Математическая статистика

Выборочный метод

1. Предмет МС

МС – раздел математики, который изучает методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Предметом МС является изучение СВ (случайных событий, процессов) по результатам наблюдений.

Задачи МС

1. Указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.

2. Разработать методы анализа статистических данных в зависимости от целей исследования

а) оценка интересующих характеристик наблюдаемой СВ;

б) проверка статистических гипотез, т.е. решение вопроса согласования результатов оценивания с опытными данными.

Результаты исследования статистических данных методами МС используются для принятия решения (в задачах планирования, управления, прогнозирования и организации производства, при контроле качества продукции и др.)

2. Генеральная и выборочная совокупность

Пусть наблюдается некоторая СВ *X*. Например, партия изделий проверяется на стандартность, наблюдают количество покупателей, которых успеет обслужить данный продавец за единицу времени и т.д.

Проводить сплошной обследование часто невозможно, либо сопряжено с определенными трудностями. Тогда из всей совокупности объектов отбирают ***случайным образом*** ограниченное число объектов и подвергают их исследованию.

Выборочной совокупностью (***выборкой***) называют совокупность случайно отобранных объектов.

Генеральной совокупностью называют совокупность объектов из которых производилась выборка

Объемом совокупности (генеральной или выборочной) называют количество объектов этой совокупности.

Выборка может быть ***повторной*** и ***бесповторной***.

Если объем генеральной совокупности достаточно велик, а выборка составляет лишь незначительную ее часть, то различие между повторной и бесповторной выборке практически стираются.

Для того, чтобы по выборке правильно судить о генеральной совокупности, надо, 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 ее осуществлять случайно.

На практике применяют различные способы отбора. Их можно разделить на две группы:

1) отбор, не требующий разделения генеральной совокупности на части: простой случайный повторный или бесповторный отбор;

2) отбор, при котором генеральная совокупность разделяется на части: типический, механический, серийный.

На практике применяют комбинированный отбор.

Все объекты, составляющие генеральную совокупность, должны иметь хотя бы один общий признак, позволяющий классифицировать объекты, сравнивать их. Наличие общего признака является основой для образования статистической совокупности.

Статистическая совокупность представляет результаты описания или измерения общих признаков объектов исследования.

Предметом изучения в статистике являются изменяющиеся (варьирующие) признаки, которые иногда называют статистическими признаками. Они делятся

<table>
<tr><th>качественные</th><th colspan="2">количественные</th></tr>
<tr><td rowspan="3">признаки, которыми объект обладает или не обладает. Они не поддаются непосредственному измерению (цвет волос, квалификация, национальность и т.д)</td><td colspan="2">результаты подсчета или измерения</td></tr>
<tr><td>дискретные</td><td>непрерывные</td></tr>
<tr><td>могут принимать лишь отдельные значения из некоторого ряда чисел (число промахов и попаданий, число пришедших людей и т.д.)</td><td>могут принимать любые значения в некотором интервале (время работы механизма, скорость движения и т.д.)</td></tr>
</table>

Эмпирические распределения

1. Статистическое распределение

Пусть из генеральной совокупности извлечена выборка, в которой x_1 встречается n_1 раз $x_2 - n_2$ раз, …, $x_k - n_k$ раз, причем

$\sum_{i=1}^{k} n_i = n -$ объем выборки.

Наблюдаемые значения x_i называют ***вариантами.***

$n_i -$ ***частоты.***

$\frac{n_i}{n} = \omega_i -$ ***относительные частоты*** или ***частости.***

Вся совокупность полученных объектов представляет собой ***первичный материал***, который подлежит дальнейшей обработке.

Операцию расположения значений СВ по неубыванию называют ***ранжированием*** статистических данных.

Вариационный ряд – ранжированный в порядке возрастания (убывания) ряд вариант.

Статистический ряд (статистическое распределение) – ранжированный в порядке возрастания (убывания) ряд вариант с соответствующими им весами (частотой, частостью и т.д.)

Статистическое распределение бывает ***дискретным*** и ***интервальным***.

Статистическое распределение записывают в виде таблицы.

Графическим представление статистического распределения является ***полигон частот*** (***относительных частот***)

Пример. В супермаркете проводилось наблюдение над числом X покупателей, обратившихся в кассу за один час. Наблюдения в течении 20 часов (10 дней с 9.00 до 10.00 и с 10.00 до 11.00) дали следующий результат:

70; 75; 100; 120; 75; 60; 100; 120; 70; 60;

65; 100; 65; 100; 70; 75; 60; 100; 100; 120.

Составьте вариационный ряд и статистическое распределение.

Наблюдаемая величина X является ДСВ

Ранжируем полученные данные

60; 60; 60; 65; 65; 70; 70; 70; 75; 75; 75;

100; 100; 100; 100; 100; 100; 120; 120; 120.

60; 60; 60; 65; 65; 70; 70; 70; 75; 75; 75;

100; 100; 100; 100; 100; 100; 120; 120; 120.

Составим статистическое распределение выборки (вариационный ряд).

x_i						
n_i						
ω_i						

$n = 20$

$$\omega_i = \frac{n_i}{n}$$

Статистическое распределение выборки является статистической оценкой неизвестного распределения.

В соответствии с теоремой Бернулли при $n \to \infty$ относительные частоты ω_i сходятся по вероятности к соответствующим вероятностям. Поэтому при больших n статистическое распределение мало отличается от истинного.

В случаях, если число значений признака *X* велико или признак является непрерывным, составляют ***интервальный статистический ряд***.

Интервальный статистический ряд задают в виде последовательности непересекающихся интервалов и соответствующих им частот (частостей) (в качестве частоты принимают сумму частот вариант, попавших в интервал).

Число интервалов выбирают произвольно

$$[5;10] \le k \le [20;25]$$

Или рассчитывают по формуле Стерджесса (Стёрждеса)

$$n = 1 + [\log_2 N]$$

$$n = 1 + [3{,}322 \lg N]$$

Длину интервалов, как правило, выбирают одинаковой, равной

$$h \approx \frac{x_{max} - x_{min}}{k}$$

$[x_{i-1};\ x_i)$	$[x_0;\ x_1)$	$[x_1;\ x_2)$	...	$[x_{k-1};\ x_k]$
$x_i^* = \frac{x_i + x_{i-1}}{2}$	x_1^*	x_2^*	...	x_k^*
n_i	n_1	n_2	...	n_k
$\omega_i = \frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$

Гистограммой частот (относительных частот) называют ступенчатую фигуру, составленную из прямоугольников, построенных на интервалах группировки так, что площадь каждого прямоугольника равна соответствующей данному интервалу частоте (относительной частоте) (высоты прямоугольников равны

$$\frac{n_i}{h} \quad \left(\frac{n_i}{nh} \right).$$

Площадь гистограммы относительных частот равна 1, гистограммы частот - объему выборки.

2. Эмпирическая функция распределения

Эмпирической функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения варианты x относительную частоту наблюдения значений, меньших x:

$$F^*(x) = \sum_{x_i^* < x} \frac{n_i}{n}$$

Эмпирическую функцию распределения используют в качестве оценки теоретической функции распределения $F(x) = P(X < x)$.

$F^*(x)$ обладает теми же свойствами, что и функция распределения дискретной СВ в теории вероятностей

1) $0 \le F^*(x) \le 1$;

2) $F^*(x)$ – неубывающая функция;

3) $F^*(x)$ – непрерывная слева кусочно-постоянная функция;

4) если $x_{\min}$ – наименьшее, и $x_{\max}$ – наибольшее выборочные значения, то $F^*(x) = 0$ при $x \le x_{\min}$ и $F^*(x) = 1$ при $x > x_{\max}$.

Статистические оценки параметров распределения

1. Основные определения

Любое значение параметра распределения, найденное на основе ограниченного числа опытов, содержит элемент случайности.

Статистической оценкой (оценкой) θ^* параметра θ теоретического распределения называют его приближенное значение, зависящее от данных выбора.

Любая из таких оценок случайна и при ее использовании неизбежны ошибки. Желательно выбрать такую оценку, чтобы эти ошибки были по возможности минимальны.

Обозначим θ^* оценку параметра θ, найденную по данным выборки.

$$\theta^* = \theta(X_1, X_2, \ldots X_n)$$

θ^* является СВ. Ее закон распределения зависит от закона распределения наблюдаемой СВ X и от числа опытов.

Функцию результатов наблюдений (т.е. функцию выборки) называют ***статистикой***.

К статистической оценке предъявляют требования:

1. $\lim\limits_{n\to\infty} P\left(\left|\theta-\theta^*\right|<\varepsilon\right)=1$

Статистическая оценка должна быть ***состоятельной***.

2. $M\left(\theta^*\right)=\theta$

Статистическая оценка не должна содержать систематической ошибки (ошибки одного знака). Оценку, удовлетворяющую такому условию называют ***несмещенной***.

3. $D\left(\theta^*\right)\to\min$

Оценка должна быть ***эффективной***.

2. Точечные оценки

Точечной называют статистическую оценку, которая определяется одним числом.

Пусть статистическое распределение выборки имеет вид:

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2		n_k

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}_в\right)^2 n_i$$

$$\sigma_в = \sqrt{D_в}$$

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i\right)^2 n_i - \left(\overline{x}_в\right)^2$$

Теорема. $\overline{X}_в = \frac{1}{n}\sum_{i=1}^{n} X_i$ – несмещенная, состоятельная оценка MX.

Замечание. Можно показать, что при нормальном распределении СВ X эта оценка будет так же и эффективной.

Пусть изучают СВ X с математическим ожиданием $MX = a$ и дисперсией DX, оба параметра неизвестны.

Пусть $x_1,\ x_2, \ldots, x_n$ – выборка, полученная в результате n независимых наблюдений.

$$X_1,\ X_2, \ldots, X_n$$

$$D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i)^2 n_i - \left(\overline{x}_в\right)^2$$

Выборочная дисперсия – смещенная оценка генеральной дисперсии. Поэтому, выборочную дисперсию «исправляют», умножая на $\frac{n}{n-1}$.

$$s^2 = \frac{n}{n-1}D_в \qquad s = \sqrt{s^2}$$

Можно доказать, что s^2 также является состоятельной оценкой генеральной дисперсии.

3. Интервальные оценки

Интервальной называют статистическую оценку, которая определяется двумя числами – концами интервала.

Интервальные оценки позволяют установить *точность* и *надежность* оцениваемого параметра.

Пусть θ^* – статистическая оценка неизвестного параметра θ.

θ^* тем точнее определяет θ, чем меньше $|\theta - \theta^*|$.

$$|\theta - \theta^*| < \delta, \ \delta > 0$$

δ – ***точность*** оценки θ^*.

Статистические методы не позволяют категорически утверждать, что θ^* удовлетворяет неравенству $|\theta - \theta^*| < \delta$. Можно лишь говорить о вероятности, с которой это неравенство выполняется.

Надежностью (***доверительной вероятностью***) оценки θ^* называют вероятность γ с которой выполняется неравенство $|\theta - \theta^*| < \delta$.

Надежность задают наперед значениями, равными 0,95; 0,99; 0,999.

Пусть $P\left(|\theta - \theta^*| < \delta\right) = \gamma$.

$$P\left(\theta^* - \delta < \theta < \theta^* + \delta\right) = \gamma$$

$$P\left(\theta^* - \delta < \theta < \theta^* + \delta\right) = \gamma$$

Интервал $\left(\theta^* - \delta; \theta^* + \delta\right)$, который покрывает неизвестный параметр θ с заданной надежностью γ называют ***доверительным интервалом.***

Пусть наблюдаемая СВ X распределена нормально и ***известно*** σ.

С надежностью γ интервал $\left(\overline{x}_в - \frac{t\sigma}{\sqrt{n}}; \overline{x}_в + \frac{t\sigma}{\sqrt{n}}\right)$

покрывает неизвестный параметр a. Точность оценки $\delta = \frac{t\sigma}{\sqrt{n}}$.

Пример. СВ X имеет нормальной распределение с известным сренеквадратическим отклонением $\sigma = 3$. Найдите доверительные интервалы для оценки неизвестного математического ожидания по выборочным средним $\overline{x}$, если объем выборки $n = 36$ и задана надежность $\gamma = 0{,}95$.

Пусть наблюдаемая СВ X распределена нормально и σ ***неизвестно***. Доверительный интервал для неизвестного параметра a можно найти по формуле

$$\left(\overline{x}_в - \frac{t_\gamma \sigma}{\sqrt{n}}; \overline{x}_в + \frac{t_\gamma \sigma}{\sqrt{n}} \right),$$

где t_γ находят по таблицам, зная n и γ.

Пусть наблюдаемая СВ X распределена нормально. Оценим неизвестное среднеквадратическое отклонение σ по исправленному s.

$$s(1-q)<\sigma<s(1+q)$$

Для нахождения q используют χ^2-распределение или специальные таблицы.

Замечание. В формуле $s(1-q)<\sigma<s(1+q)$ Предполагают, что $q<1$. Если $q>1$, то

$$0<\sigma<s(1+q).$$

Статистические гипотезы

1. Основные определения

Основные задачи МС разделяют на две категории:

- оценивание неизвестных параметров распределения (получение по выборке оценок, наилучших в том или ином смысле);

- проверка статистических гипотез (по выборке принять или отвергнуть некоторое предположение о распределении генеральной совокупности, из которой извлечена выборка).

Статистической гипотезой называют любое предположение о виде (***непараметрическая гипотеза***) или параметрах (***параметрическая гипотеза***) неизвестного распределения.

Статистическую гипотезу называют ***простой***, если она полностью определяет функцию распределения. В противном случае гипотезу называют ***сложной***.

Пример. Предположим, что введен новый способ производства некоторого товара. Для определения качества товара измеряют некоторую его характеристику $\xi \sim N(a_0, \sigma_0)$, где a_0, σ_0 известны.

Если необходимо выяснить, как новый способ производства влияет на качество товара, можно выдвинуть, например, такие гипотезы:

$H_1: a = a_0, \ \sigma = \sigma_0,$ т.е. распределение СВ ξ не изменилось после изменения процесса производства;

Пример. Предположим, что введен новый способ производства некоторого товара. Для определения качества товара измеряют некоторую его характеристику $\xi \sim N(a_0, \sigma_0)$, где a_0, σ_0 известны.

$H_1: a = a_0, \ \sigma = \sigma_0,$ т.е. распределение СВ ξ не изменилось после изменения процесса производства;

$H_2: a > a_0, \ \sigma = \sigma_0,$ т.е. увеличилось среднее показателя качества;

$H_3: a = a_0, \ \sigma < \sigma_0,$ т.е. разброс значений показателя качества стал меньше.

Проверяемую гипотезу называют ***нулевой*** (***основной***) и обозначают H_0. Наряду с нулевой рассматривают ***альтернативную*** (***конкурирующую***) гипотезу $H_1 \ \left(\bar{H}\right)$.

$$H_0 : \ \theta = \theta_0, \quad H_1 : \ \theta \neq \theta_0.$$

$$H_0 : \ \theta = \theta_0, \quad H_1 : \ \theta > \theta_0.$$

Правило, которое позволяет по выборке принять или отвергнуть проверяемую гипотезу, называют ***критерием проверки статистической гипотезы*** (***статистическим критерием***)

Статистическими методами нельзя доказать правильность гипотезы.

Критерий проверки статистической гипотезы позволяет отбросить гипотезу как неправильную, но не позволяет доказать, что она верна, т.е. статистические критерии указывают лишь на отсутствие опровержения со стороны имеющихся экспериментальных данных.

Если по результатам проверки статистическая гипотеза принимается, то говорят, что она ***согласуется с выборочными данными*** или, что она ***не противоречит результатам наблюдений***.

Статистический критерий обычно основывается на некоторой статистике $\overline{\theta}_n$, для которой известно ее точное или приближенное распределение.

Множество всех возможных значений этой статистики разбивают на два непересекающихся множества: ***область принятия нулевой гипотезы*** и ***область отклонения нулевой гипотезы*** (***критическая область***).

Проверяемая гипотеза H_0	H_0 принимается -	H_0 отвергается -
объективно верна	правильное решение	***ошибка 1-го рода***
объективно неверна	***ошибка 2-го рода***	правильное решение

Вероятность ошибки первого рода, т.е. вероятность отвергнуть нулевую гипотезу, когда она верна, называют ***уровнем значимости статистического критерия*** и обозначают α :

$$P\left(H_0 \text{ отвергается} \,\middle|\, H_0 \text{ верна}\right) = \alpha.$$

Проверяемая гипотеза H_0	H_0 принимается -	H_0 отвергается -
объективно верна	правильное решение	***ошибка 1-го рода***
объективно неверна	***ошибка 2-го рода***	правильное решение

Вероятность ошибки второго рода, т.е. вероятность ошибочно принять нулевую гипотезу, обозначают β:

$$P\left(H_0 \text{ принимается} \,\middle|\, H_0 \text{ не верна}\right) = \beta.$$

Проверяемая гипотеза H_0	H_0 принимается -	H_0 отвергается -
объективно верна	правильное решение	***ошибка 1-го рода***
объективно неверна	***ошибка 2-го рода***	правильное решение

Пользуясь терминологией статистического контроля качества продукции, можно сказать, что α – это риск поставщика (забраковка партии, удовлетворяющей стандарту), а β – риск потребителя (принятие партии не удовлетворяющей стандарту)

Мощностью критерия называют вероятность отклонить проверяемую гипотезу H_0, когда она неверна.

$$P\left(H_0 \text{ отвергается} \,\middle|\, H_0 \text{ не верна}\right) = \beta.$$

При построении статистических гипотез требование, чтобы ошибки обоих родов были бы минимальны противоречиво. Невозможно одновременно уменьшить обе ошибки.

Проверяемая гипотеза H_0	H_0 принимается -	H_0 отвергается -
объективно верна	правильное решение	***ошибка 1-го рода***
объективно неверна	***ошибка 2-го рода***	правильное решение

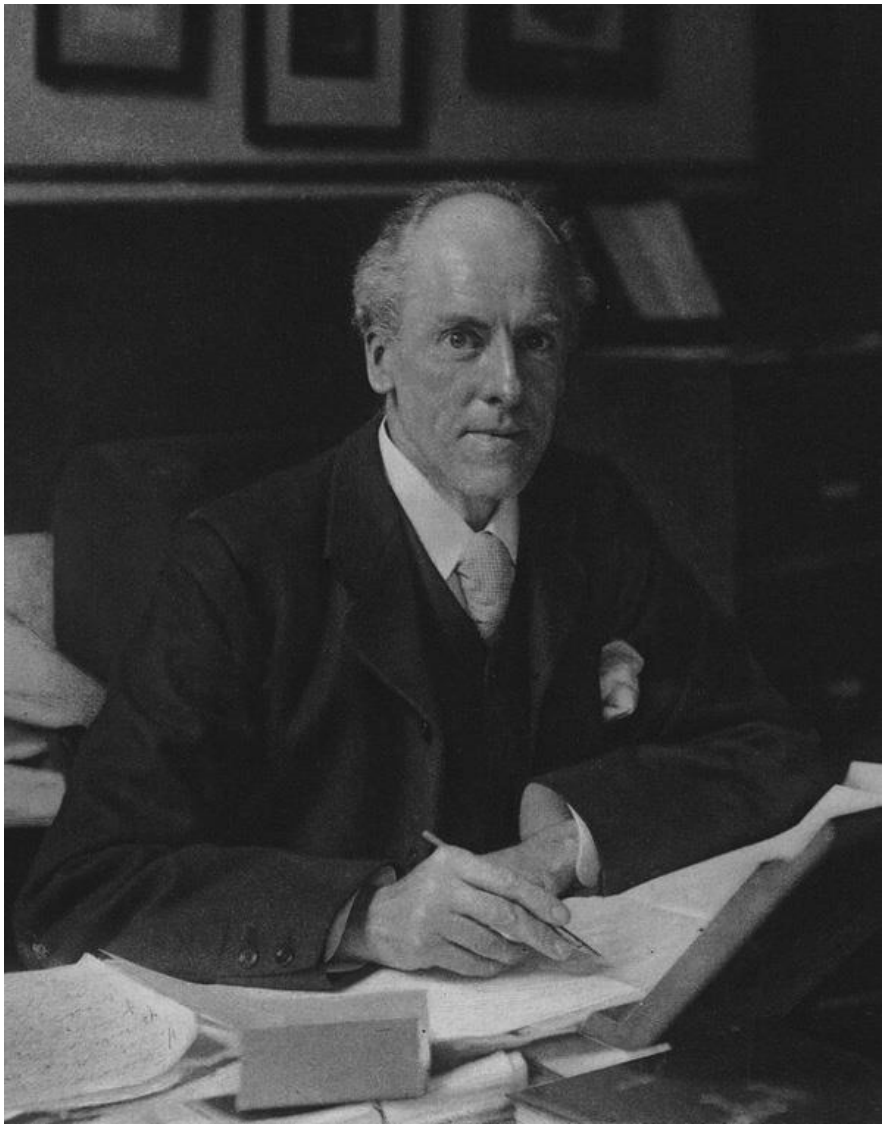
На практике поступают так: задают уровень значимости α (как правило, равный 0,05, 0,01 или 0,1) а, затем, выбирают статистический критерий так, чтобы ошибка второго рода была наименьшей.

Статистические критерии, с помощью которых проверяют гипотезы о значениях параметров распределения или о соотношениях между ними, в предположении, что тип распределения известен, называют ***критериями значимости*** (***параметрическими критериями***)

Статистические критерии, с помощью которых проверяют гипотезы о виде распределения, называют ***критериями согласия*** (***непараметрическими критериями***)

Наиболее известными являются критериями согласия являются критерий χ^2 Пирсона и критерий Колмогорова.

Карл Пирсон
27.03.1857-27.04.1936



Английский математик и биолог, основатель английской школы биометрики.

Внес существенный вклад в распространение методов статистического анализа в биологии и психологии.

Основные идеи Пирсона были опубликованы в серии из 19 статей под общим названием «Математический вклад в теорию эволюции».

Пирсон считается одним из основоположников современной статистики.

2. Критерий согласия χ^2 Пирсона

Пусть имеется выборка объема n и сгруппированный статистический ряд, в котором k групп (например, в случае непрерывной СВ это будет k интервалов).

Группы выбирают так, чтобы охватить весь диапазон значений предполагаемой СВ. Если диапазон значений СВ неограничен, то крайние интервалы должны быть расширены до $-\infty$ и $+\infty$ соответственно.

В каждый интервал должно входить не менее 5 наблюдений. Группы с малым числом наблюдений объединят с соседними.

Проверяемая гипотеза представляет собой предположение о виде распределения наблюдаемой СВ и является простой (конкретно указывает предполагаемое распределение)

H_0: функция распределения наблюдаемой СВ совпадает с $F(x)$.

H_1: функция распределения наблюдаемой СВ не совпадает с $F(x)$.

Критерий согласия Пирсона основан на сравнении эмпирических и теоретических частот попадания СВ в рассматриваемые группы (интервалы)

n_i – эмпирическая частота наблюдения значений из интервала $\left[x_{i-1}; x_i\right)$

$$n_i' = np_i = nP\left(\xi \in \left[x_{i-1}; x_i\right)\right) = n\left(F\left(x_i\right) - F\left(x_{i-1}\right)\right)$$

n_i' – теоретическое значение соответствующей частоты

По данным выборки вычисляют статистику

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}$$

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}$$

Для вычисления статистики $\chi^2_{набл}$ нужно знать сгруппированный статистический ряд и теоретическую функцию распределения $F(x)$.

При этом $F(x)$ может зависеть от одного или нескольких параметров. Пусть r – число неизвестных параметров теоретического распределения. В этом случае вместо значений параметров используют их оценки.

Теорема. Если теоретическая функция распределения зависит от r параметров и оценки этих параметров обладают свойствами асимптотической нормальности и асимптотической эффективности, то, независимо от вида теоретической функции распределения $F(x)$ в пределе (при $n \to \infty$) статистика $\chi^2_{набл}$ имеет распределение χ^2 с числом степеней свободы $k - r - 1$, где k – число интервалов группировки, r – количество параметров теоретической функции распределения, оцениваемых по данной выборке.

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

Критерий согласия χ^2 Пирсона заключается в следующем:

если $\chi^2_{набл} < \chi^2_{\alpha, k-r-1}$, где $\chi^2_{\alpha, k-r-1}$ определяют по таблице критических значений распределения χ^2, то гипотеза H_0 принимается (признается непротиворечащей экспериментальным данным; нет оснований отвергнуть гипотезу H_0) на уровне значимости α;

если $\chi^2_{набл} \geq \chi^2_{\alpha, k-r-1}$, то гипотеза H_0 отвергается (не согласуется с данными эксперимента).

Основное достоинство критерия согласия χ^2 Пирсона – его универсальность, т.е. применимость для любого закона распределения, в том числе с неизвестными параметрами

Основное недостаток – необходимость большого объема выборки (не менее 60-100 наблюдений) и произвольность группировки, влияющая на величину $\chi^2_{набл}$.