Математическая статистика

Анохин Сергей, D-053
Кокорева Ирина, D-063
Руководитель: Попова И. Г.

Северск-2005

Содержание

Введение

Математическая статистика – наука, занимающаяся методами обработки экспериментальных данных, полученных в результате наблюдений над случайными явлениями. При этом следующие *задачи*:

- описание явлений – упорядочить статистический материал, представить в удобном для экспериментатора виде (таблица, график, диаграмма);
- анализ и прогноз – приближенная оценка интересующих числовых событий (средняя, дисперсия) и погрешности этих величин;
- выработка оптимальных решений – в результате возникает задача проверки правдоподобности гипотез, решением которой является принятие или неприятие выдвинутой гипотезы.

Математическая статистика при решении своих задач опирается на размышляющий, оценивающий составляющий, аппарат экспериментатора.

Генеральная совокупность и выборка

Полный набор всех возможных значений дискретной СВ называется **генеральной совокупностью.** N – объем совокупности.

Однако в реальности провести сплошное обследование нецелесообразно и невозможно. На практике ограничиваются выборкой.

Часть генеральной совокупности из ***n*** *элементов, отобранных случайным образом называется* **выборкой.** $n \le N$

Выборка с объемом < 30 называется *выборкой малого объема.*

Способы отбора

1. Отбор, не требующий расчленения:
- *простой, бесповторный*
- *с повторениями*
2. Отбор, при котором вся генеральная совокупность делится на части
- *механический*
- *типический*
- *серийный*

Простой – отбор, при котором объекты извлекаются из совокупности по одному.

Механический – генеральная совокупность «механически» делится на группы. Выборка производится с каждой из групп.

Типический – объекты выбирают не из всей совокупности, а из каждой ее типической части.

Серийный – объекты отбираются не по одному, а сериями, которую подвергают сплошному обследованию. **Примеры**

Для того, чтобы по данным выборки можно было судить об интересующем нас признаке генеральной совокупности, нужно чтобы выборка правильно представляла пропорции генеральной совокупности, т.е. выборка должна быть *репрезентативной (представительной)*. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если *каждый объект отобран случайно и если все объекты имеют одинаковую вероятность попасть в выборку.*

Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка, причем значение x_1 встречалось n_1 и т.д., $x_k - n_k$. Наблюдаемые значения x называется *вариантами,* а последовательность вариант, записанных в возрастающем порядке – *вариационным рядом.* Число наблюдений называется *частотами.*

Относительная частота наблюдений – *отношение числа наблюдений к объему выборки.* $W_i=n_i/n$

Статистическим распределением *называют перечень вариант и соответствующих им частот или относительных частот.*

Пример

Для визуальной оценки выборочного распределения производится группировка данных. Для этого:

- располагают значения x_i по возрастанию;
- весь интервал разбивают на k последовательных непересекающихся интервалов;
- подсчитывают числа n_1 – количество попавших значений x_i в каждый интервал.

Такая таблица называется ***группированным статистическим рядом.***

$x_j \div x_{j+1}$	$x_1 \div x_2$	$x_2 \div x_3$	. . .	$x_{k-1} \div x_k$
n_j	n_1	n_2	. . .	n_{kj}

Эмпирическая функция распределения

Эмпирической функцией распределения *(функция распределения выборки) называется* $F^*(x)$*, определяющую для каждого значения* ***x*** *относительную частоту события* $X<x$.
$F^*(x)>n_x/n;\ n_x$ – число вариант, меньше $\boldsymbol{x}$, n – объем выборки.

Свойства.

- значения $F^*(x)\in[0;1]$
- $F^*(x)$ – функция неубывающая: $F^*(x_2)> F^*(x_1)$, если $x_2> x_1$
- если x_1 – наименьшая варианта, $F^*(x_1)=0$

  если x_k – наибольшая, то $F^*(x_1)=1$.

В отличие от эмпирической функции, функцию $F(x)$ генеральной совокупности называется *теоретической*. Различия между ними состоят в том, что $F(x)$ определяет вероятность события $X<x$, а $F(x)$ – относительную частоту.

Наглядным изображением статистического ряда распрделения служат *полигон* и *гистограмма.*

Полигон – *ломаная линия, соединяющая точки* $\boldsymbol{(x_i;n_i)}$.

Гистограмма – *ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат интервалы, длиной* $\boldsymbol{n}$ *, а высотой – величины* $\boldsymbol{n_i/n}$.

Если гистограмма является гистограммой частот, то ее площадь равна сумме всех частот, т.е. объему выборки.

Если гистограмма является гистограммой относительных частот, то ее площадь равна сумме всех относительных частот. $\left.\frac{n_1}{n}+\frac{n_2}{n}+\ldots+\frac{n_i}{n}\right/N$

Статистические оценки параметров распределения

- Точечные оценки
- Интервальные оценки
- Точность и надежность
- Доверительный интервал для мат.ожидания
- Доверительный интервал для оценки дисперсии

Для того, чтобы статистические оценки давали хорошее приближение оценивающих параметров, они должны удовлетворять условиям:

- объем выборки должен быть достаточным для оценивания
- оценка интересующего нас параметра есть случайная величина.

Статистические оценки:

- *Несмещенные* – есть оценка мат.ожидания, которая равна оценивающему параметру;
- *Смещенные* – оценка $M(x) \neq$ оценивающему параметру;
- *Эффективные* – оценка, имеющая при заданном объеме выборки ***n*** наименьшую дисперсию;
- *Состоятельные* – оценка, стремящаяся при $\boldsymbol{n \to 0}$ по вероятности к оцениваемому параметру.

Точечные оценки

Точечной *называют оценку, определяющую одним числом.*

Пусть требуется изучить количественный признак генеральной совокупности. Допустим, удалось установить, какое имеется распределение. Тогда возникает задача оценки параметров данного распределения.

Пример

Однако чаще всего экспериментатору не известен вид распределения, т.к. он обладает только данными выборки и тогда для оценки параметров нужно найти зависимость этих параметров от наблюдаемых величин.

Генеральная и выборочная средняя
Генеральная и выборочная дисперсии

Генеральная и выборочная средняя

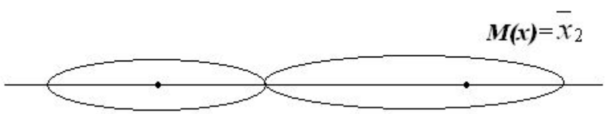
Генеральная средняя – *среднее арифметическое значений генеральной совокупности* $\overline{x}_2$

$$\overline{x}_2 = \frac{1}{N}\sum_{i=1}^{n} x_i$$

$$\overline{x}_2 = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots}{N}$$ с повторениями

Генеральная средняя есть среднее взвешенное значений генеральной совокупности с их весами, равными соответствующим частотам.

Если рассматривать ***x*** генеральной совокупности как СВ, то $M(x) = \overline{x}_2$



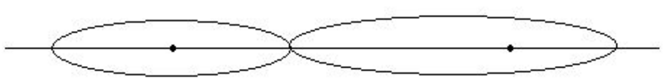
Выборочная средняя – *среднее арифметическое значений выборки.*

Пусть имеется выборка объема ***n***. Тогда выборочная средняя равна: $\overline{x}_в = \frac{1}{N}\sum_{i=1}^{n} x_i$

Выборочная средняя по данным одной выборки есть определенное число.

Если извлекать другие выборки такого же объема из генеральной совокупности, то выборочная средняя меняется от выборки к выборке.

Выборочная средняя есть *несмещенная* оценка генеральной средней.



При увеличении объема выборки ***n*** выборочная средняя стремится к генеральной средней.

Генеральная и выборочная дисперсии

Генеральной дисперсией *называют среднее арифметическое квадратов отклонений значений генеральной совокупности от их среднего значения.*

$$D_г = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_г)^2; \qquad D_г = \sum_{i=1}^{N} N_i (x_i - \bar{x}_г)^2 \Big/ N; \qquad D_г = \sum_{i=1}^{N} N(x_i - \bar{x}_г)^2 \cdot p_i;$$

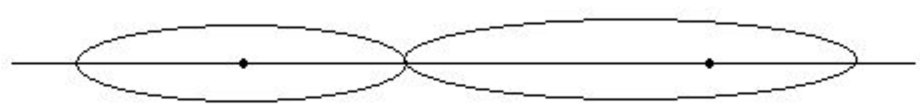
Кроме дисперсий для характеристики рассеивания значений генеральной совокупности вокруг своего среднего пользуются другой характеристикой – *средним квадратическим отклонением.*

Выборочной дисперсией *называют среднее арифметическое квадратов отклонений наблюдаемых значений выборки от их среднего значения.*

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_г)^2 = S^2$$

$S^2 = \sum_{i=1}^{n} n_i (x_i - \bar{x}_г)^2 \Big/ n$ – с повторениями

Для оценки рассеивания выборки служит выборочное среднеквадратическое отклонение.



Интервальные оценки

Интервальной оценкой *называют оценку, определяющуюся двумя концами интервала.*

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, что приводит к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться другими оценками.

Интервальные оценки позволяют определить точность и надежность оценок.

Точность и надежность

Пусть найденная по данной выборке статистическая характеристика θ^* служит оценкой неизвестного параметра θ генеральной совокупности. Будем считать θ постоянным числом. θ^* будет тем точнее определять параметр θ, чем меньше абсолютная величина разности $|\theta\theta^*|<\varepsilon$. Чем меньше ε, тем точнее оценка.

Однако статистические методы не позволяют категорически утверждать, что θ^* удовлетворяет условию $|\theta\theta^*|<\varepsilon$, а можно лишь говорить о вероятности, с которой это неравенство осуществляется: $P(|\theta\theta^*|<\varepsilon)=\beta$

Надежностью *(доверительной вероятностью) оценки θ по θ^* называется вероятность β, с которой осуществляется неравенство $|\theta\theta^*|<\varepsilon$.*

Обычно надежность оказывается заранее заданным числом, близким к 1. Наиболее частые значения β: 0,95; 0,98; 0,99; 0,999.

Соотношение $P(|\theta-\theta^*|<\varepsilon)=\beta$ означает вероятность того, что интервал $(\theta^*-\ \varepsilon;\ \theta^*+\varepsilon)$ заключает в себя (покрывает) неизвестный параметр θ, равна доверительной вероятности β.

Доверительным интервалом *называется интервал $(\theta^*-\varepsilon;\ \theta^*+\varepsilon)$, покрывающий неизвестный параметр θ с надежностью β.*

Иногда вместо доверительной вероятности β используют обратную величину – уровень значимости $\alpha = 1-\beta$. Если β – вероятность, что оцениваемый параметр попадет в интервал, α – вероятность, что не попадет. В статистических таблицах указывается именно α.

Доверительный интервал для мат. ожидания

Рассмотрим нахождение доверительного интервала для M(X) нормально распределенной СВ, т.е. нужно найти такой интервал, чтобы выполнялось следующее неравенство $\bar{X} - \varepsilon < M(X) < \bar{X} + \varepsilon$, т.е. данный интервал ширину 2ε. Он обладает симметрией.

Вероятность того, что $\left| X - \bar{X} \right| < \varepsilon$ определяется:

- законом нормального распределения, если известна $D(x)=\sigma^2$
- или распределения Стьюдента, если D(x) неизвестна, а подсчитана ее несмещенная оценка S^2.

Критерий Стьюдента определяется таким параметром как степень свободы $\upsilon=n-1$.

Для расчетов доверительных интервалов для M(X) используют два подхода:

- *когда D(x) известна*
- *когда D(x) неизвестна*

Расчет доверительных интервалов при известной дисперсии

Будем рассматривать выборочную среднюю как случайную величину. Примем без доказательств, что если СВ ***X*** распределена нормально, то и выборочная средняя по независимым наблюдениям распределена нормально.

$P(|X - M(x)| < \varepsilon) = 2\Phi(x) = \beta$.

Таким образом для нормального распределения $2\Phi(x) = \beta$.

Параметр ***t*** в функции Лапласа: $t = \varepsilon\sqrt{n}/\sigma$

Таким образом для отыскания границ доверительного интервала:

1. по таблицам функции Лапласа находим значение аргумента ***t***, для которого $\Phi(t) = \beta/2$.
2. зная значение ***t*** из условия $t = \varepsilon\sqrt{n}/\sigma$ находим ***ε*** (граница интервала): $\varepsilon = t\sigma/\sqrt{n}$
3. Записываем доверительный интервал: $(\bar{X} - \varepsilon; \bar{X} + \varepsilon)$

Пример

Расчет доверительных интервалов при неизвестной дисперсии

Если D(x) неизвестна, а ее несмещенная оценка S^2, то в этом случае β покрытия M(x) интервалом $(\overline{X}-\varepsilon;\overline{X}+\varepsilon)$ вычисляют по закону распределения Стьюдента со степенью свободы ***υ=n–1*** и ***α=1–β***.

Имеются таблицы, которые по заданным уровням значимости и степеням определяют значения критерия Стьюдента $t_{\alpha,\upsilon}$ по формуле $t_{\alpha,\nu}=\varepsilon\sqrt{n}/S$

Таким образом доверительный интервал для *M(x)* при неизвестной дисперсии строится в виде $(\overline{X}-\frac{t_{\alpha,\nu}\cdot S}{\sqrt{n}};\overline{X}+\frac{t_{\alpha,\nu}\cdot S}{\sqrt{n}})$

Данный интервал определяет, что с доверительной вероятностью β он покрывает истинное значение *M(x)*.

Пример

Доверительный интервал для оценки дисперсии

Доверительный интервал строится на основании того, что величина $(n–1)S^2/\sigma$ распределена по закону «хи-квадрат»(χ^2) со степенями υ=n–1.

Выборочная дисперсия D(x) и нормального распределения связаны следующим соотношением:

$\chi= (n–1)S^2/\sigma$

$\chi^2 \sigma^2= (n–1)S^2$

Для заданной доверительной вероятности β или, что тождественно, для заданного уровня значимости α=1–β. Потребуем, чтобы выполнялось следующее соотношение

$P(\chi^2_{\alpha/2,\upsilon} < \delta)= P(\chi^2_{1-\alpha/2} < \delta)= \alpha/2; \;\; \delta_2 < \chi^2_{\alpha/2,\upsilon} < \delta_2$

Из этого соотношения следует, что границы доверительного интервала в явном виде выглядят следующим образом:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}$$

Пример

Проверка статистических гипотез

Статистической гипотезой *называют, гипотезу о видах неизвестного распределения или о параметрах известного распределения.*

Проверка статистической гипотезы заключается в сопоставлении некоторых статистических показателей, вычисленным по данным выборки со значениями этих же показателей, определенными теоретически в предположении, что проверяемая гипотеза верна.

Классификация гипотез.

В результате проверки могут быть приняты два неправильных решения, т.е. допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.

Вероятность совершить ошибку первого рода принято обозначать **α**. На практике, наиболее часто используют α=0,05, это означает, что в 5 случаях из 100 имеется риск допустить ошибку первого рода, т.е. отвергнуть правильную гипотезу.

Статистический критерий, статистическая область

Сравнение двух дисперсий

Сравнение математических ожиданий

Проверка гипотезы о равенстве средних при и известных дисперсиях

Проверка гипотезы о равенстве средних при неизвестных дисперсиях

Классификация гипотез

- *Статистические, нестатистические*
- *Выдвинутая, конкурирующая.*

Выдвинутую гипотезу называют нулевой (основной) и обозначают $\mathbf{H_0}$. *Конкурирующая* гипотеза $\mathbf{H_1}$ – это гипотеза альтернативная нулевой, т.е. противоречащая основной. Пример

По количеству предположений: *простые, сложные*.

Простая – это гипотеза содержащая только одно предположение. *Сложная* – гипотеза состоящая из конечного или бесконечного числа простых гипотез.

Статистический критерий, статистическая область

Для проверки H_0, используют специально подобранную слу-чайную величину, точное или приближенное значение которой из-вестно. Эту величину обозначают через U или Z, если она распреде-лена нормально; F или υ^2 - по закону Фишера; χ^2 - по закону «хи квадрат»; T или t - по распределению Стьюдента.

Статистическим критерием *называют случайную величину служащую для проверки H_0.*

Наблюдаемым значением критерия *называют, значение критерия выраженное по данным выборки.*

После выбора определенного критерия, множество всех его возможных значений разбивается на два подмножества:

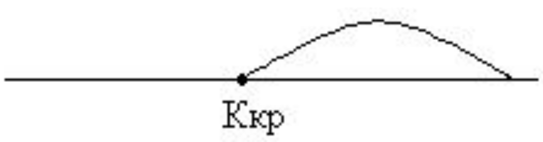
- содержит значения критерия при котором H_0 отвергается;
- содержит значения критериев при которых H_0 принимается.

Критической областью *называют, совокупность значений критерия при которых H_0 отвергается.*

Областью принятия гипотезы *(областью допустимых значений), называют совокупность критерия при которой H_0 принимают.*

Основной принцип проверки статистических гипотез: если наблюдаемое значение критерия принадлежит критической области, то гипотезу отвергают; если наблюдаемое значение критерия принадлежит области покрытия гипотезы, то гипотезу принимают.

Критическая область и область покрытия гипотез – это интервалы, следовательно существует точка которая их разделяет.



Критической точкой *(границей), называют точку определяющую критическую область от области принятия гипотез.*

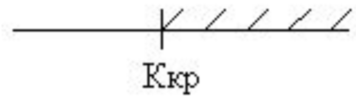
Различают:

1. Одностостороннюю критическую область

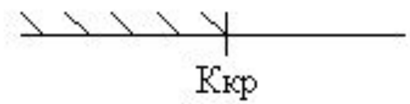
- *левостороннюю*
- *правостороннюю*

2. Двустороннюю критическую область

Правостороннюю называют критическую область определяемую неравенством **$К>К_{кр}$**



Левостороннюю называют критическую область определяемую неравенством **$К<К_{кр}$**



Двустороннюю называют критическую область определяемая двумя неравенствами **$К<К_1$ и $К>К_2$; $К_1>К_2$**

При отыскании критической области задают α (уровень значимости) и ищут критические точки исходя из требований, что критерий К примет значение лежащее в критической области, при этом вероятность такого события равна принятому уровню значимости α, т.е. для правосторонней области $P(К>К_{кр})= \alpha$; для левосторонней области $P(К<К_{кр})= \alpha$; для двусторонней области $P(К>|К_{кр}|)= \alpha/2$

Если наблюдаемое значение критерия принадлежит критичес-кой области, нулевую гипотезу отвергают, если не принадлежит, то нет оснований отвергать H_0.

Для многих критериев составлены таблицы:

- Стьюдента;
- χ^2;
- Фишера

Сравнение двух дисперсий

Рассмотрим гипотезу о параметрах нормального распределе-ния. Пусть имеется две серии опытов, регистрирующая значение некоторой случайной величины.

X: $x_1, x_2 \dots\ x_n$

Y: $y_1, y_2 \dots y_n$

Осуществим проверку нулевой гипотезы о равенстве диспер-сий <u>при неизвестных математических ожиданиях.</u>

H_0: $D_x = D_y$

Постановка задачи.

Пусть даны две случайные величины X и Y, распределенные нормально. По данным выборки объем их nx и ny подсчитаны выбо-рочные дисперсии S, Sтребуется.

Механизм проверки.

Цель работы при заданном уровне значимости α проверить ну-левую гипотезу о равенстве дисперсий.

Такая задача возникает при сравнении точности двух прибо-ров, или при сравнении различных методов измерения. Т.е. когда выборочные дисперсии отличаются, возникает вопрос значимости или не значимости это различие.

Если *различие неразличимо*, то имеет место нулевая гипотеза, т.е. приборы, например имеют одинаковую точность. А различия вы-борочных дисперсий объясняется случайными причинами.

Механизм проверки

По данным выборок значений $\boldsymbol{n_x}$ и $\boldsymbol{n_y}$, вычисляют наблюдаемое значение критерия как отношение большей дисперсии к меньшей:

$$F_{набл} = \frac{S^2_{большая}}{S^2_{меньшая}} \qquad F_{набл} = \frac{\max(S_x^2, S_y^2)}{\min(S_x^2, S_y^2)}$$

Критическая область строится в зависимости от конкурирующей гипотезы. По таблицам распределения Фишера, по заданному уровню значимости **α** и вычисленным степеням свободы υ_x, υ_y находят табличное значение критерия:

- для альтернативной гипотезы H_1: $D_x > D_y$

  $F_{кр}$ в зависимости от параметров $F_{кр}\ (\alpha, \upsilon_x, \upsilon_y)$
- для альтернативной гипотезы H_1: $D_x \neq D_y$

  $F_{кр}$ в зависимости от параметров $F_{кр}\ (\alpha/2, \upsilon_x, \upsilon_y)$

Если $F_{наб} > F_{кр}$, то H_0 отвергают.

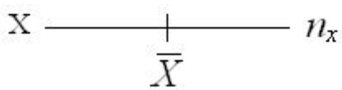
Если $F_{наб} < F_{кр}$, то нет оснований отвергать H_0, предположение о том что D_x, D_y, принимается с уровнем **α**, в 95% случаях – с доверительной вероятностью.

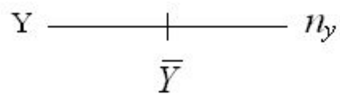
Пример

Сравнение мат.ожиданий

Для проверки гипотезы, соответствие двух выборок принад-лежности к одной и той же генеральной совокупности, рассмотрим вопрос о значимости расхождений между выборочным значением математических ожиданий. Выдвинем нулевую гипотезу о равенстве математических ожиданий.

H_0: $M_x = M_y$

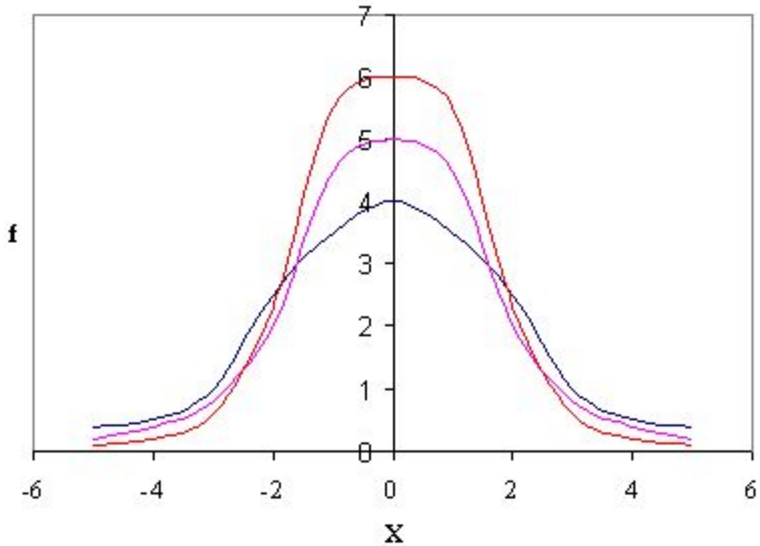




Тестирование такой гипотезы основано:

- на нормальном распределении в случае большого объема выборок (n>30), когда дисперсии считаются известными
- на распределении Стьюдента в случае малого объема выборок (n<30) когда дисперсии являются неизвестными.

Сравнительные графики плотностей распределения нормального и Стьюдента приведены на рисунке:



ı розовой линиями показано
ıределение Стьюдента,
сной – нормальное

Проверка гипотезы о равенстве средних при известных дисперсиях

Для того чтобы при заданном уровне значимости α =0.05 проверить нулевую гипотезу H_0: $M_x=M_y$ о равенстве математических ожиданий двух больших нормальных выборок с известными дисперсиями D_x и D_y, необходимо:

1. Вычислить наблюдаемое значение критерия:

$$Z_{набл} = \overline{X} - \overline{Y} \Big/ \sqrt{\frac{D_x}{n_x} + \frac{D_y}{n_y}}$$

Построить критическую область в зависимости от конкурирую-щей гипотезы:

- при конкурирующей гипотезе H_1: $M_x \neq M_y$ по таблице функции Лапласа находят критическую точку $z_{кр}$ из равенства $\Phi(z_{кр}) = (1 - \alpha)/2$.
  Если $|Z_{набл}| < z_{кр}$, то нет оснований отвергать нулевую гипотезу.
  Если $|Z_{набл}| > z_{кр}$, то нулевую гипотезу отвергают.
- при конкурирующей гипотезе H_1: $M_x > M_y$ по таблице функции Лапласа находят критическую точку zkp из равенства
  $\Phi(z_{кр}) = (1 - 2\alpha)/2$.
  Если $Z_{набл} < z_{кр}$, то нет оснований отвергать нулевую гипотезу.
  Если $Z_{набл} > z_{кр}$, то нулевую гипотезу отвергают.
- при конкурирующей гипотезе H_1: $M_x < M_y$ по таблице функции Лапласа находят «вспомогательную критическую точку» $z_{кр}$ из равенства
  $\Phi(z_{кр}) = (1 - 2\alpha)/2$.
  Если $Z_{набл} > -z_{кр}$, то нет оснований отвергать нулевую гипотезу.
  Если $Z_{набл} < -z_{кр}$, то нулевую гипотезу отвергают.

Проверка гипотезы о равенстве средних при неизвестных дисперсиях

Постановка задач: пусть генеральные совокупности распределены нормально, причем их дисперсии Dx и Dy заранее не известны. Взяты две выборки малого объема, требуется сравнить средние этих генеральных совокупностей.

Методика проверки задач: заключается в использовании критерия Стьюдента при условии, что генеральные дисперсии не известны, однако в предположении, что они равны между собой.

Такая задача возникает: если сравниваются средние размеры двух партий деталей, изготовленных на одном и том же станке. Естественно будет предположить, что дисперсии контролируемых размеров одинаковы.

Алгоритм проверки

Алгоритм проверки

1) Прежде чем сравнивать средние требуется проверить **H_0: D_x=D_y**

2) Если гипотеза подтвердилась нужно вычислить наблюдаемое значение критерия:

$$T_н = \frac{\overline{X} - \overline{Y}}{\sqrt{\nu_x \cdot S_x^2 + \nu_y \cdot S_y^2}} \cdot \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}$$

3) Строим критическую область в зависимости от конкурирующей гипотезы

а) Если **H_1: $M_x \neq M_y$** – двусторонняя критическая область строится исходя из условия чтобы вероятность попадания наблюдаемого значения критерия в эту область была равна принятому уровню значимости α взятого из таблицы Стьюдента для числа степеней свободы в верхней части таблицы, т.е. для двусторонней критической области при условии **$|T_{набл}| < t_{кр}(\alpha,\upsilon)$**, то нет основания отвергать нулевую гипотезу; если **$|T_{набл}| > t_{кр}(\alpha,\upsilon)$**, то нулевую гипотезу отвергают.

б) Если **H_1: $M_x > M_y$** строится правосторонняя критическая область, а критическую точку находят по таблице Стьюдента из нижней части.

Если **$T_{набл} < t_{кр}$**, то нет основания отвергать нулевую гипотезу .

Если **$T_{набл} > t_{кр}$**, то нулевую гипотезу отвергают.

в) При конкурирующей гипотезе **H_1: $M_x < M_y$** по таблице критических точек распределения Стьюдента, по заданному уровню значимости α,помещенному в нижней строке таблицы ,и числу степеней свободы **$k= n_x + n_y - 2$** найти «вспомогательную критическую точку» **$t_{кр}$** односторонней критической области.

Если **$T_{набл} < - t_{кр}$**, то нет основания отвергать нулевую гипотезу.

Если **$T_{набл} > - t_{кр}$**, то нулевую гипотезу отвергают.

$T_{набл}$ и число степеней свободы.

$$T_{набл} = \overline{X} - \overline{Y} \Big/ \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \qquad \nu = \frac{S_x^2}{n_x} + \frac{S_y^2}{n_y} \Bigg/ \frac{\left(\frac{S_x^2}{n_x}\right)^2}{\nu_x} + \frac{\left(\frac{S_y^2}{n_y}\right)^2}{\nu_y}$$

Проверка гипотезы о законе распределения генеральной совокупности

Если закон распределения не известен, но есть основание предположить, что он имеет определенный вид (А), то проверяют нулевую гипотезу:

H_0: генеральная совокупность распределена по закону А.

Проверка гипотезы о предполагаемом законе распределения производится так же как и проверка гипотезы о параметрах распределения, т.е. при случайно отобранной случайной величине – критерия согласия.

Критерием согласия *называют критерий проверки гипотезы о предполагаемом законе распределения.*

Имеется несколько критериев согласия:

- критерий Пирсона;
- критерий Колмогорова;
- критерий Смирнова.

Критерий Пирсона

Пусть по выборке объема n получены эмпирические частоты, т.е. мы имеем предполагаемое распределение.

x_i	x_1	x_2	x_3	x_4	x_5
n_i	n_1	n_2	n_3	n_4	n_5

Допустим, что в предположении нормального распределения генеральной совокупности вычислены теоретические частоты ().

При уровне значимости α требуется проверить гипотезу: генеральная совокупность распределена нормально.

В качестве критической проверку нулевой гипотезы примем случайную величину:

$$\chi^{2\,(*)}_{расчет} = \sum \frac{(n_i - n_i')^2}{n_i'}$$

Эта величина случайная, т.к. в различных опытах она принимает различные, заранее не известные, значения. Чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия => он характеризует близость эмпирических и теоретических распределений.

Доказано, что при закон распределения случайной величины **(*)** не зависит от того, какому закону распределения подчинена генеральная совокупность, а стремится к закону распределения χ^2 с числом степеней свободы: **υ=k–1–r** , где

k – число групп (интервалов) выборки

r – число параметров предполагаемого распределения.

А т.к. для нормального распределения нам интересно M(x) и D(x), то число степеней свободы определяется **υ=k–3**

Правила проверки.

Правила проверки

Для того, чтобы при заданном уровне значимости α проверить H0: “генеральная совокупность распределена нормально”, необходимо:

1. вычислить теоретические частоты;
2. вычислить наблюдаемое значение критерия: $\chi^2_{набл} = \sum \frac{(n_i - n'_i)^2}{n'_i}$
3. по таблицам критических точек распределения χ2 по заданному уровню значимости и числу степеней свободы **υ=k–3**, найти критическую точку: $\boldsymbol{\chi^2_{кр}=(\alpha,\upsilon)}$;
4. сравнить 2 имеющихся критерия:

   - если $\chi^2_{набл} < \chi^2_{кр}$ - нет основания отвергать нулевую гипотезу о нормальном распределении.
   - если $\chi^2_{набл} > \chi^2_{кр}$ - нулевую гипотезу о нормальном распределении отвергают.

Замечание:

- объем выборки должен быть достаточно велик (более 50);
- малочисленные группы следует объединять в одну, суммируя частоты;
- т.к. возможные ошибки первого и второго рода, то в окончательном выводе следует проявить осторожность:
  можно повторить опыт;
  увеличить число наблюдений;
  для проверки воспользоваться другими критериями;
  построить график распределения;
  вычислить эксцесс и асимметрию.
- для контроля вычислений формулу $\chi^2_{набл} = \sum \frac{(n_i - n'_i)^2}{n'_i}$ преобразуют к виду $\chi^2_{набл} = \left(\sum \frac{n_i^2}{n'_i}\right) - n$

Корреляционно-регрессионный анализ

- Корреляционная зависимость
- Корреляционный момент
- Коррелированность и зависимость случайных величин
- Выборочное корреляционное отношение
- Простейшие случаи криволинейной корреляции
- Метод наименьших квадратов

Во многих задачах требуется установить или оценить зависимость изучаемо случайной величины Y от одной или нескольких других случайных величин.

Две случайные величины могут быть связаны:

- функциональной зависимостью

- статистической

- независимой

Строгая функциональная зависимость реализуется редко, т.к. обе случайных величины или одна подвержены действию других случайных величин.

Статистической *называется зависимость, при которой изменение одной из величин влечет изменение распределения другой.*

В частности она проявляется в том что изменение одной из величин влечет изменение среднего значения другой. Такая статистическая зависимость называется *корреляционной*.

Корреляционная зависимость

Предположим изучается связь между случайными величинами X и Y. Пусть каждому значению X соответствует несколько значений Y.

Условным средним $\overline{Y}_X$ называется среднее арифметическое случайной величины Y соответствующее значению случайной величины X равное x.

Если каждому значению X соответствует одно значение $\overline{Y}_X$, то очевидно что она – функция от x. В этом случае говорят, что случайная величина Y зависит от X корреляционно.

Корреляционной зависимостью Y_x *называют функциональную зависимость* $\overline{Y}_X$ *от значений x.* $\overline{Y}_X = f(x)$ *- уравнение регрессии Y на X, а график – линией регрессии Y на X. f(x) – функция регрессии.*

Аналогично определяется условная средняя X на Y: $\overline{X}_Y = f(y)$.

Две основные задачи теории корреляции:

1. Оценить тесноту (силу) корреляционной связи
2. Если связь существует, то нужно установит ее форму – вид функциональной зависимости между и величиной X.

Для решения первой задачи существует *коэффициент корреляции.*

Коэффициент корреляции.

Выборочным коэффициентом корреляции называется отношение разности между М(X) произведения случайных величины и произведением математических ожиданий этих случайных величин к σ^2 случайных величин X и Y.

$$r_в = \frac{M(X \cdot Y) - M(X) \cdot M(Y)}{\sigma_x \cdot \sigma_y}$$

Он служит для оценки тесноты линейной корреляционной зависимости.

Свойства коэффициента корреляции

1. $\mathbf{r_в}$ по абсолютной величине ≤ 1
2. Если $\mathbf{r_в} = 0$, то X и Y не связаны линейной зависимостью, а другая может при этом существовать.
3. Если $|\mathbf{r_в}| = 1$, то X и Y связаны строго корреляционной зависимостью.
4. Т.к. rв характеризует степень тесноты линейной связи, то она проявляется в том, что при возрастании одной случайной величины другая имеет тенденцию возрастать, т.е. наблюдается положительная корреляция, rв>0; если при возрастании одной случайной величины другая – убывает, $r_в<0$.

Remark: Зависимость тем ближе к линейному закону, чем $|\mathbf{r_в}|$ ближе к единице.

Для описания системы двух случайных величин или зависимости между двумя случайными величинами, кроме математического ожидания и дисперсии используют и другие величины. Коэффициент корреляции rв является частным случаем такой характеристики, более частным случаем выступает корреляционный момент $\boldsymbol{\mu_{xy}}$.

Корреляционный момент

Корреляционным моментом μ_{xy} *случайных величин X и Y, называют математическое ожидание при отклонении этих величин.*

$\mu_{xy} = M[(X - M(X))*(Y - M(Y))]$

$\mu_{xy} = M(X*Y) - M(X)*M(Y)$

$r_в = \mu_{xy} / \sigma x*\sigma y$

Свойства корреляционного момента:

- корреляционный момент служит для характеристики связи между случайными величинами X и Y;
- корреляционный момент двух независимых величин равен нулю;
- корреляционный момент имеет размерность равную произведению размерности величин X и Y;
- размерность корреляционного момента является недостатком при сравнении зависимости двух случайных величин, чтобы избежать это, был введен коэффициент корреляции.

Коррелированность и зависимость случайных величин

- Две случайных величин называются коррелированными, если их корреляционный момент (или что то же самое коэффициент корреляции) равен нулю;

- X и Y называются некоррелированными случайными величина-ми, если их корреляционный момент равен нулю;

- Две коррелированные величины также зависимы. Обратное предположение не всегда имеет место.

Выборочное корреляционное отношение

Для оценки тесноты <u>нелинейной</u> корреляционной связи служат такие характеристики как: выборочное корреляционное отношение $\boldsymbol{\eta}_{yx}$ (игрек к икс) и $\boldsymbol{\eta}_{xy}$ (икс к игрек).

Выборочным корреляционным отношением *называется, отношение межгруппового среднеквадратического отклонения к общему среднеквадратическому отклонению.*

$$\eta_{yx} = \sigma_{межгр} / \sigma_{общ}$$

Свойства корреляционного отношения:

- если корреляционное отношение равно нулю, X и Y не связаны друг с другом;
- если корреляционное отношение = 1, X и Y не связаны корреляционной зависимостью;
- значение корреляционного отношения удовлетворяет двойному неравенству $0 \le \eta_{yx} \le 1$;
- корреляционное отношение всегда меньше или равно коэффициенту корреляции $\eta_{yx} \le |\mathbf{r}_{в}|$.

Достоинства корреляционного отношения.

Корреляционное отношение служит мерой тесноты связи любой, в том числе и линейной. В этом его достоинство перед коэффициентом корреляции, который оценивает степень тесноты только линейной связи.

Недостатки корреляционного отношения.

Корреляционное отношение не позволяет судить на сколько близко расположены точки найденным по данным наблюдения к кривой определенного вида (гипербола, парабола, синусоида и т.д.). Это объясняется тем, что при определении корреляционного отношения вид связи не учитывается.

Простейшие случаи криволинейной корреляции

Если график регрессии Y на X изображен кривой линией, то корреляция называется криволинейной.

Примеры функции регрессии

1. Параболическая корреляция второго порядка $\bar{y}_x$= ах² + вх + с
2. Параболическая корреляция третьего порядка $\bar{y}_x$=а+вх²+сх + d
3. Гиперболическая $\bar{y}_x$= а/х + в

Для определения вида функции регрессии строят на графике точки с координатами (х, $\bar{y}_x$). По их расположению делают заключение о примерном виде функции регрессии.

При окончательном решении принимают во внимание особенности, вытекаемые из сущности решаемой задачи.

Теория криволинейной корреляции решает те же задачи что и теория линейной корреляции, т.е. устанавливает формы и тесноты корреляционной зависимости.

Неизвестные параметры уравнения регрессии ищут методом «наименьших квадратов».

Метод наименьших квадратов

Метод наименьших квадратов заключается в том, что сумма квадратов отклонений теоретических данных от экспериментальных, должна быть наименьшей.

Будем искать уравнение регрессии в виде ***y = f(x).***

Предположим, что имеет место линейная зависимость $\varphi(x)= a_0 + a_1 x$

a_0, a_1 – коэффициенты линейной зависимости

$\varphi(x)$ – предполагаемая теоретическая зависимость

Используя метод наименьших квадратов построим функцию равную сумме квадратов отклонений экспериментальных данных от теоретических данных:

$$S = \sum_{i=1}^{n} [\varphi(x) - f(x)]^2 \to \min$$

По методу наименьших квадратов для нахождения коэффициентов a_0, a_1, необходимо составить систему двух уравнений, для этого необходимо взять частные производные от функции S и приравнять их к нулю.

$$\begin{cases} \partial S / \partial a_0 = 0 \\ \partial S / \partial a_1 = 0 \end{cases}$$

Учитывая число функций $\varphi(x)=a_0+a_1x$

$$\begin{cases} 2\sum(a_1 + a_1 x_i - f(x_i)) = 0 \\ 2\sum(a_0 + a_1 x - f(x_i)) \cdot x_i = 0 \end{cases}$$

Окончательной системой уравнений по методу наименьшего квадрата имеет вид

$$\begin{cases} \mathrm{n} * \mathrm{a}_0 + \mathrm{a}_1 * \sum \mathrm{x}_\mathrm{i} = \sum f(x_i) \\ \sum \mathrm{x}_\mathrm{i} * \mathrm{a}_0 + \mathrm{a}_1 * \sum \mathrm{x}_\mathrm{i}^2 = \sum f(x_\mathrm{i}) * x_\mathrm{i} \end{cases}$$

Решая полученную систему найдем неизвестные коэффициенты a_0 и a_1, запишем уравнение регрессии в виде $\varphi(x) = a_0 + a_1 x$

Для того чтобы найти погрешность данного метода необходимо вычислить

$$\varepsilon = \frac{\sqrt{S/n}}{Y_{max}} * 100\%; \qquad S = \sum (f(x) - y)^2$$

Если предположили нелинейную корреляцию, то уравнение связи пытаемся искать в виде $\varphi(x)=a_0+a_1x+a_2x^2$, то аналогично по методу наименьших квадратов найдем функцию

$$S = \sum_{i=1}^{n} [a_0 + a_1 x + a_2 x_i^2 - f(x)]^2 \to \min$$

Находим частные производные данной функции

$$\begin{cases} \partial S/\partial a_0 = 0 \\ \partial S/\partial a_1 = 0 \\ \partial S/\partial a_2 = 0 \end{cases}$$

Запишем систему уравнений

$$\begin{cases} n * a_0 + a_1 * \sum x_i + a_2 \sum x_i^2 = \sum f(x_i) \\ \sum x_i * a_0 + a_1 * \sum x_i^2 + a_2 \sum x_i^3 = \sum f(x_i) x_i \\ \sum x_i^2 * a_0 + a_1 * \sum x_i^3 + a_2 \sum x_i^4 = \sum f(x_i) x_i^2 \end{cases}$$

Аналогично находим a_0, a_1, a_2 и записываем уравнение регрессии в виде $\varphi(x)=a_0+a_1x+a_2x^2$

Определяем погрешность с помощью ε.

Заключение о реальной зависимости между случайными величинами X и Y делаем путем графического представления.

- Если нужно отобрать 20% изготовленных деталей, то отбирают каждую пятую.
- Детали изготавливаются на разных станках. Выборка производится с каждого станка.
- Изделия изготавливаются станками-автоматами. Обследованию подвергается продукция нескольких автоматов.

Задано распределение частот выборки.

x	2	6	12
n_i	3	10	7

Определить объем, написать распределение относительных частот.

n=3+10+12=20

x	2	6	12
W	3/20	10/20	7/20

$\Sigma\ W_i=1$

$\Sigma\ n_i=n$

Пусть имеется нормальное распределение. Тогда нужно оценить, найти $M(x)$ и σ. Для показательного распределения нужно оценить параметр λ. $f(x)= \lambda \cdot e^{-\lambda x}$.

Найти доверительный интервал с надежностью 0.9 неизвестного *M(X)* нормально распределенной СВ X, если известны $\bar{X}$=20.9, σ=2, n=16, β=0.9.

2Φ(t)=σ

Φ(t)=β/2=0.45

t=1.645

$\varepsilon = t\sigma / \sqrt{n}$ =0.82

(20.9–0.82; 20.9+0.82)

(20.08; 21.72)

По данным выборки, объема 50, найдена =-0.155, S=936. Найти доверительный интервал для неизвестной дисперсии, β=0.95. n=50, υ=49, α=0.05.

По таблицам распределения Стьюдента для уровня значимости α=0.05 и числа степеней свободы υ=49 найдем значение критерия Стьюдента $t_{\alpha,\upsilon}$=2.009

Запишем значение границы интервала $\varepsilon = \frac{t_{\alpha,\nu} \cdot S}{\sqrt{n}}$=0.27

Запишем границы доверительного интервала (-0.425;0.115).

С вероятностью 0.95 истинное значение *M(x)* лежит в пределах (-0.425;0.115).

При доверительной вероятности 90% найти доверительный интервал для D(x), если для выборки, объемом 5 выборочная D(x)=6.6, а выборочная средняя 0.4.

По таблицам распределения χ^2 найдем значение критерия χ^2 для уровня значимости $\upsilon=4$ и $\alpha=/2=0.05$

$\chi^2_{0.05,4}=9.5$

$\chi^2_{\alpha/2,\upsilon} = \chi^2_{0.95,4}=0.711$

$4 \cdot 6.6/9.5 < \sigma^2 < 4 \cdot 6.6/0.711$

$2.78 < \sigma^2 < 37.13$

Если H_0 состоит в предположении, что математическое ожидание M(X) нормального распределения равно 10, то H_1 может состоять в предположении, что M(X) не равно 10.

H_0: $M(X)=10$

H_1: $M(X)\neq 10$

По двум малым независимым выборкам объемов $n_x=11$ и $n_y=14$ из нормальных распределений найдены исправленные выборочные дисперсии $S^2_x=0.76$ и $S^2_y=0.38$. При уровне значимости $\alpha=0.05$ проверить нулевую гипотезу H_0: $D_x=D_y$ о равенстве диспер-сий при конкурирующей гипотезе H_1: $Dx>Dy$.

Решение: Найдем отношение большей исправленной дисперсии к меньшей:

$F_{набл} = S^2_б / S^2_м = 0.76 / 0.38 = 2$

По условию конкурирующая гипотеза имеет вид H_1: $Dx>Dy$, поэтому критическая область – правосторонняя. По таблице критических то-чек распределения Фишера, по уровню значимости $\alpha=0,05$ и числам степеней свободы $k_1 = n_x - 1 = 11 - 1 = 10$ и

$k_2 = n_y - 1 = 14 - 1 = 13$ находим критическую точку:

$F_{кр}(\alpha, k_1, k_2) = F_{кр}(0.05,10,13) = 2.67$

Так как $F_{набл} = 2. < F_{кр} = 2.67$, то нет оснований отвергать Ho о равенстве дисперсий. Другими словами, исправленные выборочные дисперсии различаются незначимо.