Раздел 18. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

§ 1. Выборка и ее описание

Теория вероятностей и математическая статистика занимаются анализом закономерностей случайных массовых явлений. В теории вероятностей определяются вероятности тех или иных событий по известным вероятностям более простых событий, числовые характеристики случайных величин или вероятности, связанные с этими величинами, по известным законам распределения этих случайных величин. На практике для нахождения законов распределения случайных величин необходимо использовать экспериментальные данные.

Основной задачей математической статистики является разработка методов получения вероятностных характеристик случайных явлений на основе результатов наблюдений или эксперимента.

Математическая статистика опирается на теорию вероятностей и в свою очередь служит основой для разработки методов обработки и анализа статистических результатов в конкретных областях человеческой деятельности.

Понятия генеральной совокупности и выборки

Исходными понятиями математической статистики являются понятия генеральной и выборочной совокупностей.

Опр. 1. ***Выборка (случайная выборка, выборочная совокупность)*** – множество значений результатов наблюдений над одной и той же случайной величиной при одних и тех же условиях. Элементы выборки называются ***выборочными значениями***. Количество проведенных наблюдений называется ***объемом выборки***.

Опр. 2. ***Генеральной совокупностью*** называется множество всех возможных наблюдений над случайной величиной при данном комплексе условий.

В большинстве случаев генеральная совокупность бесконечна (можно производить сколь угодно много наблюдений).

При контроле качества данной партии товаров объем генеральной совокупности равен объему этой партии. Если обследование всей партии невозможно (например, обследование объекта связано с его уничтожением или требует больших материальных затрат), то о качестве партии судят по случайной выборке товаров из этой партии.

Назначение статистических методов в том, чтобы по выборке ограниченного объема сделать вывод о свойствах генеральной совокупности в целом.

Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем нас признаке генеральной совокупности, необходимо, чтобы объекты выборки «правильно» его представляли.

Опр. 3. Выборка называется ***репрезентативной*** (представительной), если она достаточно хорошо отражает изучаемые свойства генеральной совокупности.

Считается, что это требование выполняется, если объем выборки достаточно велик и все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку, т.е. при отборе сохраняется принцип случайности. Такую выборку называют ***случайной выборкой***.

Опр. 4. Выборка называется ***повторной***, если каждый выбранный элемент перед отбором следующего возвращается в генеральную совокупность. Если такого возвращения не происходит, выборка называется ***бесповторной***.

Например, в задаче контроля качества, как правило, рассматриваются бесповторные выборки, а если производится несколько измерений некоторой величины, то выборка считается повторной.

Статистический ряд и его графическое изображение

Пусть имеется выборка объема n: x_1; x_2; …; x_n.

Опр. 5. ***Вариационным рядом*** выборки x_1, x_2, …, x_n называется способ ее записи, при котором ее элементы упорядочены (как правило, в порядке неубывания).

Пример 1. Дана выборка: 2; 4; 7; 3; 1; 1; 3; 2; 7; 3.

Запишем ее вариационный ряд: 1; 1; 2; 2; 3; 3; 3; 4; 7; 7. •

Опр. 6. Разность W между максимальным и минимальным элементами называется ***размахом выборки***: $W = x_{\max} - x_{\min}$.

Как правило, некоторые выборочные значения могут совпадать, поэтому часто выборку представляют в виде статистического ряда.

Опр. 7. Пусть в выборке элемент x_i встречается n_i раз. Число n_i называется ***частотой*** выборочного значения x_i, а $\frac{n_i}{n}$ – ***относительной частотой***.

Очевидно, что $\sum_{i=1}^{k} n_i = n$, где k – число различных элементов выборки.

Опр. 8. Последовательность пар $(x_i^*; n_i)$, где $x_1^*, x_2^*, \ldots, x_k^*$ – различные выборочные значения, а $n_1, n_2, \ldots, n_k$ – соответствующие им частоты, называется ***статистическим рядом***.

Обычно статистический ряд записывается в виде таблицы, первая строка которой содержит различные выборочные значения x_i^*, а вторая – их частоты n_i (или относительные частоты $\frac{n_i}{n}$, иногда и те, и другие):

x_i^*	x_1^*	x_2^*	…	x_k^*
n_i	n_1	n_2	…	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$

При большом объеме (больше 30) выборки ее элементы объединяют в группы (разряды), представляя результаты опытов в виде ***интервального*** (***группированного***) ***статистического ряда***. Для этого интервал, содержащий все элементы выборки, разбивают на k непересекающихся интервалов. Число интервалов выбирается произвольно и, как правило, $5 \div 10 \le k \le 20 \div 25$. Вычисления значительно упрощаются, если интервалы имеют одинаковую длину $h \approx \frac{W}{k}$. (В дальнейшем будет рассматриваться именно этот случай.) После того, как частичные интервалы выбраны, опреде-

ляют частоты n_i – количество элементов выборки, попавших в i-й интервал (элемент, совпадающий с верхней границей интервала, относится к последующему интервалу) и относительные частоты $\frac{n_i}{n}$. Полученные данные сводятся в таблицу:

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

Для наглядного представления выборки используют полигон (для дискретных статистических рядов) и гистограмму (для интервальных статистических рядов) частот (или относительных частот).

Опр. 9. ***Полигоном частот*** называется ломаная с вершинами в точках $(x_i^*; n_i), 1 \le i \le k$ (см. рис. 1); ***полигоном относительных частот*** – ломаная линия с вершинами в точках $\left(x_i^*; \frac{n_i}{n} \right), 1 \le i \le k.$

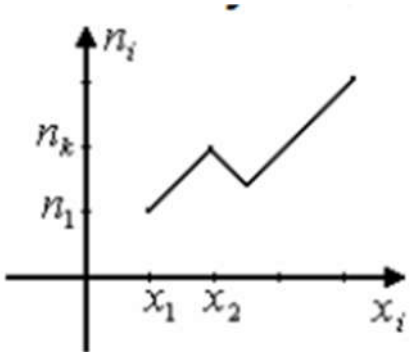

Рис. 1. Полигон частот

Опр. 10. ***Гистограммой относительных частот*** (***частот***) называют ступенчатую фигуру, составленную из прямоугольников, построенных на интервалах группировки так, что площадь каждого прямоугольника равна соответствующей данному интервалу относительной частоте (частоте) (высоты прямоугольников равны $\frac{n_i}{nh}$ в случае гистограммы относительных частот и $\frac{n_i}{h}$ в случае гистограммы частот, см. рис. 2).

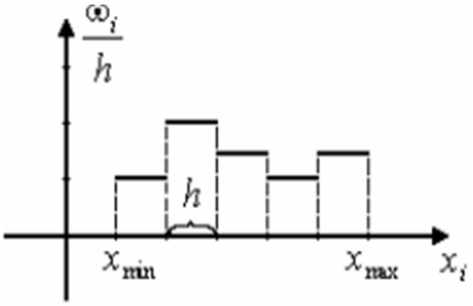

Рис. 2. Гистограмма относительных частот

Гистограмма относительных частот обладает тем свойством, что ее площадь равна 1. Площадь гистограммы частот равна объему выборки n.

При достаточно большом объеме выборки n и достаточно малых интервалах группировки h гистограмма относительных частот является хорошим приближением графика плотности распределения наблюдаемой случайной величины. Поэтому по виду гистограммы можно выдвинуть предположение (гипотезу) о распределении изучаемой случайной величины.

Эмпирическая функция распределения

Опр. 11. ***Эмпирической функцией распределения*** называется функция $F_n^*(x)$, определяющая для каждого значения x относительную частоту наблюдения значений, меньших x:

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

Слово «эмпирический» означает «полученный по экспериментальным (опытным) данным», т.е. по выборке. По этой причине иногда употребляют термин «выборочная функция распределения».

Из определения эмпирической функции распределения видно, что она обладает такими же свойствами, как функция распределения дискретной случайной величины в теории вероятностей, а именно:

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

2) $F_n^*(x)$ – неубывающая функция;

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

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

Пример 1 (продолжение). Для выборки $2; 4; 7; 3; 1; 1; 3; 2; 7; 3$ запишем эмпирическую функцию распределения и построим ее график.

Объем выборки $n = 10$. Составим статистический ряд:

x_i^*	1	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	0,2	0,2	0,3	0,1	0,2

Запишем эмпирическую функцию распределения, накапливая относительные частоты:

$$F_n^*(x) = \begin{cases} 0 & \text{при } x \le 1, \\ 0,2 & \text{при } 1 < x \le 2, \\ 0,4 & \text{при } 2 < x \le 3, \\ 0,7 & \text{при } 3 < x \le 4, \\ 0,8 & \text{при } 4 < x \le 7, \\ 1 & \text{при } x > 7. \end{cases}$$

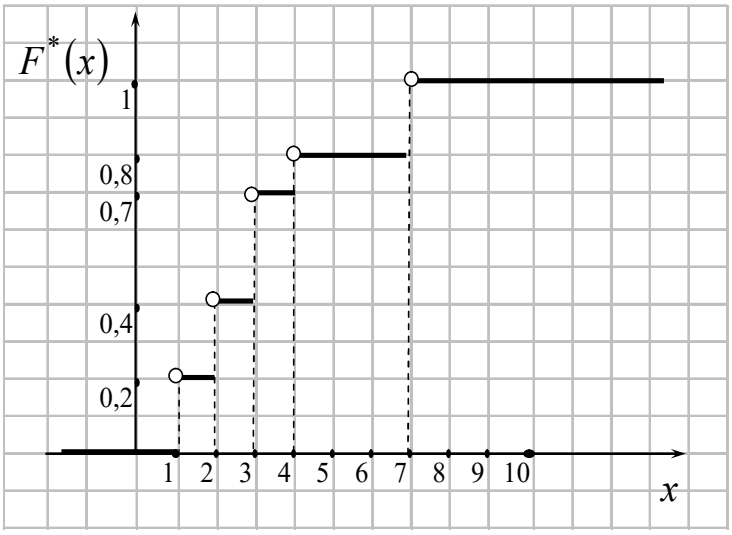

Рис. 3. График эмпирической функции распределения

График $F_n^*(x)$ представлен на рис. 3. •

Замечание. Если график $F_n^*(x)$ строится по интервальному статистическому ряду, то скачки происходят в точках, соответствующих серединам интервалов группировки.

Основное значение эмпирической функции распределения в том, что она используется в качестве оценки теоретической функции распределения $F(x) = P(\xi < x)$ наблюдаемой случайной величины ξ.

Пусть имеется выборка наблюдений над случайной величиной ξ. Значение $F_n^*(x)$ эмпирической функции распределения в точке x равно относительной частоте наблюдения значений, меньших x, т. е. относительной частоте события $\{\xi < x\}$. Согласно закону больших чисел (в форме Я. Бернулли), при $n \to \infty$ относительная частота стремится к вероятности события, т. е., в данном случае, к $P(\xi < x) = F(x)$.

§ 2. Точечное оценивание параметров распределения. Свойства точечных оценок

Основные задачи статистической обработки одной выборки:

1) оценивание параметров распределения;

2) проверка статистических гипотез о виде или параметрах распределения.

Пусть имеется выборка объема n: x_1; x_2; ...; x_n. Выборка представляет собой ряд наблюдений над одной и той же случайной величиной. При статистическом анализе выборки в первую очередь стремятся оценить математическое ожидание и дисперсию. По результатам этого ограниченного числа наблюдений невозможно *вычислить* числовые характеристики наблюдаемой случайной величины, а можно только *оценить* их.

Опр. 1. Любая функция $\hat{\theta}_n = \hat{\theta}_n(x_1; x_2; ...; x_n)$, зависящая от выборочных значений, называется ***статистикой*** или ***выборочной функцией***.

Опр. 2. ***Точечной оценкой*** параметра θ называется любая статистика $\hat{\theta}_n$, предназначенная для оценки этого параметра и определяемая одним числом.

Подчеркнем, что точечная оценка практически никогда не совпадает с истинным значением параметра, она может только оценивать его с большей или меньшей точностью.

Для любого параметра можно предложить разные оценки. Так, в качестве оценки для математического ожидания можно использовать первый элемент выборки x_1, среднее арифметическое наибольшего и наименьшего элементов выборки, среднее арифметическое всех элементов выборки и т. д.

Задача статистического оценивания параметров заключается в том, чтобы из всего множества оценок выбрать в некотором смысле наилучшую. Это означает, что распределение случайной величины $\hat{\theta}_n(x_1; x_2; ...; x_n)$ должно концентрироваться около истинного значения параметра θ.

Замечание. Если, имея выборку x_1; x_2; ...; x_n значений некоторой случайной величины, повторно провести n независимых наблюдений над этой случайной величиной, то новая выборка $x_1'; x_2'; \ldots; x_n'$, вообще говоря, не будет совпадать с первоначальной. Поэтому выборочные значения можно рассматривать как случайные величины.

Основное предположение математической статистики: выборочные значения x_1; x_2; ...; x_n являются независимыми в совокупности одинаково распределенными случайными величинами. Следовательно, любая оценка $\hat{\theta}_n(x_1; x_2; ...; x_n)$ также являются случайной величиной.

Основные характеристики точечных оценок

Качество точечной оценки характеризуется следующими основными свойствами.

Опр. 3. Оценка $\hat{\theta}$ называется ***несмещенной*,** если ее математическое ожидание равно оцениваемому параметру: $M\hat{\theta} = \theta$.

Требование несмещенности гарантирует отсутствие систематических ошибок при оценивании. Оно особенно важно при малом числе наблюдений (в случае выборок объема не более 30).

Опр. 4. Оценка $\hat{\theta}_n$ называется ***состоятельной***, если при увеличении объема выборки n оценка $\hat{\theta}_n$ сходится по вероятности к θ: $\hat{\theta}_n \xrightarrow{P} \theta$ при $n \to \infty$.

Это свойство означает, что при большом объеме выборки практически достоверно, что $\hat{\theta}_n \approx \theta$. Чем больше объем выборки, тем более точные оценки можно получить.

Опр. 5. Пусть $\hat{\theta}_1$ и $\hat{\theta}_2$ – две различные *несмещенные* оценки параметра. Если для их дисперсий выполняется условие $D\hat{\theta}_1 < D\hat{\theta}_2$, то говорят, что оценка $\hat{\theta}_1$ более эффективна, чем оценка $\hat{\theta}_2$. Оценка с наименьшей дисперсией называется ***эффективной*.**

Это означает, что распределение эффективной оценки наиболее тесно сконцентрировано около истинного значения параметра.

Замечание. Не всегда можно найти оценки, которые имели бы все указанные свойства.

Точечные оценки для математического ожидания и дисперсии

Для негруппированной выборки	Для группированного статистического ряда
Выборочное среднее	
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i$
Выборочная дисперсия	

$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ $D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2$	$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x})^2\, n_i$ $D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i^*)^2 n_i - (\overline{x})^2$
Несмещенная оценка дисперсии $s^2 = \frac{n}{n-1}D_{\text{в}}$	
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n(\overline{x})^2\right)$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i^* - \overline{x})^2\, n_i$ $s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i^*)^2\, n_i - \frac{n}{n-1}(\overline{x})^2$

Выборочное среднее $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ (среднее арифметическое элементов выборки) характеризует центр распределения (рассеивания) изучаемой случайной величины.

Выборочная дисперсия $D_{\text{в}}$ характеризует степень разброса (рассеяния) выборочных значений относительно среднего.

Свойства выборочных среднего и дисперсии как оценок для математического ожидания и дисперсии

Утв. 1. Выборочное среднее является *несмещенной* и *состоятельной*, а в случае выборки из нормального распределения и *эффективной* оценкой для математического ожидания наблюдаемой случайной величины.

Докажем несмещенность выборочного среднего как оценки для математического ожидания.

Пусть $x_1, x_2, \ldots, x_n$ – некоторая выборка, т. е. $x_1, x_2, \ldots, x_n$ – независимые СВ, имеющие одинаковое распределение. Обозначим $Mx_i = a$. Надо показать, что $M\overline{x} = a$. По свойствам математического ожидания,

$$M\overline{x} = M\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \frac{1}{n}\sum_{i=1}^{n}Mx_i = \frac{1}{n}\sum_{i=1}^{n}a = \frac{1}{n}na = a.$$

Состоятельность выборочного среднего как оценки математического ожидания следует из закона больших чисел. ◁

Утв. 2. Выборочная дисперсия $D_{\text{в}}$ является *состоятельной*, но *смещенной* оценкой дисперсии изучаемой случайной величины:

$$MD_{\text{в}} = \frac{n-1}{n}\sigma^2.$$

Докажем смещенность выборочной дисперсии как оценки для дисперсии. Пусть выборочные значения $x_1, x_2, \ldots, x_n$ – независимые СВ, имеющие одинаковое распределение с $Mx_i = a,\ Dx_i = \sigma^2$.

Покажем сперва, что для любого постоянного c справедливо равенство

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - c)^2 - n(\overline{x} - c)^2.$$

Действительно,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}\big((x_i - c) - (\overline{x} - c)\big)^2 =$$

$$= \sum_{i=1}^{n}\Big((x_i - c)^2 - 2(x_i - c)(\overline{x} - c) + (\overline{x} - c)^2\Big) =$$

$$= \sum_{i=1}^{n}(x_i - c)^2 - 2(\overline{x} - c)\sum_{i=1}^{n}(x_i - c) + \sum_{i=1}^{n}(\overline{x} - c)^2 =$$

$$= \sum_{i=1}^{n}(x_i - c)^2 - 2(\overline{x} - c)\left(\sum_{i=1}^{n}x_i - nc\right) + n(\overline{x} - c)^2 =$$

$$= \sum_{i=1}^{n}(x_i - c)^2 - 2(\overline{x} - c)(n\overline{x} - nc) + n(\overline{x} - c)^2 =$$

$$= \sum_{i=1}^{n}(x_i - c)^2 - 2n(\overline{x} - c)^2 + n(\overline{x} - c)^2 = \sum_{i=1}^{n}(x_i - c)^2 - n(\overline{x} - c)^2.$$

Таким образом, $D_{\text{в}} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2.$

Используя определение дисперсии СВ, а также свойства математического ожидания и дисперсии, имеем:

$$MD_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}M(x_i - a)^2 - M(\overline{x} - a)^2 = \frac{1}{n}\sum_{i=1}^{n}Dx_i - D\overline{x} =$$

$$= \frac{1}{n}\sum_{i=1}^{n} Dx_i - D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} Dx_i - \frac{1}{n^2}\sum_{i=1}^{n} Dx_i =$$

$$= \frac{1}{n} n\sigma^2 - \frac{1}{n^2} n\sigma^2 = \sigma^2\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}\sigma^2. \ \triangleleft$$

Итак, выборочная дисперсия $D_в$ является *смещенной* оценкой дисперсии изучаемой случайной величины (дает заниженное значение). В связи с этим вместо нее вводится другая статистика – ***исправленная выборочная дисперсия*** $s^2 = \frac{n}{n-1} D_в$, которая является ***несмещенной оценкой дисперсии.***

Утв. 3. Исправленная дисперсия s^2 является *несмещенной состоятельной* оценкой дисперсии. В случае выборки из нормального распределения s^2 является также *асимптотически эффективной* оценкой дисперсии.

§ 3. Интервальное оценивание параметров распределения

Точечные оценки не дают информации о степени близости оценки к истинному значению оцениваемого параметра. Чтобы получить информацию о точности и надежности оценки, используют интервальные оценки.

Опр. 1. ***Интервальной оценкой*** **(*доверительным интервалом*)** параметра θ называется интервал, границы которого $\hat{\theta}_1 = \hat{\theta}_1(x_1; x_2; \ldots; x_n)$ и $\hat{\theta}_2 = \hat{\theta}_2(x_1; x_2; \ldots; x_n)$ являются функциями выборочных значений и который с заданной вероятностью γ накрывает истинное значение оцениваемого параметра θ:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = \gamma.$$

Интервал $(\hat{\theta}_1;\ \hat{\theta}_2)$ называется ***доверительным интервалом***; число γ – ***доверительной вероятностью*** или ***надежностью*** интервальной оценки; значение $\alpha = 1 - \gamma$ – ***уровнем значимости***.

Величина доверительного интервала существенно зависит от объема выборки (уменьшается с ростом n, т. е. *чем больше объем выборки, тем более точную оценку можно получить*) и от доверительной вероятности γ (величина доверительного интервала увеличивается с приближением γ к 1, т. е. *чем более надежный вывод*

мы хотим получить, тем меньшую точность мы можем гарантировать).

Выбор доверительной вероятности определяется конкретными условиями. Обычно используются значения 0,90; 0,95; 0,99; 0,9973, т. е. такие, чтобы получить интервал, который с большой вероятностью накроет истинное значение оцениваемого параметра.

Приведем примеры доверительных интервалов для параметров нормального распределения. Для этого нам нужно знать, каким распределениям подчиняются статистики $\bar{x}$ и s^2. (Напомним, что запись $\xi \sim \mathcal{N}(a;\sigma)$ означает, что СВ ξ имеет нормальное распределение с $M\xi = a, D\xi = \sigma^2$. Тот факт, что СВ ξ имеет закон распределения $\mathcal{P}$, будем символически обозначать $\xi \sim \mathcal{P}$.)

Построение доверительного интервала для математического ожидания в случае выборки из нормального распределения с известной дисперсией σ^2

Утв. 1. Пусть имеется выборка объема n из нормального распределения с математическим ожиданием a и дисперсией σ^2, т. е. $x_1, x_2, \ldots, x_n \sim \mathcal{N}(a;\sigma)$. Тогда статистика $\bar{x}$ распределена по нормальному закону с параметрами a и $\frac{\sigma}{\sqrt{n}}$, а статистика $\frac{\bar{x}-a}{\sigma/\sqrt{n}}$ имеет стандартное нормальное распределение:

$$\bar{x} \sim \mathcal{N}\left(a;\frac{\sigma}{\sqrt{n}}\right); \qquad \frac{\bar{x}-a}{\sigma/\sqrt{n}} \sim \mathcal{N}(0;1).$$

Отметим, что

$$M\bar{x} = M\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} Mx_i = \frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n}na = a;$$

$$D\bar{x} = D\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n^2}\sum_{i=1}^{n} Dx_i = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Это означает, в частности, что $\bar{x}$ является более точной, чем одиночное наблюдение, оценкой для математического ожидания,

поскольку чем меньше дисперсия, т. е. разброс значений, тем точнее оценка.

Утв. 2. Доверительный интервал для математического ожидания a в случае выборки из нормального распределения с *известной* дисперсией σ^2 определяется соотношением

$$P\left(\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}} < a < \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \qquad (1)$$

где n – объем выборки; $\overline{x}$ – выборочное среднее; α – уровень значимости; u_α – *квантиль* нормального распределения уровня α, т. е. такое число, что для случайной величины $\xi \sim \mathcal{N}(0;1)$, имеющей стандартное нормальное распределение, $P(|\xi| \geq u_\alpha) = \alpha$.

Квантиль u_α определяется по таблице функции Лапласа из соотношения $\Phi(u_\alpha) = \dfrac{1-\alpha}{2}$.

Формула (1) означает, что при достаточно большом количестве выборок одного и того же объема n примерно в $100(1-\alpha)\%$ выборок интервал $\left(\overline{x} - u_\alpha \dfrac{\sigma}{\sqrt{n}}; \overline{x} + u_\alpha \dfrac{\sigma}{\sqrt{n}}\right)$ накрывает истинное значение математического ожидания a.

Доказательство. Из утверждения 1 следует, что

$$P\left(\left|\frac{\overline{x} - a}{\sigma / \sqrt{n}}\right| < u_\alpha\right) = 1 - \alpha;$$

$$P\left(-u_\alpha < \frac{a - \overline{x}}{\sigma / \sqrt{n}} < u_\alpha\right) = 1 - \alpha;$$

$$P\left(-u_\alpha \frac{\sigma}{\sqrt{n}} < a - \overline{x} < u_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha;$$

$$P\left(\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}} < a < \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \triangleleft$$

Выборочные распределения

С нормальным распределением связаны также следующие три наиболее часто использующиеся в статистике распределения: χ^2-распределение, *t*-распределение Стьюдента и *F*-распределение Фишера.

Опр. 2. ***Распределением*** χ^2 ***с*** *k* ***степенями свободы*** называется распределение суммы квадратов *k* независимых СВ, распределенных по нормальному закону с параметрами 0 и 1, т. е. если $\xi_1, \xi_2, \ldots, \xi_k \sim \mathcal{N}(0;1)$, то $\eta = \xi_1^2 + \xi_2^2 + \ldots + \xi_k^2 \sim \chi_k^2$.

Плотность распределения χ^2 с *k* степенями свободы имеет вид

$$f_{\chi_k^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}} & \text{при } x > 0, \\ 0 & \text{при } x \le 0. \end{cases}$$

Здесь $\Gamma\left(\dfrac{k}{2}\right)$ – значение ***гамма-функции*** которая определяется для $\alpha > 0$ формулой $\Gamma(\alpha) = \int\limits_0^{+\infty} x^{\alpha-1}\mathrm{e}^{-x}dx$ и обладает свойствами: 1) $\Gamma(1) = \Gamma(2) = 1$; 2) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$; 3) $\Gamma(\alpha) = (\alpha-1)!$, если α – натуральное; 4) $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$.

На рис. 4 изображены кривые плотностей распределения χ^2 при числе степеней свободы $k = 1, 2, 3, 5$.

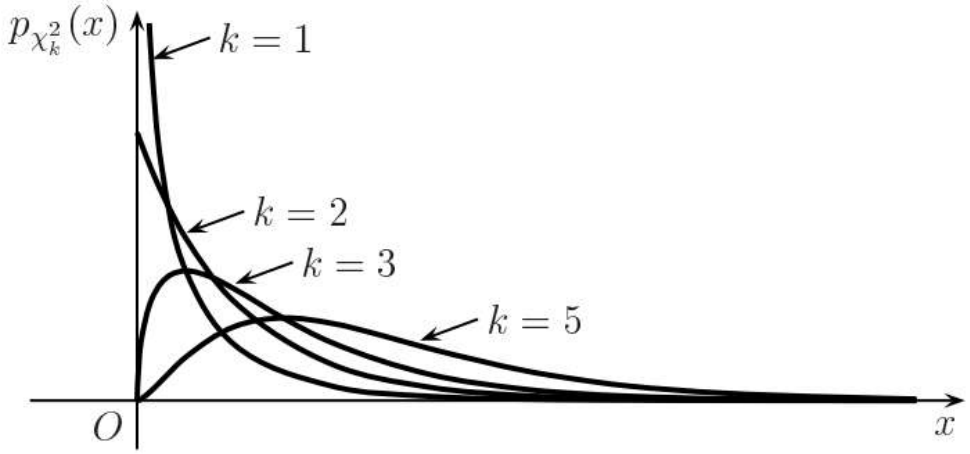

Рис. 4. Графики плотности распределения χ^2 с числом степеней свободы $k = 1, 2, 3, 5$

В частном случае при $k = 2$ распределение χ^2_k совпадает с показательным (экспоненциальным) распределением.

С ростом *k* распределение χ^2_k приближается к нормальному.

Считается, что при $k > 30$ оно практически не отличается от нормального.

Утв. 3. В случае выборки объема *n* из нормального распределения с *известным* математическим ожиданием статистика $\frac{nD_{\text{в}}}{\sigma^2}$ имеет χ^2-распределение с *n* степенями свободы:

$$\frac{nD_{\text{в}}}{\sigma^2} \sim \chi^2_n;$$

в случае выборки объема *n* из нормального распределения с *неизвестным* математическим ожиданием статистика $\frac{(n-1)s^2}{\sigma^2}$ имеет χ^2-распределение с числом степеней свободы $n-1$:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Доказательство. Докажем утверждение для случая известного математического ожидания. Пусть имеется выборка объема *n* из нормального рас-

пределения, т. е. СВ $x_1, x_2, \ldots, x_n$ независимы и $x_i \sim \mathcal{N}(a;\sigma)$. Тогда в силу свойств нормального распределения СВ $\dfrac{x_i - a}{\sigma} \sim \mathcal{N}(0;1)$.

Покажем, что в случае, когда математическое ожидание известно, статистика $\dfrac{nD_{\text{в}}}{\sigma^2}$ представляет собой сумму квадратов этих СВ. В этом случае для расчета $D_{\text{в}}$ используется известное значение математического ожидания: $D_{\text{в}} = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - a)^2$, поэтому

$$\frac{nD_{\text{в}}}{\sigma^2} = \frac{n\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - a)^2}{\sigma^2} = \frac{\sum\limits_{i=1}^{n}(x_i - a)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - a}{\sigma}\right)^2,$$

что в силу определения 2 доказывает, что $\dfrac{nD_{\text{в}}}{\sigma^2} \sim \chi^2_n$. ◁

Упражнение 1. Доказать утверждение 3 для случая неизвестного математического ожидания.

Замечание. В общем случае число степеней свободы, соответствующее той или иной оценке дисперсии, определяется как количество независимых наблюдений, по которым вычисляется данная оценка дисперсии, минус число параметров, которые оцениваются по этой выборке, кроме дисперсии. В случае, когда математическое ожидание неизвестно, одна степень свободы «расходуется» на вычисление оценки $\overline{x}$.

Опр. 3. ***Распределением Стьюдента с*** k ***степенями свободы*** называется распределение СВ $t = \dfrac{\xi}{\sqrt{\eta / k}} \sim t_k$, где СВ $\xi \sim \mathcal{N}(0;1)$ и $\eta \sim \chi^2_k$ независимы.

Уи́льям Си́ли Го́ссет
(англ. *William Sealy Gosset*)
(1876–1937)
британский химик и статистик, работавший на пивоваренном заводе «Гиннесс» (Arthur Guinness Son & Co). Один из основоположников теории статистических оценок и проверки гипотез.

«Гиннесс» был передовым предприятием, ориентированным на использование новейших достижений науки для принятия экономических и технических решений, благодаря чему Госсет имел полную свободу в проведении научных исследований. В частности, для контроля качества продукции необходимо было понять, как с точки зрения математики можно обосновать достоверность выводов, полученных при исследовании малой выборки, и правомерность применения этих выводов к выборке большой. В результате этих исследований Госсет разработал математическое обоснование «закона ошибок» для малых статистических выборок.

Госсет более известен под своим псевдонимом Стьюдент (Student), поскольку по условиям контракта с корпорацией «Гиннесс» не имел права открыто публиковать результаты своих исследований (таким способом охранялась коммерческая тайна, «ноу-хау» в виде вероятностно-статистических методов, разработанных Госсетом).

Первым, кто понял значение работ Госсета по оценке параметров малой выборки, был английский статистик Р. Э. Фишер (1890–1962), считавший, что Госсет совершил «логическую революцию» в математической статистике.

W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W

Плотность распределения Стьюдента с k степенями свободы имеет вид

$$f_{t_k}(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}};$$

ее график представлен на рис. 5.

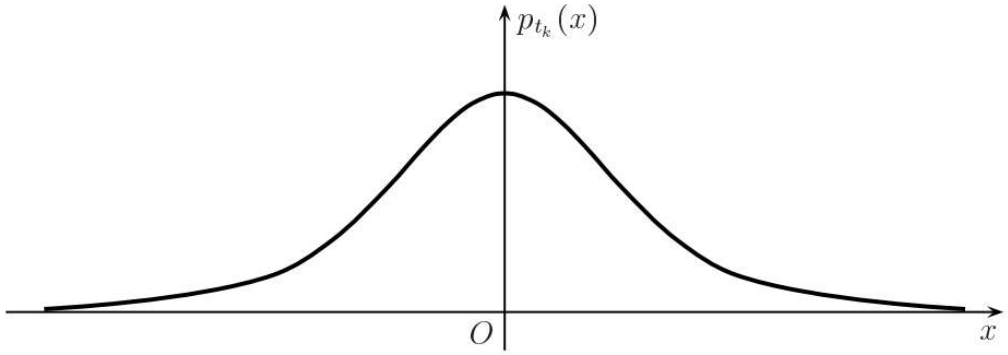

Рис. 5. График плотности распределения Стьюдента

При $k \to \infty$ график приближается к графику плотности нормального распределения. Практически уже при $k > 30$ можно считать *t*-распределение приближенно нормальным.

Утв. 4. В случае выборки объема *n* из нормального распределения с *неизвестной* дисперсией статистика $\dfrac{\overline{x} - a}{\sqrt{s^2 / n}}$ имеет распределение Стьюдента с числом степеней свободы $n - 1$:

$$\frac{\overline{x} - a}{\sqrt{s^2 / n}} \sim t_{n-1}.$$

Упражнение 2. Сравнить с формулировкой утверждения 1.

Доказательство. Пусть имеется выборка объема *n* из нормального распределения, т. е. СВ $x_1, x_2, \ldots, x_n$ независимы и $x_i \sim \mathcal{N}(a; \sigma)$. Тогда в силу утверждений 1 и 3 СВ $\xi = \dfrac{\overline{x} - a}{\sigma / \sqrt{n}} \sim \mathcal{N}(0; 1)$, СВ $\eta = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, причем можно доказать также, что эти СВ независимы.

Представив рассматриваемую статистику в виде

$$\frac{\overline{x} - a}{\sqrt{s^2 / n}} = \frac{\dfrac{\overline{x} - a}{\sigma / \sqrt{n}}}{\dfrac{\sqrt{s^2 / n}}{\sigma / \sqrt{n}}} = \frac{\dfrac{\overline{x} - a}{\sigma / \sqrt{n}}}{\sqrt{\dfrac{s^2}{\sigma^2}}} = \frac{\xi}{\sqrt{\dfrac{\eta}{n-1}}},$$

согласно определению 3 заключаем, что она подчиняется распределению Стьюдента с числом степеней свободы $n - 1$. ◁

Опр. 4. ***F*-распределением Фишера с числами степеней свободы** k_1 ***и*** k_2 называется распределение СВ

$$F = \frac{\eta_1 / k_1}{\eta_2 / k_2} \sim F_{k_1;k_2},$$

где СВ $\eta_1 \sim \chi^2_{k_1}$ и $\eta_2 \sim \chi^2_{k_2}$ независимы.

W W W В И К И С П Р А В К А W W W W W W W W W W W W W W W W W W W W W W W W W

Ро́нальд Э́йлмер Фи́шер
(англ. *Ronald Aylmer Fisher*)
(1890–1962)

английский статистик, биолог-эволюционист и генетик, «отец современной статистики», один из основоположников математической генетики, по образованию математик и физик-теоретик.

С его именем связаны многие понятия математической статистики, он построил теорию точечных и интервальных статистических оценок, внес существенный вклад в создание современной теории проверки статистических гипотез, положил начало использованию статистических процедур при планировании научного эксперимента. Фундаментальный в теории вероятностей термин «дисперсия» также был введен Фишером в 1916 г.

Большинство методов Фишера имеют общий характер и применяются в естественных науках, в экономике и в других областях деятельности. Его книга «Статистические методы для исследователей», опубликованная в 1925 г., переиздавалась в течение 50 лет.

F-распределение исследовалось и было названо в честь Фишера его учеником Джорджем Снедекором (1881–1974), хотя сам Фишер рассматривал величину $\zeta = \frac{1}{2}\ln F$, распределение которой сейчас называют *z*-распределением Фишера. В современной статистической практике предпочитают использовать *F*-распределение, имеющее более простые свойства.

W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W W

Плотность распределения Фишера с k_1 и k_2 степенями свободы имеет вид

$$f_{F_{k_1;k_2}}(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{k_1+k_2}{2}\right)}{\Gamma\left(\dfrac{k_1}{2}\right)\Gamma\left(\dfrac{k_2}{2}\right)}\left(\dfrac{k_1}{k_2}\right)^{\frac{k_1}{2}}x^{\frac{k_1}{2}-1}\left(1+\dfrac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}} & \text{при } x>0,\\ 0 & \text{при } x\le 0.\end{cases}$$

На рис. 6 представлены кривые плотностей распределения Фишера в зависимости от значений k_1 и k_2.

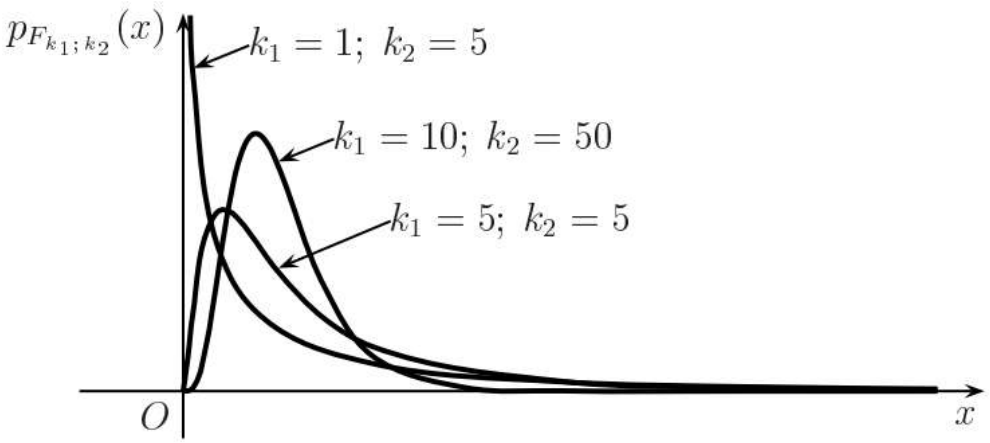

Рис. 6. Графики плотностей распределения Фишера в зависимости от значений k_1 и k_2

Замечание. Существуют таблицы распределений χ^2, Стьюдента, Фишера. Однако при использовании статистических таблиц необходимо обращать пристальное внимание на то, какие величины затабулированы и какие нужны при вычислениях.

Построение доверительного интервала для математического ожидания в случае выборки из нормального распределения с неизвестной дисперсией

Утв. 5. Доверительный интервал для математического ожидания a в случае выборки из нормального распределения с *неизвестной* дисперсией σ^2 определяется соотношением

$$P\left(\bar{x}-t_{\alpha;n-1}\frac{s}{\sqrt{n}}<a<\bar{x}+t_{\alpha;n-1}\frac{s}{\sqrt{n}}\right)=1-\alpha,$$

где n – объем выборки; $\bar{x}$ – выборочное среднее; s^2 – несмещенная оценка дисперсии; α – уровень значимости; $t_{\alpha;n-1}$ – *квантиль* уровня α распределения Стьюдента с числом степеней свободы $k = n-1$, т. е. такое число, что для случайной величины ξ, имеющей распределение Стьюдента с числом степеней свободы $k = n-1$, имеет место $P(|\xi| \geq t_{\alpha;\, n-1}) = \alpha$.

Упражнение 3. Доказать аналогично доказательству утверждения 2.

Квантиль $t_{\alpha;\, n-1}$ определяется по таблице распределения Стьюдента.

При малых выборках ($n < 30$) распределение Стьюдента дает не вполне определенные результаты (широкий доверительный интервал). Это объясняется тем, что малая выборка содержит малую информацию об интересующем нас признаке. С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормальному.

Построение доверительного интервала для дисперсии в случае выборки из нормального распределения с неизвестным математическим ожиданием

Утв. 3. Доверительный интервал для дисперсии σ^2 в случае выборки из нормального распределения с *неизвестным* математическим ожиданием a определяется соотношением

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2;\, n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;\, n-1}} \right) = 1-\alpha,$$

где n – объем выборки; s^2 – несмещенная оценка дисперсии; α – уровень значимости; $\chi^2_{\alpha/2;\, n-1}$ и $\chi^2_{1-\alpha/2;\, n-1}$ – *квантили* распределения χ^2 с числом степеней свободы $k = n-1$, определяемые соотношением $P(\xi \geq \chi^2_{\alpha;\, n-1}) = \alpha$ для случайной величины ξ, имеющей распределение χ^2 с числом степеней свободы $k = n-1$.

Квантили $\chi^2_{\alpha/2;\, n-1}$ и $\chi^2_{1-\alpha/2;\, n-1}$ определяются по таблице распределения χ^2.

Доказательство. Из утверждения 3 следует, что для выборки из нормального распределения с неизвестным математическим ожиданием

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. В силу несимметричности графика плотности распределения χ^2 для построения доверительного интервала будут использованы две квантили $\chi^2_{\alpha/2;\,n-1}$ и $\chi^2_{1-\alpha/2;\,n-1}$ (см. рис. 7), такие, что для СВ $\xi \sim \chi^2_{n-1}$ имеют место соотношения $P(\xi \geq \chi^2_{\alpha/2;\,n-1}) = \frac{\alpha}{2}$ и $P(\xi \leq \chi^2_{1-\alpha/2;\,n-1}) = \frac{\alpha}{2}$.

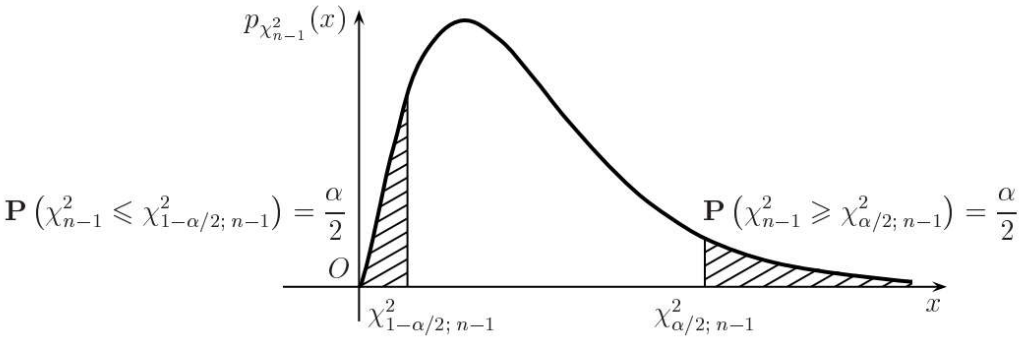

Рис. 7. К построению доверительного интервала для дисперсии в случае выборки из нормального распределения с неизвестным математическим ожиданием

Тогда

$$P\left(\chi^2_{1-\alpha/2;n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2;n-1}\right) = 1-\alpha;$$

$$P\left(\frac{1}{\chi^2_{\alpha/2;n-1}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{1-\alpha/2;n-1}}\right) = 1-\alpha;$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2;\,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;\,n-1}}\right) = 1-\alpha. \triangleleft$$

§ 4. Проверка статистических гипотез

Основные задачи математической статистики разделяют на две категории, тесно связанные между собой, но отличающиеся постановкой задач: оценивание параметров и проверка статистических гипотез. Основной задачей оценивания параметров являет-

ся получение по выборке оценок, наилучших в том или ином смысле. При проверке гипотез задача ставится иначе: требуется по выборке принять или отвергнуть некоторое предположение о распределении генеральной совокупности, из которой извлечена выборка.

Опр. 1. ***Статистической гипотезой*** называется любое предположение о виде (***непараметрическая гипотеза***) или параметрах (***параметрическая гипотеза***) неизвестного распределения.

Опр. 2. Статистическая гипотеза называется ***простой***, если она полностью определяет функцию распределения. В противном случае гипотеза называется ***сложной***.

Пример 1. Предположим, что введен новый способ производства некоторого товара. Для определения качества товара измеряется некоторая его характеристика $\xi \sim \mathcal{N}(a_0; \sigma_0)$, 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;

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

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

Гипотеза H_1 является простой, а гипотезы H_2 и H_3 – сложными. •