Математическая статистика

Задачи математической статистики

- Оценка неизвестной функции распределения.
- Оценка неизвестных параметров распределения.
- Статистическая проверка гипотез.

Выборочный метод. Генеральная совокупность. Выборка

- ***Опр.*** Исследуемая совокупность N объектов наз. генеральной совокупностью (N - очень велико, в некоторых случаях количество значений, образующих генеральную совокупность, можно считать и бесконечным).

- ***Опр***. Совокупность объектов n, отобранных случайным образом из генеральной совокупности наз. выборочной совокупностью (выборкой), где $n << N$.

- Число n наз. объемом выборки.

- Метод основанный на том, что по выборочной совокупности выделенной из данной генеральной совокупности делается заключение о всей генеральной совокупности наз. **выборочным методом**

Виды выборок

Собственно-случайная

- Выборка образованная случайным выбором элементов без расчленения на части или группы.

Механическая

- Выборка, в которую элементы из генеральной совокупности отбираются через определенный интервал. Например, если объем выборки должен составлять 10% (10%-я выборка), то отбирается каждый 10-й элемент.

Типическая

- Выборка, в которую случайным образом отбираются элементы из типических групп, на которые по некоторому признаку разбивается генеральная совокупность.

Серийная

- Выборка, в которую случайным образом отбираются не элементы, а целые группы совокупности(серии), а сами серии подвергаются сплошному наблюдению.

Способы образования выборки

Повторный отбор

- Каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть повторно отобран.

Бесповторный

- Отобранный элемент не возвращается в общую совокупность

Статистический ряд. Статистическое распределение. Эмпирическая функция распределения

- Варианты:

$$x_1, x_2, x_3, \ldots, x_n.$$

- Вариационный ряд:

$$x_1 < x_2 < x_3 < \ldots < x_n$$

- или

$$x_1 > x_2 > x_3 > \ldots > x_n.$$

- Из генеральной совокупности извлечена выборка объема n:
- x_1 наблюдалась n_1 раз;
- x_2 наблюдалась n_2 раза;
- x_3 наблюдалась n_3 раза;
- .....
- x_k наблюдалась n_k раз.
- Причем $\sum_{i=1}^{k} n_i = n$.

- Числа

$$n_1, n_2, \ldots, n_k$$

называются <u>частотами</u>.

- Числа $w_i = \frac{n_i}{n}$, где $i = 1,2,\ldots,k$

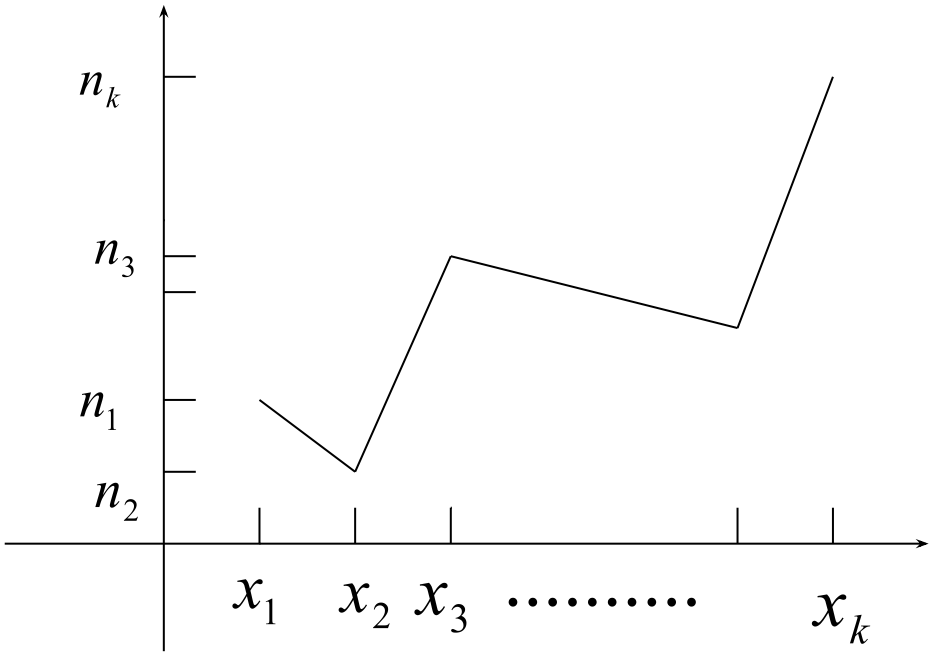
наз. <u>относительными частотами</u>.

Статистическое распределение выборки

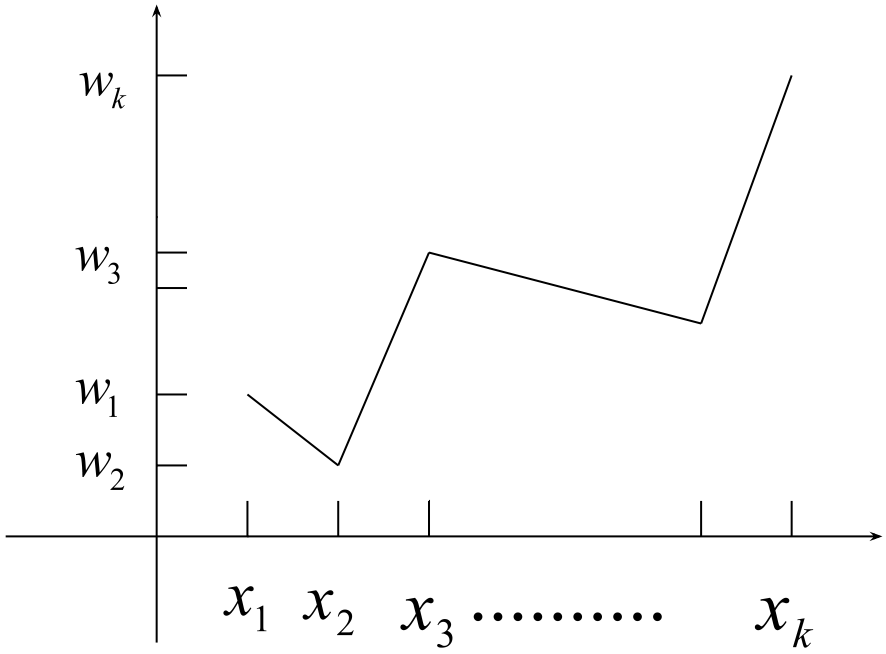
x_1	x_2	x_3	.....	x_k
n_1	n_2	n_3	.....	n_k

$$\sum_{i=1}^{k} n_i = n$$

Полигон частот


n_k
n_3
n_1
n_2
x_1 x_2 x_3 ..... x_k

Полигон относительных частот


w_k
w_3
w_1
w_2
x_1 x_2 x_3 ..... x_k

Эмпирическая функция распределения

- Эмпирическая функция распределения это функция равная отношению числа вариант, меньших x , к объему выборки:

.

$$F^*(x) = \frac{n(x)}{n}$$

Свойства эмпирической функции распределения

- 1) $0 \le F^*(x) \le 1;$
- 2) $F^*(x)$ - неубывающая;
- 3) если x_1 наименьшая варианта, то $F^*(x) = 0,$ при $x \le x_1;$

  4) если x_k наибольшая варианта, то $F^*(x) = 1,$ при $x > x_k.$

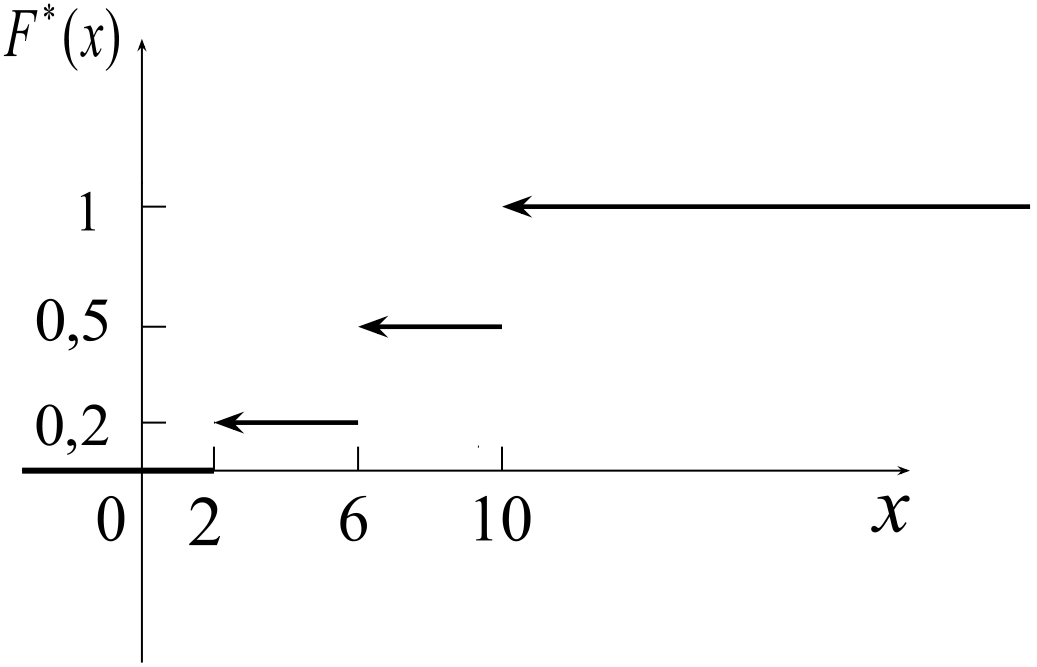
Пример.
По данному распределению выборки построить эмпирическую функцию.

x_i	2	6	10
n_i	12	18	30

$$n = \sum_{i=1}^{3} n_i = 60$$

$$F^*(x) = \begin{cases} 0, & x \le 2; \\ \dfrac{12}{60}, & 2 < x \le 6; \\ \dfrac{12+18}{60}, & 6 < x \le 10;. \\ 1, & x > 10. \end{cases}$$

$$F^*(x)=\begin{cases}0, & x\le 2;\\ 0{,}2, & 2<x\le 6;\\ 0{,}5, & 6<x\le 10;\\ 1, & x>10.\end{cases}$$


$F^*(x)$
1
0,5
0,2
0
2
6
10
x

Статистическая совокупность

$[x_0; x_1]$	$(x_1; x_2]$	$(x_2; x_3]$	.....	$(x_{k-1}; x_k]$
n_1	n_2	n_3	.....	n_k

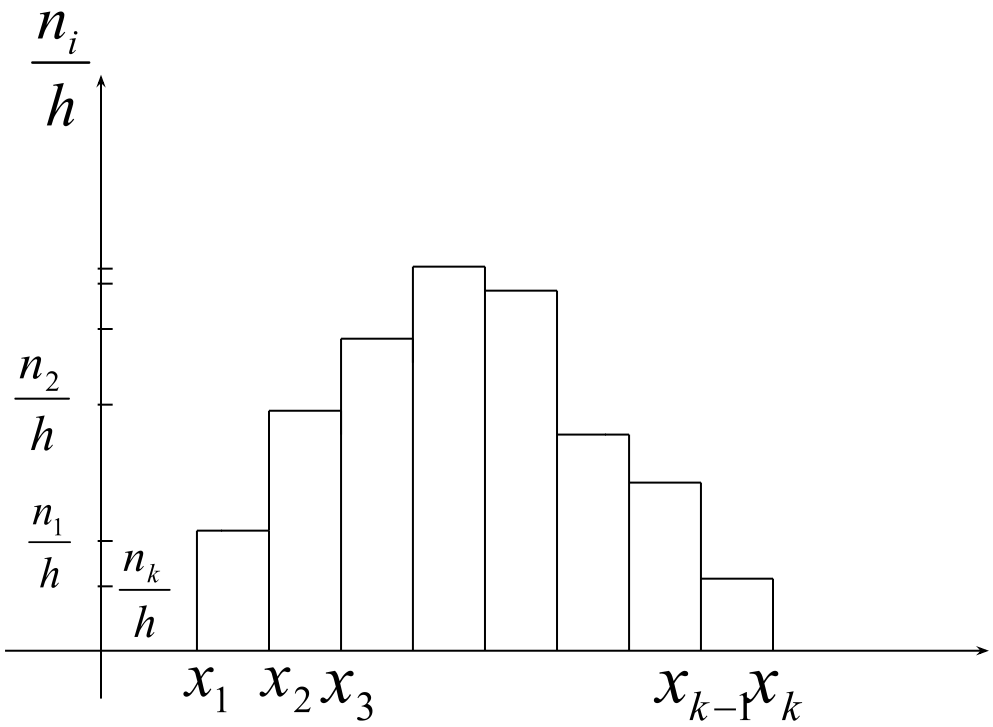
$$h = x_1 - x_0 = x_2 - x_1 = ... = x_k - x_{k-1}$$

- Число интервалов определяется по формуле Стерджеса

$$k = 1 + 3{,}22 \cdot \lg n$$

Гистограмма частот

- Ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частот).


$\frac{n_i}{h}$
$\frac{n_2}{h}$
$\frac{n_1}{h}$
$\frac{n_k}{h}$
x_1 x_2 x_3
x_{k-1} x_k

- Площадь гистограммы частот

$$S = \sum_{i=1}^{k} \Delta S_i ,$$

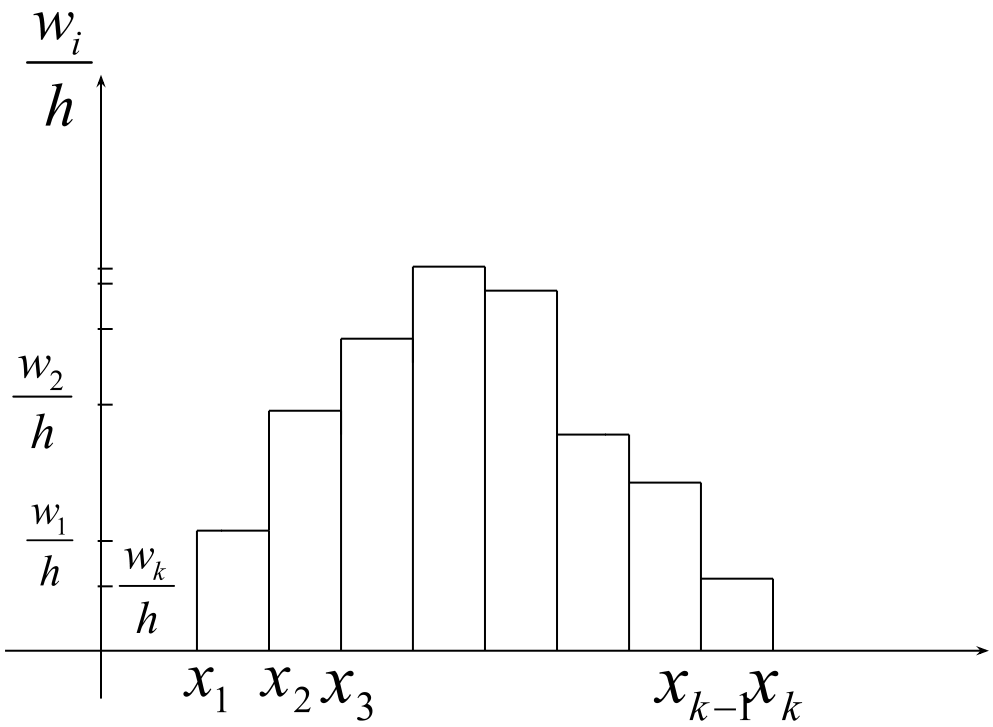
тогда

$$\Delta S_i = \frac{n_i}{h} \cdot h = n_i ,$$

$$S = \sum_{i=1}^{k} n_i = n.$$

Гистограмма относительных частот

- Ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{w_i}{h}$ (плотность относительных частот).


$\frac{w_i}{h}$
$\frac{w_2}{h}$
$\frac{w_1}{h}$
$\frac{w_k}{h}$
x_1 x_2 x_3
x_{k-1} x_k

Площадь гистограммы относительных частот

$$S = \sum_{i=1}^{k} \Delta S_i ,$$

$$\Delta S_i = \frac{w_i}{h} \cdot h = w_i ,$$

тогда

$$S = \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{\sum_{i=1}^{k} n_i}{n} = \frac{n}{n} = 1.$$

Статистические оценки параметров распределения

Точечные оценки

- Оценка, которая определяется одним число, наз. ***точечной.***

Интервальные оценки

- Оценка, которая определяется двумя числами, являющимися концами интервала, содержащего неизвестный параметр, называется **интервальной**.

Свойства точечных оценок

Несмещенность

- Статистическая оценка θ^* наз. несмещенной, если её математическое ожидание равно оцениваемому параметру θ при любом объеме выборки:

$$M(\theta^*) = \theta.$$

Эффективность

- Статистическая оценка θ^* наз. эффективной, если она имеет наименьшую возможную дисперсию.

Состоятельность

- Статистическая оценка θ^* наз. состоятельной, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру θ :

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

- Теорема. Если дисперсия несмещенной оценки при $n \to \infty$ стремится к нулю, то такая оценка состоятельна.
- Док-во: Оценка θ^* параметра θ несмещенная, т.е. $M\theta^* = \theta$, поэтому при $\forall \varepsilon$ из неравенства Чебышева

$$P\left(\left|\theta^* - M\theta^*\right| < \varepsilon\right) \geq 1 - \frac{D\theta^*}{\varepsilon^2}$$

следует

$$P\left(\left|\theta^* - \theta\right| < \varepsilon\right) \geq 1 - \frac{D\theta^*}{\varepsilon^2}.$$

Но $D\theta^* \to 0$ при $n \to \infty$.

Значит при $n \to \infty$, для каждого фиксированного ε :

$$\frac{D\theta^*}{\varepsilon^2} \to 0,$$

а $$1 - \frac{D\theta^*}{\varepsilon^2} \to 1.$$

Но тогда $P\left(\left|\theta^* - \theta\right| < \varepsilon\right) \to 1$ при $n \to \infty$.

Генеральная средняя

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

или

$$\overline{x}_г = \frac{\sum_{i=1}^{k} x_i \cdot N_i}{N}.$$

Выборочная средняя

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

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

Генеральная дисперсия

$$D_г = \frac{\sum_{i=1}^{N} (x_i - \overline{x}_г)^2}{N}$$

или

$$D_г = \frac{\sum_{i=1}^{k} (x_i - \overline{x}_г)^2 \cdot N_i}{N}.$$

Выборочная дисперсия

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

$$D_{в} = \frac{\sum_{i=1}^{k} \left(x_i - \overline{x}_{в} \right)^2 \cdot n_i}{n},$$

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

- Выборочная средняя является несмещенной и состоятельной:

1.Рассмотрим выборочную среднюю, как случайную величину

$$\overline{X_в} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$M(X_1) = M(X_2) = \ldots = M(X_n) = M(X) = \overline{x}_г,$$

$$D(X_1) = D(X_2) = \ldots = D(X_n) = D(X) = \sigma_г^2.$$

$$M(\overline{X}_в) = M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} \cdot M\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) =$$

$$= \frac{1}{n} \cdot n \cdot M(X) = M(X) = \overline{x}_г,$$

- т.е.

$$M(\overline{x}_в) = \overline{x}_г.$$

$$D(\overline{X}_в) = D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} \cdot D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} D(X_i) =$$

$$= \frac{1}{n^2} \cdot n \cdot D(X) = \frac{1}{n} \cdot D(X) = \frac{\sigma_г^2}{n}.$$

2.Используем неравенство Чебышева:

$$P\left(\left|X - m\right| < \varepsilon\right) \geq 1 - \frac{D}{\varepsilon^2}.$$

$$P\left(\left|\overline{X_в} - M(\overline{X_в})\right| < \varepsilon\right) \geq 1 - \frac{D(\overline{X_в})}{\varepsilon^2};$$

$$P\left(\left|\overline{X_в} - \overline{x_г}\right| < \varepsilon\right) \geq 1 - \frac{\sigma_г^2}{n \cdot \varepsilon^2}.$$

Пусть $n \to \infty$ тогда $P\left(\left|\overline{X_в} - \overline{x_г}\right| < \varepsilon\right) \to 1,$

т.е. $\lim\limits_{n\to\infty} P\left(\left|\overline{X_в} - \overline{x_г}\right| < \varepsilon\right) = 1.$

- Значит выборочная средняя является статистической оценкой генеральной средней.

- Выборочная дисперсия является смещенной оценкой:

$$M(D_в) = M\left(\frac{1}{n}\cdot\sum(X_i - \overline{x}_г)^2 - (\overline{X}_в - \overline{x}_г)^2\right) =$$

$$= \frac{1}{n}\cdot\sum M(X_i - \overline{x}_г)^2 - M(\overline{X}_в - \overline{x}_г)^2 =$$

$$= \frac{1}{n} \cdot \sum M(X_i - M(X_i))^2 - M(\overline{X}_в - M(\overline{X}_в))^2 =$$

$$\frac{1}{n}\cdot n\cdot D(X)-D(\overline{X}_в)=\sigma_г^2-\frac{\sigma_г^2}{n}=\frac{n-1}{n}\cdot\sigma_г^2=\frac{n-1}{n}\cdot D_г\neq D_г.$$

$$M(D_в)\neq D_г.$$

- Несмещенная оценка генеральной дисперсии - исправленная выборочная дисперсия:

$$S^2 = \frac{n}{n-1} \cdot D_в .$$

Статистические характеристики

Мода

$$M_0 = x_k + \frac{(n_k - n_{k-1}) \cdot h}{(n_k - n_{k-1}) + (n_k - n_{k+1})}$$

Медиана

$$M_e = x_i + h \cdot \frac{\frac{n}{2} - T_{i-1}}{n_i}.$$

Асимметрия

- Асимметрия распределения характеризуется тем, что вариант, меньших и больших моды неодинаковое число.

$$A = \frac{\mu_3}{\sigma_в^3}, \qquad \mu_3 = \frac{\sum n_i \cdot (x_i - \overline{x}_в)^3}{n}.$$

- При $M_0 < \bar{x} \Rightarrow A > 0 \Rightarrow$

асимметрия положительная;

При $M_0 > \bar{x} \Rightarrow A < 0 \Rightarrow$

асимметрия отрицательная.

- Если $|A| < 0{,}1$, то распределение почти симметрично;

если $|A| > 0{,}5$, то распределение сильно асимметрично.

Эксцесс

- Эксцесс характеризует крутовершинность кривой распределения.

$$E = \frac{\mu_4}{\sigma_в^4} - 3, \qquad \mu_4 = \frac{\sum n_i \cdot (x_i - \overline{x}_в)^4}{n}.$$

- Если $|E| < 0{,}1$ то распределение считается близким к нормальному;

- если $|E| > 0{,}5$ то распределение значительно отклоняется от нормального.

Метод произведений

u_i-условные варианты,

$$u_i = \frac{x_i - C}{h},$$ C -условный нуль.

$$x_i = u_i \cdot h + C,$$

$$\overline{x}_в = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_i \cdot (C + u_i \cdot h) = C \cdot \frac{\sum_{i=1}^{k} n_i}{n} + h \cdot \frac{\sum_{i=1}^{k} n_i \cdot u_i}{n} =$$

$$= C + h \cdot M_1^*,$$

$$M_k^* = \frac{\sum_{i=1}^{n} n_i \cdot u_i^k}{n}.$$

$$\overline{x}_в = C + h \cdot M_1^*$$

$$D_в = (M_2^* - (M_1^*)^2) \cdot h^2$$

$$\mu_3 = h^3 \cdot (M_3^* - 3M_1^* \cdot M_2^* + 2(M_1^*)^3),$$

$$\mu_4 = h^4 \cdot (M_4^* - 4M_1^* \cdot M_3^* + 6M_2^* \cdot (M_1^*)^2 - 3(M_1^*)^4).$$

Статистическая проверка статистических гипотез

- ***Нулевая гипотеза*** (H_0) - выдвинутая гипотеза.

- ***Конкурирующая гипотеза*** (H_1) -

- гипотеза, которая противоречит нулевой гипотезе.

Простая гипотеза – гипотеза, содержащая одно предположение:

$$H_0: \quad \lambda = 5,$$

где λ – *параметр распределения Пуассона.*

Сложная гипотеза – гипотеза, которая состоит из конечного или бесконечного числа простых гипотез:

$$H_0: \quad \lambda > 5,$$

$$где\ \lambda - параметр \quad распределения \quad Пуассона.$$

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

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

- ***Уровень значимости*** (α) – вероятность совершить ошибку первого рода.

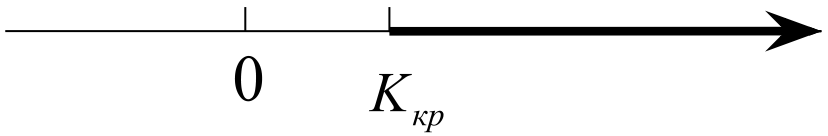
- ***Статистический критерий*** (K) - случайная величина, которая служит для проверки нулевой гипотезы.

- ***Наблюдаемым значением*** $(K_{набл})$ - значение критерия, вычисленное по выборке.

- ***Критическая область*** – совокупность значений критерия, при которых нулевую гипотезу отвергают.
- ***Область принятия гипотезы*** - совокупность значений критерия, при которых нулевую гипотезу принимают.
- ***Критические точки*** $(K_{кр})$ - точки, отделяющие критическую область от области принятия гипотезы.

- ***Правосторонняя критическая область*** – критическая область определяющаяся неравенством:

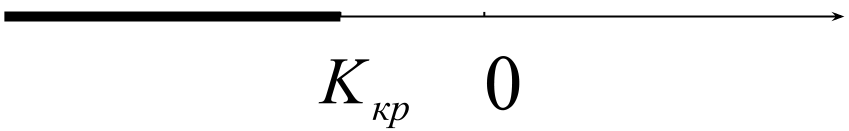
$$K > K_{кр}, \quad K_{кр} > 0$$



$K_{кр}$ ищут, исходя из требования чтобы

$$P(K > K_{кр}) = \alpha.$$

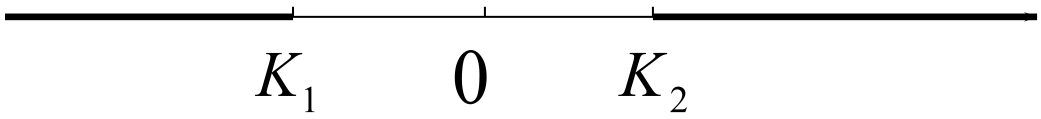
- ***Левосторонняя критическая область*** – критическая область, определяющаяся неравенством: $K < K_{кр}, \quad K_{кр} < 0.$



$K_{кр}$ ищут, исходя из требования чтобы

$$P(K < K_{кр}) = \alpha.$$

- ***Двусторонняя критическая область*** – критическая область, определяющаяся неравенством: $K < K_1$, $K > K_2$.



K_1, K_2 ищут, исходя из требования чтобы

$$P(K < K_1) + P(K > K_2) = \alpha.$$

- Если распределение критерия симметрично относительно 0 и имеются основания выбрать симметричные относительно нуля точки: $-K_{кр}$ *и* $K_{кр}$ $(K_{кр} > 0)$, то

$$P(K < -K_{кр}) = P(K > K_{кр}).$$

Тогда $P(K < K_1) + P(K > K_2) = \alpha$

заменится

$$P(K < -K_{кр}) + P(K > K_{кр}) = \alpha$$

или

$$P(K > K_{кр}) = \alpha / 2.$$

- ***Доверительная вероятность (надежность)***- вероятность с которой осуществляется неравенство $\left|\theta - \theta^*\right| < \delta$, т.е.

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

- ***Доверительный интервал*** – интервал, который покрывает неизвестный параметр θ с заданной надежностью γ .

Доверительный интервал для оценки математического ожидания нормального распределения при известном σ.

$$\bar{x}_в - \frac{t \cdot \sigma}{\sqrt{n}} < a < \bar{x}_в + \frac{t \cdot \sigma}{\sqrt{n}}$$

Число t определяется из равенства

$$\Phi(t) = \frac{\gamma}{2}.$$

Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном σ.

$$\bar{x}_в - \frac{t_\gamma \cdot S}{\sqrt{n}} < a < \bar{x}_в + \frac{t_\gamma \cdot S}{\sqrt{n}}$$

Число t_γ определяется по таблице

$$t_\gamma = t(\gamma, n).$$

- ***Критерий согласия*** – критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

- ***Критерии согласия:*** χ^2 (хи квадрат) Пирсона, Колмогорова, Смирнова и др.

Проверка гипотезы о нормальном распределении генеральной совокупности

Критерий Пирсона

- В качестве критерия проверки H_0 примем случайную величину

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

где n_i -эмпирические частоты;

n_i' -теоретические частоты.

- Строим правостороннюю критическую область, исходя из требования, что

$$P(\chi^2 > \chi^2_{кр}(\alpha; k)) = \alpha$$

в предположении справедливости H_0,

где α - уровень значимости;

k - число степеней свободы.

- Число степеней свободы находят по формуле $k = s - r - 1,$

где s - число групп(частичных интервалов) выборки;

r - число параметров предполагаемого распределения, которые оценены по данным выборки.

Если предполагаемое распределение нормальное, то оценивают два параметра и тогда $k = s - 2 - 1,\ k = s - 3.$

- Если обозначить $\chi^2_{набл} = \sum \frac{(n_i - n'_i)^2}{n'_i}$, то

при $\chi^2_{набл} < \chi^2_{кр}$ гипотезу H_0 принимают;

при $\chi^2_{набл} > \chi^2_{кр}$ гипотезу H_0 отвергают.

Критерий согласия Колмогорова

- Если функция распределения

$$F(x)$$

случайной величины X непрерывна, то

практически ее эмпирическая функция

$$F^*(x)$$

распределения при $\eta \to \infty$ сходится к .

$$F(x)$$

- Если $F(x)$ непрерывна, то функция

распределения величины

$$D_n(D_n = \max\left|F_n(x) - F(x)\right| \cdot \sqrt{n})$$

при $n \to \infty$ имеет пределом функцию

$$K(\lambda) = \sum (-1)^k \cdot e^{-2k^2\lambda^2},$$

которая не зависит от вида функции

$$F(x)$$

- По таблице найдем значение функции $K(\lambda)$ и затем значение функции

$$P(\lambda) = 1 - K(\lambda) = \beta.$$

Если $\beta \to 1$, то расхождение между эмпирическими и теоретическими функциями распределения несущественно, если $\beta \to 0$, то расхождение существенно.

Сравнение двух дисперсий нормальных генеральных совокупностей

- В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий примем случайную величину , причем отношение большей исправленной дисперсии к меньшей:

$$F = \frac{S_б^2}{S_м^2}.$$

- Величина F при условии справедливости H_0 имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 - объем выборки, по которой вычислена большая исправленная дисперсия.

Элементы теории корреляции

Основные задачи теории корреляции

О форме корреляционной связи между X и Y в виде некоторой функциональной зависимости, которая хотя бы приближенно изображала расплывчатую корреляционную зависимость.

Об оценке тесноты корреляционной связи между X и Y, т.е. о степени близости корреляционной зависимости к функциональной.

Регрессии

- Регрессией Y от X называется функциональная зависимость между значениями x и соответствующими условными средними значениями $y(x)$.
- Регрессии можно представить геометрически в виде ломанных линий, соединяющих или точки $A\,(\,x\,;\,y(x))$, или точки $B\,(x(y)\,;\,y\,)$.

- Эти линии называются эмпирическими (полученными из опыта) ломаными линиями регрессии.
- Плавную кривую можно получить и иначе, – если ломаную линию регрессии “сгладить” посредством какой-либо известной линии (прямой, параболы, гиперболы и т.п.).

- Уравнение сглаживающей линии даст хотя и приближенно, но аналитическое – в виде формулы – выражение регрессии. Подобные формулы называют эмпирическими

Задача отыскания эмпирической формулы распадается на две

- 1. Выбор типа линии, выравнивающей ломанную регрессии, т.е. типа линии, около которой группируются экспериментальные точки $A\ (x\ ;\ \overline{y}(x))$ или $B\ (\overline{x}(y)\ ;\ y\ )$.

- 2. Определение параметров, входящих в уравнение линии выбранного типа, таким образом, чтобы из множества линий этого типа взять ту, которая наиболее близко проходит около точек ломаной регрессии.

Выбор типа линии, выравнивающей ломаную линию регрессии

- Для выбора типа линии, выравнивающей ломаную линию регрессии, необходимо хорошо знать простейшие виды линий и их уравнения.

Определения параметров в уравнении выравнивающей линии выбранного типа

- Метод средних применяют в тех случаях, когда выбранный тип уравнения выравнивающей линии содержит лишь один параметр.

- Метод проб используют, когда выбранная формула содержит несколько параметров .

- Метод выровненных (или выбранных) точек состоит в выборе по чертежу нескольких точек (не обязательно совпадающих с точками линии регрессии), через которые проводят выравнивающую линию и определяют ее уравнение по координатам этих выбранных точек.
- Метод наименьших квадратов служит для оценки неизвестных величин по результатам измерений, содержащим случайные погрешности.

Метод наименьших квадратов

- Необходимо минимизировать сумму

$$S = \sum_{i=1}^{n} (\overline{y}(x_i) - y_i)^2$$

где x_i , y_i – значения опытных данных;

$\overline{y}(x_i)$ – значение функции, взятое из эмпирической зависимости в точке x_i;

n – число опытов.

- В случае линейной эмпирической формулы сумма принимает вид

$$S(a;b)=\sum_{i=1}^{n}(ax_i+b-y_i)^2\ ,$$

а в случае квадратической зависимости – следующий вид:

$$S(a;b;c)=\sum_{i=1}^{n}(ax_i^2+bx_i+c-y_i)^2\ .$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \end{cases}$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i, \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i. \end{cases}$$

Оценка тесноты корреляционной зависимости

- Для оценки тесноты корреляционной зависимости служит корреляционное отношение:

$$\eta = \sqrt{\frac{\sigma^2(\bar{y}_x)}{\sigma^2(y)}}$$

где $\sigma^2(y)$ – выборочная дисперсия случайной величины Y, вычисленная по всей таблице;

$\sigma^2(\bar{y}_x)$ – дисперсия условных средних относительно общей средней, так называемая внешняя дисперсия.

Критерий Фишера

$$F_{эмп} = \frac{\sigma^2_{ост}}{\sigma^2_{воспр.\ ср}},$$

- где $\sigma^2_{ост} = \frac{1}{n-l}\sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2$ – остаточная дисперсия;

  l – число коэффициентов в уравнении регрессии;

  $\hat{y}_i$ – ордината линии регрессии в точке x_i;

  $\sigma^2_{воспр.\,ср}$ – дисперсия воспроизводимости средних, равная исправленной внутренней дисперсии, деленной на число m экспериментов , по которым вычислялись условные средние $\bar{y}_i$:

$$\sigma^2_{воспр.ср.} = \frac{1}{m} \cdot \frac{m}{m-1} \cdot \sigma^2_{внутр.} = \frac{1}{m-1} \cdot \sigma^2_{внутр.}$$

- Величина $F_{эмп}$ имеет распределение Фишера с $k_1 = n - l$ и $k_2 = n(m-1)$ числами степеней свободы (n – число задаваемых экспериментатором значений величины X, m – число проводимых опытов, l – число коэффициентов в уравнении регрессии).

Из таблицы критических точек распределения Фишера находим .

- Если $F_{эмп} < F_{крит}$, уравнение регрессии адекватно.

- Если $F_{эмп} > F_{крит}$ расхождение между теоретической и эмпирической линиями регрессии значимо, уравнение не адекватно, следует взять многочлен более высокого порядка.

Линейная корреляция

- Из всех корреляционных зависимостей надо особо выделить линейную корреляцию, т.е. такую, когда точки регрессии располагаются вблизи некоторой прямой линии.

Виды регрессии

- 1) регрессия Y на X в виде функциональной зависимости

$$\overline{y}_x = \rho_{yx} x + b \quad ;$$

- 2) регрессия X на Y в виде функциональной зависимости

$$\overline{x}_y = \rho_{xy} y + d \quad .$$

Выборочный коэффициент корреляции

$$r_в = \frac{\sum n_{xy} xy - n \cdot \bar{x} \cdot \bar{y}}{n \cdot \sigma_x \cdot \sigma_y}$$

Выборочное уравнение прямой линии регрессии Y на X

$$\bar{y}_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$-1 \le r_в \le 1$$

Выборочное уравнение прямой линии регрессии X на Y

$$\bar{x}_y - \bar{x} = r_в \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$-1 \le r_в \le 1$$

- Если данные наблюдений над признаками X и Y заданы в виде корреляционной таблицы с равноотстоящими вариантами, то целесообразно перейти к условным вариантам :

$$U_i = \frac{x_i - C_1}{h_1} \quad , \quad V_j = \frac{y_j - C_2}{h_2}$$

Выборочный коэффициент корреляции

$$r_в = \frac{\sum n_{uv} uv - n \cdot \bar{u} \cdot \bar{v}}{n \cdot \sigma_u \cdot \sigma_v}$$

$$\bar{u} = \frac{\sum n_u u}{n}, \bar{v} = \frac{\sum n_v v}{n},$$

$$\sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2}, \sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2}.$$

$$\bar{x} = \bar{u} \cdot h_1 + C_1,\ \bar{y} = \bar{v} \cdot h_2 + C_2$$

$$\sigma_x = \sigma_u \cdot h_1,\ \sigma_y = \sigma_v \cdot h_2.$$