$$

3. Общая сумма рангов должна совпадать с расчетной, которая определяется по формуле

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

где n – общее количество ранжируемых значений. Несовпадение реальной и расчетной сумм рангов будет свидетельствовать об ошибке, допущенной при начислении рангов или их суммировании.

Расчет функциями MS Excel: =(n+1-РАНГ(число; данные; 0) +РАНГ(число; данные;1))/2, где <число> – значение, для которого рассчитывается ранг, <данные> – набор ранжируемых данных.

Для Calc (Open Office): =(n+1-RANK(число;данные;0) +RANK(число; данные;1))/2

на начало

Распределения функция эмпирическая (статистическая)

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 ($x \in R$).

Эмпирическая функция распределения $F_n^*(x)$ является *оценкой* вероятности события $\{X < x\}$, т.е. оценкой теоретической функции распределения $F(x)$ случайной величины X.

Регрессия

Функция, оценивающая характер связи между переменными величинами. Парная регрессия характеризует связь между двумя признаками: результативным и факторным. Аналитически связь между ними может описываться линейным уравнением (линейная регрессия) или нелинейно – например, гиперболой.

на начало

Регрессия линейная
linear regression

(лат.: linea – линия, черта; regressio – обратное движение, возвращение).

Регрессия, аппроксимируемая (приближенно описывающаяся) линейной функцией $y = kx + b$. (•)

Угловой коэффициент k называется коэффициентом регрессии Y на X.

Если уравнение (•) отыскивается по выборочным данным, то оно называется выборочным уравнением регрессии. Соответственно, k – выборочный коэффициент регрессии Y на X, а b – выборочный свободный член уравнения. Коэффициент регрессии измеряет вариацию Y, приходящуюся на единицу вариации X. Параметры уравнения регрессии (коэффициенты k и b) находятся методом наименьших квадратов.

Репрезентативность (представительность) выборки

Выборка является *репрезентативной* (или *представительной),* если она достаточно полно представлять изучаемые признаки генеральной совокупности. Условием обеспечения репрезентативности выборки является, согласно закону больших чисел, соблюдение случайности отбора, т.е. все объекты генеральной совокупности должны иметь равные вероятности попасть в выборку.

на начало

Анализ репрезентативности выборки особенно важен на начальном этапе исследований, когда численность генеральной совокупности неизвестна, но известны некоторые параметры опыта, позволяющие оценить репрезентативность.

В этом случае достаточная численность выборки определяется по формуле

$$N = \left(\frac{t_\infty \sigma}{\Delta}\right)^2,$$

где t_∞ – значение обратной функции распределения Стьюдента с "числом степеней свободы" для заданной стандартной доверительной вероятности,

σ – стандартное отклонение, рассчитанное по выборке,

Δ – заданная абсолютная погрешность определения среднего арифметического значения, введенная в именованных числах, т.е. в тех же единицах измерения, что и варианты выборки.

Например, при подсчете количества неделимых объектов исследования (например, количество листьев) абсолютная погрешность может быть установлена равной 1.

Рошера критерий

См. *Томпсона правило*

на начало

Случайная величина

Переменная, которая в результате испытания в зависимости от случая принимает одно из множества возможных значений (заранее неизвестное). Более строго, случайная величина *X* определяется как функция, заданная на множестве элементарных исходов (или в пространстве элементарных событий).

Непрерывная случайная величина – случайная величина, функция распределения которой непрерывна в любой точке и дифференцируема всюду, кроме отдельных точек; множество возможных значений случайной величины бесконечно или несчетно.

Смирнова (Колмогорова-Смирнова) двухвыборочный критерий однородности

Двухвыборочным критерием однородности Смирнова (Колмогорова-Смирнова) 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.

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 $y_1, y_2, \dots, y_m$

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

на начало

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

Событие

Результат испытания, который регистрируется как факт.

Среднее значение выборочное, среднеарифметическое выборочное

Выборочное среднее значение – наиболее часто применяемый статистический показатель, характеризующий середину эмпирической совокупности.

Вычисление среднего значения выборки производится по формуле:

$$\overline{x} = \frac{1}{n}\sum_i x_i$$

где n – численность выборки,

x_i – значения вариант выборки.

Если все варианты разнесены по М классам, каждый из которых характеризуется определенным значением вариант и частотой f, то среднее можно вычислить по следующей

на начало

формуле

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{M} f_j x_j$$

где f_j – частота класса,

x_j – значение класса.

Строго говоря, вычисление среднего значения законно только для таких эмпирических выборок, которые не противоречат гипотезе о нормальности статистического распределения.

Оценку ошибки $\overline{x}$ среднего значения см. в *стандартная ошибка по выборке*.

Функция MS Excel: СРЗНАЧ(данные);

Функция Calc (Open Office): AVERAGE(данные).

Среднее отклонение выборочное

Выборочное среднее отклонение (выборочная оценка среднего отклонения), подобно стандартному отклонению, характеризует разброс эмпирической выборки относительно среднего значения и вычисляется по формуле

$$\overline{d} = \frac{1}{n}\sum_{i} |x_i - \overline{x}|,$$

n – численность выборки,

x_i – значения вариант выборки,

$\overline{x}$ – выборочное среднее значение.

на начало

Среднее отклонение отражает так называемый модульный подход к вычислению меры отклонения между величинами в противоположность тому, что стандартное отклонение отражает квадратический подход. Подобный выбор возникает перед исследователем не только в описательном статистическом анализе, а и во многих других областях математики. До сих пор квадратический подход находит большее применение из-за удобства дифференцирования квадратического функционала.

Функция MS Excel: СРОТКЛ(данные);

Функция Calc (Open Office): AVEDEV(данные).

Стандартная ошибка по выборке

Стандартная ошибка $\overline{x}$ среднего значения определяется по формуле

$$m = \frac{\sigma}{\sqrt{n}},$$

где σ – стандартное отклонение,

n – объем выборки.

При описании результатов экспериментального исследования в медико-биологических науках стандартную ошибку принято обозначать символом m. Обычно используется понятная большинству исследователей традиционная запись, характеризующая *среднее значение* и его стандартную ошибку, в виде $\overline{x} \pm m$.

Замечание. Статистическая ошибка не имеет ничего общего с ошибкой точности. Смысл m – это, насколько исследователь ошибается, принимая данные выборочной совокупности за генеральную.

на начало

Стандартного отклонения по выборке ошибка

Ошибка стандартного отклонения вычисляется по формуле:

$$m_s = \frac{\sigma}{\sqrt{2n}}.$$

См. замечание к *стандартная ошибка*.

Стандартное отклонение по выборке

См. *дисперсия выборки*.

Статистика описательная

Эмпирические (опытные, экспериментальные) выборки (совокупности) состоят из отдельных вариант (элементов), которые объединены общностью некоторых свойств (признаков, переменных). Выборки могут быть получены в результате медико-биологического или технического эксперимента, научного опыта, социологического опроса и т.п. Источник появления выборок для статистического анализа значения не имеет. Единственное требование к анализируемым данным определяется используемыми методами расчета. Они применимы только к количественным выборкам, т.е. к таким эмпирическим выборкам, варианты которых измерены в количественной шкале. Если варианты выборок измерены в порядковой или номинальной шкале, следует применять иные методы расчета описательной статистики.

на начало

Количество вариант совокупности в источниках называют по-разному. Так, если речь идет об эмпирической выборке, количество ее элементов может называться численностью, величиной или размером. Термин "размерность" употреблять в значении "численность" не следует, т.к. он зарезервирован для описания так называемых многомерных совокупностей. Традиционными в отечественной статистической литературе являются термины "выборка", "варианта" и "численность", поэтому по возможности следует придерживаться их.

Как правило, рассчитываются следующие выборочные точечные и интервальные статистические показатели описательной статистики:

- показатели положения: среднее значение и его стандартная ошибка, доверительный интервал, медиана;
- показатели разброса (рассеяния): стандартное отклонение, среднее отклонение, размах, коэффициент вариации, средняя разность Джини, межквартильный размах;
- показатели формы распределения: коэффициент асимметрии, эксцесс.

Кроме перечисленных показателей, рассчитывается *достаточная численность выборки* из анализа заданных и рассчитанных выборочных показателей.

на начало

Степень(и) свободы

Для того чтобы свести к минимуму ошибки, в таблицах критических значений статистических критериев в общем количестве данных не учитывают те, которые можно вывести методом дедукции. Оставшиеся данные составляют так называемое число степеней свободы, т. е. то число данных из выборки, значения которых могут быть случайными.

Так, если среднее трех данных равно 2, то первые два из них могут принимать любые значения, и если они определены, то третье значение фактически становится известным. Если, например, значение первого данного равно 2, а второго – 1, то третье может быть равным только 3. Таким образом, в такой выборке имеются только две степени свободы. В общем случае для выборки в n данных существует $df=n-1$ степень свободы. Если имеются две независимые выборки, то число степеней свободы для первой из них составляет (n_1-1), а для второй – (n_2-1). А поскольку при определении достоверности разницы между ними опираются на анализ каждой выборки, число степеней свободы, по которому нужно будет находить критическое значение t, будет равно $(n_1+n_2)-2$.

Если же речь идет о двух зависимых выборках, то в основе расчета лежит вычисление суммы разностей, полученных для каждой из n пар данных (т.е., например, разностей между результатами до и после воздействия на одного и того же испытуемого). Поскольку одну (любую) из этих разностей можно вычислить, зная остальные разности и их сумму, число степеней свободы для определения того же критерия t будет равно $n-1$.

на начало

Sturgess'a правило

Используется при вычислении числа классов для выборок умеренной численности. Правило Sturgess'a выражается формулой

$$M = 1 + \log_2 n \quad или \quad M \approx 1 + 1{,}44 \cdot \ln n \quad или \quad M \approx 1 + 3{,}32 \cdot \lg n,$$

где M – число классов, n – численность выборочной совокупности.

См. также *группировка*.

Стьюдента критерий для независимых выборок для равных дисперсий (гомоскедастический тест) (two-group unpaired-test)

Критерий Стьюдента для независимых выборок (two-group unpaired-test) представляет собой обобщение критерия Стьюдента на случай двух эмпирических выборок. Он предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае равных (неизвестных) дисперсий. Поэтому перед применением теста рекомендуется проверить нулевую гипотезу о равенстве дисперсий сравниваемых совокупностей с помощью критерия Фишера.

на начало

Вычисление статистики критерия производится по формуле

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

где $\overline{x}$ *и* $\overline{y}$ – средние значения совокупностей,

n_x и n_y – численности совокупностей,

Статистика критерия подчиняется t-распределению с числом степеней свободы $df = (n_x + n_y - 2)$.

Нулевая гипотеза отвергается в том случае, когда значение достигнутого уровня статистической значимости P для вычисленного t-критерия окажется меньше заданного критического уровня. При исследовании сравниваются модуль рассчитанной величины t-критерия Стьюдента с критическим для числа степеней свободы $df = (n_x + n_y - 2$. Если эта величина больше критического значения, то нулевая гипотеза отвергается.

Функции MS Excel: TTECT(данные 1ой выборки; данные 2ой выборки;2;2) {функция TTEST из Calc (Open Office)} возвращают вероятность, соответствующую критерию Стьюдента выборок с равными дисперсиями (гомоскедастичные данные) для двустороннего распределения.

Этот же тест совпадает с так называемым "двухвыборочным t-тестом с одинаковыми дисперсиями", представленным в пакете "Анализ данных" Microsoft Excel.

на начало

Стьюдента критерий для независимых выборок для разных 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.

Когда дисперсии неизвестны и их равенство не предполагается ($\sigma_x^2 \neq \sigma_y^2$), используется так называемая t статистика критерия Беренса-Фишера

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

Данное распределение близко к распределению Стьюдента с "пересчитанным" числом степеней свободы df.

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

Здесь n_x и n_y – численности совокупностей,

$\overline{x}$, $\overline{y}$ – выборочные средние значения, σ_x^2, σ_y^2 – дисперсии.

на начало

Стьюдента критерий парный
two-group paired-test

Критерий Стьюдента для связанных выборок (парного критерия Стьюдента, two-group paired-test) предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае неравных неизвестных дисперсий. В источниках критерий может называться одновыборочным критерием Стьюдента. Это название вызвано тем обстоятельством, что на самом деле, исходя из схемы расчета, анализируется действительно одна выборка, составленная из попарных разностей вариант исходных связанных выборок. Понятно, что в данном случае проверяется нулевая гипотеза о равенстве среднего значения полученной выборки известному значению, а именно – нулю.

Используется следующая формула вычисления t-критерия Стьюдента:

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

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 анализируется альтернативная.

на начало

Стьюдента критерий равенства среднего значения

Критерий Стьюдента предназначен для проверки нулевой гипотезы о равенстве среднего значения выборочной совокупности заданному математическому ожиданию. Вычисление производится по формуле

$$t = \frac{|x - \lambda_0|\sqrt{n}}{\sigma},$$

где $\overline{x}$ – среднее значение совокупности,

λ_0 – заданное математическое ожидание,

n – численность совокупности,

σ^2 – оценка выборочной дисперсии.

Статистика критерия Стьюдента подчиняется t-распределению с числом степеней свободы df=n-1.

При практических вычислениях в средах Excel (Microsoft Office) или Calc (Open Office) используется функция

СТЬЮДРАСПОБР(вероятность; степени_свободы)

TINV(вероятность; степени_свободы),

возвращающие t-значение распределения Стьюдента как функцию вероятности и числа степеней свободы. Параметры суть следующее.

Вероятность – вероятность, соответствующая двустороннему распределению Стьюдента.

Степени_свободы – число степеней свободы, характеризующее распределение.

на начало

Одностороннее *t*-значение может быть получено при замене аргумента вероятность на 2×вероятность. Для вероятности 0,05 и числа степеней свободы равного 10 двухстороннее значение вычисляется СТЬЮДРАСПОБР(0,05; 10) и равно 2,28139. Одностороннее значение для той же вероятности и числа степеней свободы может быть вычислено формулой СТЬЮДРАСПОБР(2*0,05;10) и равняется 1,812462.

Томпсона правило, Рошера критерий

В правиле Томпсона (критерий Рошера) для исключения выбросов используется статистика

$$t = \frac{|x - \overline{x}|}{\sigma},$$

i=1,…,n,

где x_i – результаты наблюдений,

$\overline{x}$ – выборочное среднее значение,

σ – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии.

Величина статистики критерия сравнивается с критическим значением

$$T = \sqrt{\frac{(n-1)t^2_{\left(1-\frac{\alpha}{2}\right),(n-2)}}{n-2+t^2_{\left(1-\frac{\alpha}{2}\right),(n-2)}}}$$

где $t_{\left(1-\frac{\alpha}{2}\right),(n-2)}$ – значение обратной функции *t*-распределения с параметрами (1-α/2) и (n-2),

α– заданный уровень значимости, обычно 0,05.

При величине статистики, большей критического значения, наблюдение исключается. Процедура повторяется для каждого наблюдения.

на начало

Фишера F-критерий, (критерий Фишера-Снедекора)

F-критерий Фишера применяют для сравнения дисперсий двух выборочных совокупностей. Критерий часто называют просто статистикой Фишера. Критерий Фишера чувствителен к нарушению предположения о нормальности.

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 (предложенная Снедекором) 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. Критические значения можно вычислить по нижеприведенным функциям.

Функции MS Excel: FРАСПОБР(α; $(n_1\text{-}1)$; $(n_2\text{-}1)$), α – заданный уровень значимости;
Функция Calc (Open Office): FINV(α; $(n_1\text{-}1)$; $(n_2\text{-}1)$).

Функция распределения случайной величины интегральная integral distribution function

(лат.: integer – нетронутый, незатронутый, невредимый, целый; integratio – восстановление).

Функция $F(x)$, выражающая для каждого x вероятность того, что случайная величина X примет значение, меньшее x, то есть $F(x) = P(X < x)$.

на начало

Распределение вероятностей дискретной случайной величины может быть задано перечнем всех ее возможных значений и их вероятностей. Такой способ задания неприменим для непрерывных случайных величин. Общим способом задания распределений любых типов случайных величин является интегральная функция распределения. Пусть x – действительное число. Вероятность события, состоящего в том, что случайная величина X примет значение, меньшее x то есть вероятность события $X < x$ обозначим через $F(x)$. Интегральной функцией распределения называют функцию $F(x)$, определяющую для каждого значения x вероятность того, что случайная величина X примет значение, меньшее x, то есть $F(x) = P(X < x)$. Геометрически это равенство можно истолковать так: $F(x)$ есть вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки x.

Интегральная функция распределения имеет следующие свойства.

1. Значения интегральной функции принадлежат отрезку (0,1): $0 \le F(x) \le 1$. Следовательно, график интегральной функции распределения расположен в полосе, ограниченной прямыми $y = 0, y = 1$.

2. $F(x)$– неубывающая функция, то есть $F(x_2) > F(x_1)$, если $x_2 > x_1$. Следовательно, при возрастании x в интервале (a, b), в котором заключены все возможные значения случайной величины, график интегральной функции распределения поднимается вверх.

3. Если возможные значения случайной величины принадлежат интервалу (a, b), то $F(x) = 0$ при $x < a$, $F(x) = 1$ при $x > b$. Для дискретной случайной величины график интегральной функции распределения имеет ступенчатый вид.

на начало

χ^2 критерий Пирсона
Chi-square test

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

Для нахождения χ^2 обычно используется следующая формула:

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

где $x_{i\,теор}$ – теоретически ожидаемое число или показатель для данного класса;
$x_{i\,экс}$ – фактически наблюдаемое;
M – количество классов (интервалов).

Для повышения точности при сопоставлении распределений признаков, которые принимают всего 2 значения, рекомендуется вводить поправку на непрерывность Yates'a.

С помощью метода χ^2 можно сопоставлять два эмпирических распределения.
При сравнении эмпирического распределения с гипотетическим теоретическим, содержащим s параметров, число степеней свободы определяется соотношением df = M-1-s. Для нормального распределения, имеющего два параметра (математическое ожидание и дисперсию), s=2. При $\chi^2_{экс} > \chi^2_{крит} = \chi^2_{1-\alpha}$ (критического значения для уровня значимости α и числа степеней свободы $df = r - 3$) гипотеза о согласии теоретического и экспериментального (эмпирического) распределения отвергается.

Когда χ^2 используется как критерий согласия (χ^2-goodness-of-fit test) (для таблицы r столбцов и s строк), статистика χ^2 фактически есть мера согласия (близости) наблюдаемых и ожидаемых частот.

на начало

Критическая область этой статистики – это область, лежащая выше p%-ной критической точки распределения χ^2 с $df = (r-1)(s-1)$ степеням свободы. Иначе говоря, если на принятом уровне значимости $\chi^2_{\text{экс}} > \chi^2_{\text{крит}} = \chi^2_{1-\alpha}$, то нулевая гипотеза H_0 отклоняется в пользу альтернативной H_1.

Функции MS Excel: =ХИ2ОБР(α; *df*), α – заданный уровень значимости.
Функция Calc (Open Office): CHIINV(α; *df*).

Частот (частостей) гистограмма

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.

Частота,
Частость

Числа n_i, показывающие, сколько раз встречаются варианты x_i в ряде наблюдений, называются *частотами,* а отношение их к объему выборки – *частостями* или *относительными частотами* (p_i^*), т.е.

$$p_i^* = \frac{n_i}{n}, \quad где \ n = \sum n_i .$$

на начало

Шансы (odds)

– отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Шансы и вероятности содержат одну и ту же информацию, но по-разному выражают ее.

Если вероятность того, что событие произойдет, обозначить *P*, то шансы этого события будут равны *P*/(1-*P*). Например, если вероятность выздоровления 0,3, то шансы выздороветь равны 0,3/(1-0,3) = 0,43. Для некоторых оценок шансы удобнее использовать, чем вероятности.

Шкалирование

(Лат.: scala – лестница).
Шкалирование – это операция упорядочивания исходных эмпирических данных путем перевода их в шкальные оценки. Шкала дает возможность упорядочить наблюдаемые явления, при этом каждое из них получает количественную оценку (квантифицируется). Шкалирование помогает определить низшую и высшую ступени исследуемого явления.
Шкалирование – процесс разработки шкалы для оценки характеристик объекта.
Объект может быть конкретным реальным или абстрактным, наблюдаемым или ненаблюдаемым. Наблюдаемые объекты характеризуются понятиями, концептами. Ненаблюдаемые объекты характеризуются конструктами.
Число выбранных характеристик объекта (размерность модели) может назначаться исследователем. При назначении размерности модели объекта исследования равной единице, осуществляют одномерное шкалирование.

на начало

При назначении размерности модели объекта исследования больше единицы осуществляют многомерное шкалирование. При этом используют либо аналитический, либо системный подход. В первом случае создают совокупность согласованных одномерных шкал по числу выбранных характеристик, с тем, чтобы в последующем обоснованно суммировать полученные с помощью шкал результаты.

Во втором случае используют специальные методы многомерного шкалирования относительно всех характеристик сразу.

По существу процесс шкалирования представляет собой установление по определенным правилам соответствия между объектом с одной стороны и переменными величинами, характеризующими объект – с другой.

Структуру измерения можно представить следующей иерархией атрибутов (понятий):

→ объект исследования с характеристиками и отношениями между ними
→ модель объекта (детерминистская или вероятностная)
→ характеристики модели объекта (их выбор и отношения зависят от вида модели, а их число соответствует выбранной размерности)
→ показатели, соответствующие характеристикам модели
→ случайные или неслучайные переменные величины, отображающие показатели
→ меры (пределы существования переменной величины в данном качестве)
→ шкалы (инструмент для оценки значений меры, переменной, показателя, характеристики, объекта и для отображения этих значений)
→ значения шкалы.

на начало

В соответствии с этой иерархией можно выделить следующие основные этапы шкалирования.

(1) Определение цели исследования, выбор либо вероятностной, либо детерминистской методологии исследования и определение цели шкалирования.

(2) Описание максимально возможного числа характеристик объекта исследования.

(3) Обоснованный выбор минимального и достаточного (для достижения цели исследования) числа характеристик объекта из ранее описанных и назначение размерности шкалы.

(4) Детальное описание выбранных для измерения характеристик.

(5) Определение показателей, соответствующих характеристикам объекта, выбранным для измерения.

(6) Обозначение показателей переменными и назначение их мер.

(7) Определение предварительных правил получения значений оценок переменных в процессе измерения (тестирования).

(8) Разработка совокупности предполагаемых заданий для тестирования.

(9) Оценка эффективности предполагаемых заданий и выбор из них окончательных заданий шкалы (статистическое исследование, экспертная оценка степени соответствия заданий изучаемому объекту, пробное тестирование и т.п.).

(10) Оформление шкалы и разработка окончательных правил ее применения.

на начало

Эксцесс

Эксцесс характеризует форму статистического распределения. Если эксцесс больше нуля, то форма кривой распределения островершинная по сравнению с кривой плотности нормального распределения. Если эксцесс меньше нуля, то форма кривой распределения плосковершинная по сравнению с кривой плотности нормального распределения. Эксцесс выборочной совокупности можно вычислить по формуле:

$$E = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

где n – численность выборки, x_i, i=1,2…,n – значения вариант выборки,

$\overline{x}$ – выборочное среднее значение, σ – стандартное отклонение.

Дисперсия эксцесса можно определить формулой $D_E = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$.

Эксцесс находит применение, в частности, при исследовании формы распределения выборки.

Функции MS Excel: ЭКСЦЕСС(данные); Calc (Open Office): KURT(данные).

на начало

"Ящик с усами" правило

Правило "ящик с усами" получило название от типа соответствующего графика, используемого для наглядного представления разброса эмпирических данных с нанесенными значениями медианы и квартилей.

Порядок вычисления следующий:

1. Определяются выборочные значения межквартильного размаха R_μ и медианы μ.
2. Выборочные значения, меньшие (μ-1,5R_μ) и большие (μ+1,5R_μ), называются мягкими (подозрительными) выбросами.
3. Выборочные значения, меньшие (μ-3R_μ) и большие (μ+3R_μ), называются экстремальными выбросами и должны быть исключены.

Критерий удобен для автоматической идентификации любого числа экстремально малых и больших значений выборки. Критерий достаточно популярен, однако его рекомендуется применять только в случае, если численность выборки велика.

на начало