Ответы на вопросы к экзамену по курсу
Прикладная теория вероятностей и математическая статистика

1. Понятие вероятности случайного события: аксиоматический подход. Аксиомы А. Н. Колмогорова.

В теории вероятностей существует 4 подхода к определению вероятности случайного события A: статистический, классический, геометрический и аксиоматический.

Результатом случайного эксперимента называется ***случайное событие***.

Аксиоматическое определение вероятности – единственное формальное определение понятия вероятности случайного события.

Пусть задано пространство элементарных исходов Ω некоторого случайного эксперимента. Тогда по определению, сформулированным академиком А. Н. Колмогоровым:

Каждому случайному событию А соответствует некоторое число Р(А), называемое ***вероятностью***, и такое, что выполняются следующие аксиомы:

Аксиома 1. $0 \le P(A) \le 1$.

Аксиома 2. $P(\Omega)=1$, *где* Ω – *достоверное случайное событие*.

Аксиома 3. Если $AB=\varnothing$, то $P(A+B)=P(A)+P(B)$.

Следствия из аксиом Колмогорова (Свойства вероятности):

Следствие 1. Если случайные события $A_1, A_2, \dots A_n$ образуют полную группу, то $P(A_1+A_2+\dots+A_n)=1$.

Следствие 2. $P(A)+P(\acute{A})=1$.

Следствие 3. Вероятность невозможного события равна 0: $P(\varnothing)=0$.

Следствие 4. Если $A \subseteq B$, то имеют место следующие соотношения:

1. $P(A) \le P(B)$;
2. $P(B¿) \le P(B)-(A)$.

2. Понятие случайной величины. Дискретные и непрерывные случайные величины.

Случайной величиной называют числовую функцию $\xi(\omega)$ или ξ, принимающую вещественные значения и заданную на множестве всех элементарных исходов Ω.

Случайные величины бывают дискретные и непрерывные.

Случайная величина называется ***дискретной***, если множество значений, которое она принимает конечное или счетное.

Например, кубик подбросили 7 раз. Число выпадения 5 – есть случайная дискретная величина, которая может принимать значения от 0 до 7.

Случайная величина называется ***непрерывной***, если множество значений, которое она принимает, сплошь заполняет некоторый промежуток на числовой оси.

Например, двое друзей Петя и Вася договорились встретиться в промежутке времени от 13 до 13.15. Петя пришел в 13.02, когда Васи еще не было. Время встречи с Васей – непрерывная случайная величина, которая может принимать значения от 13.02 до 13.15.

3. Функция распределения случайной величины и её свойства (без

доказательства).

Функцией распределения случайной величины $\xi(\omega)$ называют функцию действительного аргумента x, определенную равенством: $F_\xi(x)=P\{\xi<x\}, \forall\ x\in R$.

Свойства функции распределения:

1. Ограниченная функция: $0\le F_\xi(x)\le 1$.
2. Неубывающая функция: $\forall\ x_2>x_1 : F(x_2)>F(x_1)$.
3. Предельное поведение функции:

$\lim\limits_{x\to+\infty} F_\xi(x)=1$, $\lim\limits_{x\to-\infty} F_\xi(x)=0$.

4. Вероятность попадания с.в. $\xi(\omega)$ в полуоткрытый интервал ¿ $P\{a\le\xi<b\}=F_\xi(b)-F_\xi(a)$.
5. Если x_0 – точка разрыва функции распределения, то $P(\xi=x_0)=\lim\limits_{x\to+x_0} F_\xi(x)-\lim\limits_{x\to-x_0} F_\xi(x)$.
6. Функция распределения непрерывна слева, т.е. если x_0 – точка разрыва функции распределения, то за значение функции в этой точке принимают предел слева: $F_\xi(x_0)=\lim\limits_{x\to-x_0} F_\xi(x)$.

Функция распределения дискретной случной величины всегда кусочно-постоянна, а функция распределения непрерывной случной величины - обязательно непрерывная.

4. Основные числовые характеристики дискретной случайной величины (математическое ожидание, дисперсия, среднеквадратическое отклонение, центральные и начальные моменты порядка k): формулы для вычисления.

Математическое ожидание дискретной случайной величины – это число $E[\xi]$, которое вычисляют по формуле:

$$E[\xi]=\sum_{i=1}^{\infty} x_i p_i.$$

Из определения математического ожидания следует, что его значение не меньше наименьшего возможного значения случайной величины и не больше наибольшего.

Свойства математического ожидания

1. Математическое ожидание постоянной равно самой постоянной: $E(C)=C$.
2. Постоянный множитель можно выносит за знак математического ожидания: $E(C\xi)=CE(\xi)$.
3. Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий: $E(\xi\psi)=E(\xi)E(\psi)$.
4. Математическое ожидание суммы двух случайных величин (зависимых или независимых) равно сумме математических ожиданий слагаемых: $E(\xi+\psi)=E(\xi)+E(\psi)$.

Для оценки разброса значений случайной величины относительно математического ожидания используется дисперсия и среднеквадратическое отклонение.

Дисперсией случайной величины (как дискретно, так и непрерывной) называют математическое ожидание квадрата ее отклонения от ее математического ожидания – это число $V[\xi]$, которое вычисляют по формуле:

$V[\xi]=E\left[(\xi-E[\xi])^2\right]$.

Дисперсия дискретной случайной величины вычисляется по формулам:

$V[\xi]=\sum_{i=1}^{\infty}(x_i-E[\xi])^2 p_i; V[\xi]=\sum_{i=1}^{\infty} x_i^2 p_i-(E[\xi])^2$.

Свойства дисперсии:

1. Дисперсия постоянной величины равна нулю: $V[C]=0$.
2. Постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат: $V(C\xi)=C^2V(\xi)$.
3. Дисперсия суммы двух независимых случайных величин равна сумме их дисперсий: $V(\xi+\psi)=V(\xi)+V(\psi)$.

Среднеквадратическим отклонением σ называют арифметический корень из дисперсии (положительное число): $\sigma=\sqrt{V[\xi]}$.

Начальным моментом порядка k называют следующую числовую характеристику:

$a_k=E\left[\xi^k\right]$.

Центральным моментом порядка k называют следующую числовую характеристику:

$\beta_k=E\left[(\xi-E[\xi])^k\right]$.

Из определения центрального момента следует, что центральным момент второго порядка равен дисперсии.

5. Функция плотности распределения вероятностей непрерывной случайной величины и её свойства (без доказательства).

Плотностью распределения непрерывной случайной величины называется функция $f_\xi(x)$, которая удовлетворяет равенству:

$F_\xi(x)=\int_{-\infty}^{x} f_\xi(t)dt$.

Свойства функции плотности распределения:

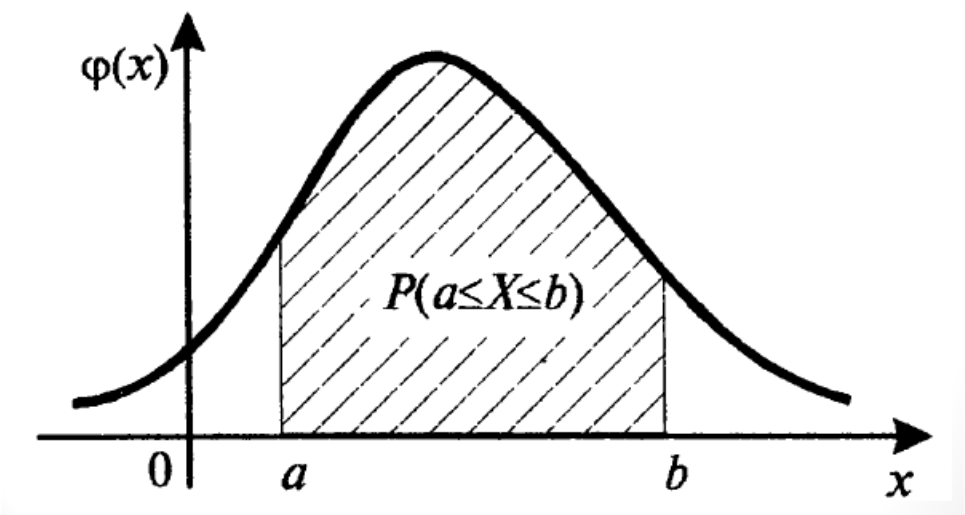
1. В точках, где функция распределения непрерывна $f_\xi(x)=F_\xi'(x)$.
2. Плотность распределения всегда неотрицательна: $f_\xi(x)\geq 0$.
3. Свойство нормировки:

$\int_{-\infty}^{+\infty} f_\xi(x)dx=1$.

4. Вероятность попадания непрерывной случайной величины в промежуток $[a;b]$ можно вычислить по формуле:

 vk.com/id446425943

$$P(\xi \in [a;b]) = \int_a^b f_\xi(x)dx.$$



Функция, для которой выполняются свойства 2 (неотрицательности) и 3 (нормировки), является плотностью какого-то распределения.

6. Основные числовые характеристики непрерывной случайной величины (математическое ожидание, дисперсия, среднеквадратическое отклонение, центральны и начальные моменты порядка k): формулы для вычисления.

Математическим ожиданием непрерывной случайной величины ξ называют число $E[\xi]$, которое вычисляют по формуле:

$$E[\xi] = \int_{-\infty}^{+\infty} x f_\xi(x)dx.$$

Дисперсию непрерывной случайной величины вычисляют по формулам

$$V[\xi] = \int_{-\infty}^{+\infty} (x - E[\xi])^2 f_\xi(x)dx; V[\xi] = \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - (E[\xi])^2.$$

Среднеквадратическим отклонением σ называют арифметический корень из дисперсии (положительное число): $\sigma = \sqrt{V[\xi]}$.

Начальным моментом порядка k называют следующую числовую характеристику:

$$a_k = \int_{-\infty}^{+\infty} x^k f_\xi(x)dx.$$

Центральным моментом порядка k называют следующую числовую характеристику:

$$\beta_k = \int_{-\infty}^{+\infty} (x - E[\xi])^k f_\xi(x)dx.$$

Ии для дискретной случайной величины, из определения центрального момента следует, что центральным момент второго порядка равен дисперсии.

7. Основные законы распределения дискретных случайных величин (биномиальный, Пуассона, геометрический).

Биноминальное распределение.

Введем ряд определений:

Успех – если в результате случайного эксперимента произойдет случайное событие А.

Неудача – если в результате случайного эксперимента произойдет случайное событие $\acute{A}$.

Обозначения: $P(A)=p, P(\acute{A})=q=1-p$.

Последовательность испытаний называется схемой Бернулли, если:

1. Все испытания проводят независимо друг от друга.

2. В каждом испытании фиксируются только два исхода – появление случайного события А и противоположного ему случайного события $\acute{A}$.

3. Вероятность наступления случайного события А (успеха) не меняется от испытания к испытанию.

Вероятность события, состоящего в том, что в n испытаниях «успех» наступил ровно k раз, вычисляется по формуле:

$$P_k(n)=C_n^k p^k q^{n-k}, \text{где } C_n^k=\frac{n!}{k!(n-k)!}.$$

Например, известно, что 30% держателей акций страховых компаний старше 50 лет, требуют возмещение страховых сумм. Имеем: $p=0,3$; $q=0,7$. Тогда вероятность того, что 4 из 10 акционеров старше 50 лет потребуют вернуть страховые суммы: $P_4(10)=C_{10}^4\cdot 0,3^4\cdot 0,7^6\approx 0,2$.

Биноминальное распределение характеризуется двумя числами: вероятностью появления события в одном испытании p и числом испытаний n: $\xi\in B(p,n)$. Математическое ожидание и дисперсия выражаются формулами: $E[\xi]=pn, V[\xi]=pnq$.

Распределение Пуассона

Если число испытаний, проводимых по схеме Бернулли, неограниченно возрастает $(n\to\infty)$, а произведение pn остается постоянным: $pn=\lambda$, то число появлений случайного события А является дискретной случайной величиной, распределенной по закону Пуассона: $\xi\in P(\lambda)$.

$$P_k(n)=\frac{\lambda^k e^{-\lambda}}{k!}.$$

Математическое ожидание и дисперсия выражаются формулами: $E[\xi]=V[\xi]=\lambda$.

Например: число выигрышных лотерейных билетов в партии из 10000 штук.

Рассмотрим последовательность независимых испытаний, которые проводятся до тех пор пока не появится случайное событие А (до первого успеха). Число испытаний, до появления случайного события А является дискретной случайной величиной, имеющей ***геометрическое распределение:*** $\xi\in G(p)$.

Вероятность того, что событие А наступил при $k-$ом испытании, вычисляется по формуле:

$P_k(n)=pq^{k-1}$.

Математическое ожидание и дисперсия выражаются формулами:

$E[\xi]=\frac{1}{p}, V[\xi]=\frac{q}{p^2}$.

Например, стрельба в тире ведется до первого попадания в мишень.

8. Основные законы распределения непрерывных случайных величин (равномерный, показательный, нормальный).

Непрерывная случайная величина ***равномерно распределена*** на отрезке $[a,b]$, если выражение для плотности распределения имеет вид:

$$f_\xi(x)=\begin{cases}0, если\, x<a\\ \frac{1}{b-a}, если\, a\le x\le b\\ 0, если\, x>b\end{cases}.$$

Функция распределения имеет вид:

$$F_\xi(x)=\begin{cases}0, если\, x<a\\ \frac{x-a}{b-a}, если\, a\le x\le b\\ 1, если\, x>b\end{cases}.$$

У этого распределения два параметра: границы отрезка a и b, $\xi\in R(a;b)$.

Математическое ожидание и дисперсия выражаются формулами:

$E[\xi]=\frac{a+b}{2}, V[\xi]=\frac{(b-a)^2}{12}$.

Непрерывная случайная величина распределена по ***показательному закону распределения*** с параметром λ ($\xi\in E(\lambda)$¿, если выражение для плотности распределения имеет вид:

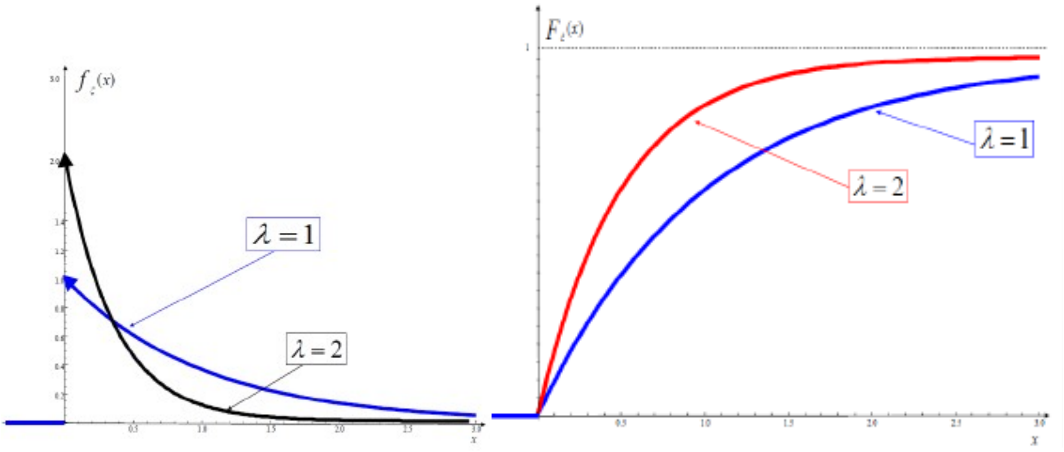
$$f_\xi(x)=\begin{cases}0, если\, x\le 0\\ \lambda e^{-\lambda x}, если\, x>0\end{cases}.$$

Функция распределения имеет вид:

$$F_\xi(x)=\begin{cases}0, если\, x\le 0\\ 1-\lambda e^{-\lambda x}, если\, x>0\end{cases}.$$

Математическое ожидание и дисперсия выражаются формулами:

$E[\xi]=\frac{1}{\lambda}, V[\xi]=\frac{1}{\lambda^2}$.



 vk.com/id446425943

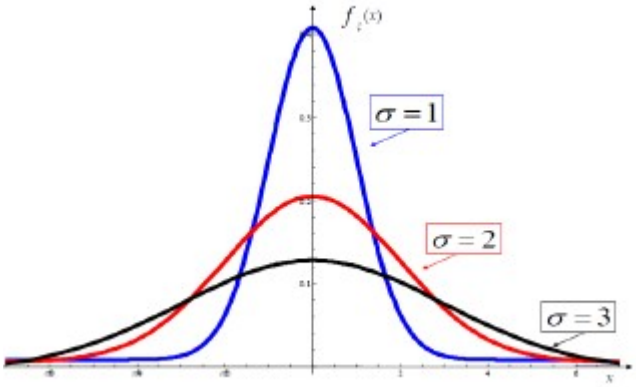
Непрерывная случайная величина нормально распределена с параметрами $m \in R$ и $\sigma > 0$ ($\xi \in N(m,\sigma)$¿, если функция плотности имеет вид:

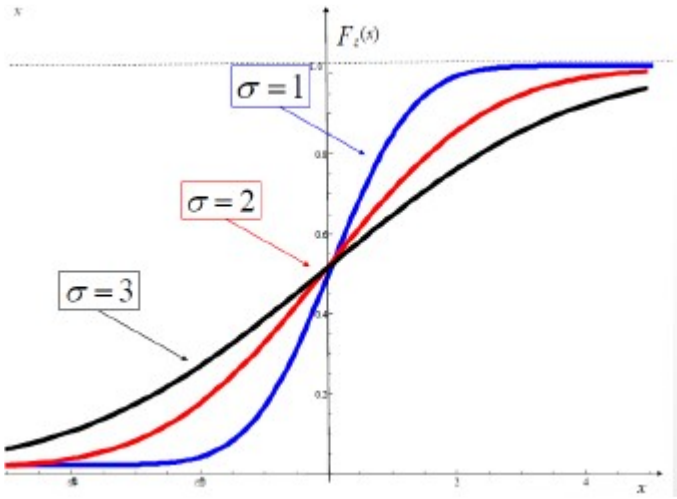
$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}.$$

Функция распределения случайной величины:

$$F_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-m)^2}{2\sigma^2}}\, dt.$$

Математическое ожидание и дисперсия выражаются формулами:
$E[\xi] = m, V[\xi] = \sigma^2$.

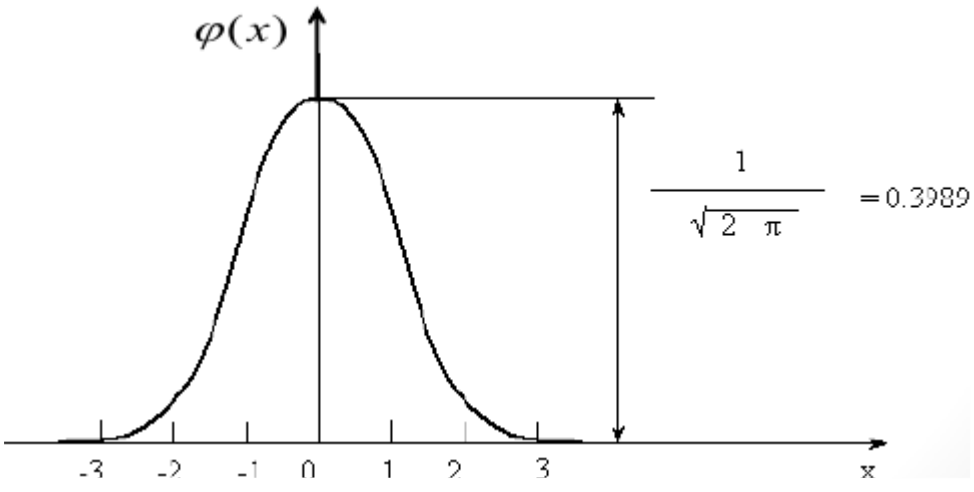




$E[\xi] = m = 0$.

Непрерывная случайная величина имеет стандартное нормальное распределение, если она распределена нормально с параметрами: $\xi \in N(0,1)$.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$



Функция распределения стандартного нормального распределения:

$$F_\xi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}}\, dt \text{ и связана с функцией Лапласа } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{\frac{-t^2}{2}}\, dt:$$

$$F_\xi(x) = \frac{1}{2} + \Phi(x).$$

Для произвольного нормального распределения:

$$F_\xi(x) = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right).$$

Функция Лапласа четная и при $x \to \infty$ равна 0,5.

9. Понятие случайного вектора.

Совокупность случайных величин $\xi_1, \xi_2, \ldots \xi_n$, заданных на одном и том же вероятностном пространстве Ω, называются $n-$¿**мерным случайным вектором** или $n-$¿мерной случайной величиной. Сами случайные величины называются при этом координатами случайного вектора. При $n=2$ случайная величина называется двумерной.

Двумерным случайным вектором η называют вектор-столбец $\eta=(\xi_1,\xi_2)^T$, составляющими которого являются случайные величины, заданные на одном и том же пространстве элементарных исходов Ω.

Двумерный случайный вектор $\eta=(\xi_1,\xi_2)^T$ будем называть ***дискретным***, если каждая его составляющая является дискретной случайной величиной.

Законом распределения двумерного случайного вектора $\eta=(\xi_1,\xi_2)^T$ будем называть таблицу вида:

X/Y	y_1	y_2	…	y_m
x_1	p_{11}	p_{12}		p_{1m}
x_2	p_{21}	p_{22}		p_{2m}
…				
x_n	p_{n1}	p_{n2}		p_{nm}

$p_{ij}=P\{\xi_i=x_i,\xi_j=y_j\}, \sum p_{ij}=1.$

10. Функция распределения двумерного случайного вектора и её свойства (без доказательства).

Функцией распределения двумерного случайного вектора η называют функцию, выражающую вероятность одновременного выполнения неравенств $\xi_1<x$ и $\xi_2<y$ и определенную равенством: $F_\eta(x;y)=P(\xi_1<x;\xi_2<y), \forall\ x,y\in R.$

Функция распределения $F_\eta(x;y)=P(\xi_1<x;\xi_2<y),\}$ в данном случае есть вероятность попасть в квадрант , верхним правым углом которого является точка $(x;y)$.

11. Маргинальные законы распределения двумерного случайного вектора.

Маргинальным (одномерным, частным) законом распределения случайной величины ξ_1 называют таблицу, в верхней строке которой перечислены все значения случайной величины ξ_1, а в нижней указаны вероятности, с которыми она их принимает:

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2		p_n

$$\text{где } p_i=\sum_{j=1}^{m} p_{ij}.$$

Аналогично записывается законом распределения случайной величины ξ_2:

y_j	y_1	y_2	…	y_m
q_j	q_1	q_2		q_m

$$\text{где } q_j=\sum_{i=1}^{n} p_{ij}.$$

12. Математическое ожидание двумерного случайного вектора.

Математическим ожиданием двумерного случайного вектора $\eta=(\xi_1,\xi_2)^T$ называют вектор вида:

$E[\eta]=(E[\xi_1],E[\xi_2])^T$, где

$$E[\xi_1]=\sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij}=\sum_{i=1}^{n} x_i p_i, E[\xi_2]=\sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij}=\sum_{j=1}^{m} y_j q_j.$$

13. Условные законы распределения двумерного случайного вектора.

Условным рядом распределения случайной величины ξ_1 при условии, что $\xi_2=y_j\ j=1,\dots m$ будем называть таблицу вида:

x_i	x_1	x_2	…	x_n
$p_{i/j}$	$p_{1/j}$	$p_{2/j}$		$p_{n/j}$

$$p_{i/j}=\frac{p_{ij}}{q_j}.$$

$$P(\xi_1=x_i/\xi_2=y_j)=\frac{P(\xi_1=x_i;\xi_2=y_j)}{P(\xi_2=y_j)}.$$

Условным рядом распределения случайной величины ξ_2 при условии, что $\xi_1=x_i, i=1,\dots n$ будем называть таблицу вида:

y_j	y_1	y_2	…	y_m
$q_{j/i}$	$q_{1/i}$	$q_{2/i}$		$q_{m/i}$

$$q_{j/i}=\frac{p_{ij}}{p_i}.$$

$$P(\xi_2=y_j/\xi_1=x_i)=\frac{P(\xi_1=x_i;\xi_2=y_j)}{P(\xi_1=x_i)}.$$

14. Условное математическое ожидание.

Условным математическим ожиданием случайной величины ξ_1 при условии ξ_2 называют число $E[\xi_1/\xi_2]$, значения которого для каждого конкретного значения $\xi_2=y_j$ определяются по формуле:

$$E[\xi_1/\xi_2=y_j]=\sum_{i=1}^{n} x_i p_{i/j}.$$

Условным математическим ожиданием случайной величины ξ_2 при

условии ξ_1 называют число $E[\xi_2/\xi_2]$, значения которого для каждого конкретного значения $\xi_1=x_i$ определяются по формуле:

$$E[\xi_2/\xi_1=x_i]=\sum_{j=1}^{m} y_j q_{j/i}.$$

15. Зависимость и независимость случайных величин: определение, критерии независимости.

Говорят, что две случайные величины ξ_1 и ξ_2 ***независимы***, если независимы связанные с ними случайные события A и B, где $A=\{\xi_1<x\}, B=\{\xi_2<y\}$.

Если события A и B зависимы, то ***зависимы*** и случайные величины ξ_1 и ξ_2.

Если случайные величины ξ_1 и ξ_2, составляющие случайный вектор, независимы, то выполняется равенство: $P(\xi_1<x;\xi_2<y)=P(\xi_1<x)P(\xi_2<y)$.

Говорят, что две случайные величины ξ_1 и ξ_2 независимы, если выполняется равенство:

$$F_\eta(x;y)=F_{\xi_1}(x)F_{\xi_2}(y).$$

Необходимое и достаточное условие того, что две случайные величины ξ_1 и ξ_2 независимы – выполнение равенства:
$p_{ij}=p_i q_j, \forall\ i=1,\dots n; j=1,\dots m.$

Необходимое и достаточное условие того, что две случайные величины ξ_1 и ξ_2 независимы – выполнение равенств:

$$p_{i/j}=p_i, \forall\ i=1,\dots n; q_{j/i}=q_j, \forall\ j=1,\dots m.$$

Если хотя бы для одной пары индексов равенство нарушается, то величины будут зависимыми.

Теорема. Если две случайные величины ξ_1 и ξ_2 независимы, то и любые функции от этих случайных величин независимы.

Теорема. Если две случайные величины ξ_1 и ξ_2 независимы, то имеют место равенства: $E[\xi_1\xi_2]=E[\xi_1]E[\xi_2]$ и $V[\xi_1+\xi_2]=V[\xi_1]+V[\xi_2]$.

16. Момент корреляции (ковариация) и его свойства (с доказательством)

Моментом корреляции (ковариацией) двух случайных величин ξ_1 и ξ_2 $cov(\xi_1,\xi_2)$ называется математическое ожидание произведения отклонений этих величин:

$$V_{\xi_1\xi_2}=E\left[(\xi_1-E[\xi_1])(\xi_2-E[\xi_2])\right].$$

При этом: если η - дискретная двумерная случайная величина, то ковариация вычисляется по формуле:

$$V_{\xi_1\xi_2}=\sum_{i=1}^{n}\sum_{j=1}^{m}(\xi_{1i}-E[\xi_1])(\xi_{2i}-E[\xi_2])p_{ij};$$

если непрерывная:

$$V_{\xi_1\xi_2}=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E[\xi_1])^2(y-E[\xi_2])^2 f_\xi(x,y)\,dxdy.$$

Если момент корреляции двух случайных величин равен нулю, то такие случайные величины называются некоррелированными.

Свойства момента корреляции:

1. Ковариация симметрична, т.е. $V_{\xi_1\xi_2}=V_{\xi_2\xi_1}$

Доказательство

$V_{\xi_1\xi_2}=E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=E\left[\left(\xi_2-E\left[\xi_2\right]\right)\left(\xi_1-E\left[\xi_1\right]\right)\right]=V_{\xi_2\xi_1}.$

2. Момент корреляции независимых случайных величин равен нулю.

Доказательство

$V_{\xi_1\xi_2}=E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=E\left[\xi_1\xi_2-E\left[\xi_1\right]\xi_2-\xi_1E\left[\xi_2\right]+E\left[\xi_1\right]E\left[\xi_2\right]\right]=¿$

$¿\,E\left[\xi_1\xi_2\right]-E\left[E\left[\xi_1\right]\xi_2\right]-E\left[\xi_1E\left[\xi_2\right]\right]+E\left[E\left[\xi_1\right]E\left[\xi_2\right]\right]=¿$

$¿\left[\text{для независимых случайных величин } E\left[\xi_1\xi_2\right]=E\left[\xi_1\right]E\left[\xi_2\right]\right]=¿$

$¿\,E\left[\xi_1\right]E\left[\xi_2\right]-E\left[\xi_1\right]E\left[\xi_2\right]-E\left[\xi_1\right]E\left[\xi_2\right]+E\left[\xi_1\right]E\left[\xi_2\right]=0.$

Что и требовалось доказать.

3. Постоянный множитель можно вынести за знак ковариаций, т.е. $V_{C\xi_1\xi_2}=C\,V_{\xi_2\xi_1}.$

Доказательство

$V_{C\xi_1\xi_2}=E\left[\left(C\,\xi_1-E\left[C\,\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=E\left[\left(C\,\xi_1-CE\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=¿$

$¿\,E\left[\left(C\left(\xi_1-E\left[\xi_1\right]\right)\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=C\,E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=C\,V_{\xi_2\xi_1}.$

4. Ковариация не изменится, если к одной из случайных величин (или к обоим сразу) прибавить постоянную, т.е. $V_{(\xi_1+C)\xi_2}=V_{\xi_1\xi_2}.$

Доказательство

$V_{(\xi_1+C)\xi_2}=E\left[\left(\xi_1+C-E\left[\xi_1+C\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=¿$

$¿\,E\left[\left(\xi_1+C-E\left[\xi_1\right]-C\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=V_{\xi_1\xi_2}.$

5. Дисперсия случайной величины, есть ковариация ее с самой собой, т.е.: $V_{\xi_1\xi_1}=V\left[\xi_1\right];V_{\xi_2\xi_2}=V\left[\xi_2\right].$

Доказательство

$V_{\xi_1\xi_1}=E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_1-E\left[\xi_1\right]\right)\right]=E\left[\left(\xi_1-E\left[\xi_1\right]\right)^2\right]=V\left[\xi_1\right].$

Аналогично доказывается $V_{\xi_2\xi_2}=V\left[\xi_2\right].$

Что и требовалось доказать.

6. Дисперсия суммы (разности) двух случайных величин равна сумме их дисперсий плюс (минус) удвоенная ковариация этих случайных величин, т. е.:

$V\left[\xi_1\pm\xi_2\right]=V\left[\xi_1\right]+V\left[\xi_2\right]\pm2\,V_{\xi_1\xi_1}.$

Доказательство

$V\left[\xi_1\pm\xi_2\right]=E\left[\left(\xi_1\pm\xi_2-E\left[\xi_1\pm\xi_2\right]\right)^2\right]=E\left[\left(\xi_1-E\left[\xi_1\right]\pm\left(\xi_2-E\left[\xi_2\right]\right)\right)^2\right]=¿$

$¿\,E\left[\left(\xi_1-E\left[\xi_1\right]\right)^2\pm2\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)+\left(\xi_2-E\left[\xi_2\right]\right)^2\right]=¿$

$¿\,E\left(\xi_1-E\left[\xi_1\right]\right)^2\pm2\,E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]+E\left(\xi_2-E\left[\xi_2\right]\right)^2=¿$

$¿\,V\left[\xi_1\right]\pm2V_{\xi_1\xi_1}+V\left[\xi_2\right].$

Что и требовалось доказать.

7. Момент корреляции двух случайных величин по абсолютной величине не превосходит произведения среднеквадратических отклонений, т.е.

$\left|V_{\xi_1\xi_2}\right|\le\sigma_{\xi_1}\cdot\sigma_{\xi_2}.$

Доказательство

Заметим, что для независимых случайных величин неравенство

выполняется (см. предыдущую теорему.). Итак, пусть случайные величины ξ_1 и ξ_2 зависимые. Рассмотрим нормированные случайные величины:

$$\xi_1^{¿}=\frac{\xi_1-E[\xi_1]}{\sigma_{\xi_1}} u \xi_2^{¿}=\frac{\xi_2-E[\xi_2]}{\sigma_{\xi_2}} u$$

вычислим дисперсию случайной величины $\xi_1^{¿}-\xi_2^{¿}$ с учётом свойства 6:

$$V[\xi_1^{¿}-\xi_2^{¿}]=V[\xi_1^{¿}]-2E\left[\left(\xi_1^{¿}-E[\xi_1^{¿}]\right)\left(\xi_2^{¿}-E[\xi_2^{¿}]\right)\right]+V[\xi_2^{¿}]$$

$$V[\xi_1^{¿}]=V[\xi_2^{¿}]=1$$

Доказательства:

$$E[\xi_1^{¿}]=E[\xi_2^{¿}]=0$$

$$E\left[\frac{\xi_1-E[\xi_1]}{\sigma_{\xi_1}}\right]=\frac{E[\xi_1]-E[\xi_1]}{\sigma_{\xi_1}}=0; V\left[\frac{\xi_1-E[\xi_1]}{\sigma_{\xi_1}}\right]=\frac{V[\xi_1-E[\xi_1]]}{\sigma_{\xi_1}^{\ 2}}=\frac{V[\xi_1]}{\sigma_{\xi_1}^{\ 2}}=1.$$

Тогда:

$$V[\xi_1^{¿}\pm\xi_2^{¿}]=1\pm 2E\left[(\xi_1^{¿})(\xi_2^{¿})\right]+1=2\left(1\pm E\left[(\xi_1^{¿})(\xi_2^{¿})\right]\right)=¿$$

$$¿ 2\left(1-E\left[\left(\frac{\xi_1-E[\xi_1]}{\sigma_{\xi_1}}\right)\left(\frac{\xi_2-E[\xi_2]}{\sigma_{\xi_2}}\right)\right]\right)=¿$$

$$¿ 2\left(1-\frac{1}{\sigma_{\xi_1}\sigma_{\xi_2}}E\left[(\xi_1-E[\xi_1])(\xi_2-E[\xi_2])\right]\right)=2\left(1-\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\sigma_{\xi_2}}\right)\geq 0,$$

так как по определению дисперсия неотрицательная величина.

$$1-\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\sigma_{\xi_2}}\geq 0\Rightarrow\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\sigma_{\xi_2}}\leq 1\Rightarrow\left|V_{\xi_1\xi_2}\right|\leq\sigma_{\xi_1}\cdot\sigma_{\xi_2}.$$

Что и требовалось доказать.

8. Ковариация двух случайных величин равна математическому ожиданию их произведения минус произведение математических ожиданий:

$$V_{\xi_1\xi_2}=E[\xi_1\xi_2]-E[\xi_1]E[\xi_2].$$

Доказательство

Из определения ковариации двух случайных величин и свойств математического ожидания:

$$V_{\xi_1\xi_2}=E\left[(\xi_1-E[\xi_1])(\xi_2-E[\xi_2])\right]=¿$$

$$¿ E[\xi_1\xi_2-\xi_1E[\xi_2]-E[\xi_1]\xi_2+E[\xi_1]E[\xi_2]]=E[\xi_1\xi_2]-E[\xi_1E[\xi_2]]-¿$$

$$-E[[\xi_1]\xi_2]+E[E[\xi_1]E[\xi_2]]=E[\xi_1\xi_2]-E[\xi_1]E[\xi_2]-E[\xi_1]E[\xi_2]\pm E[\xi_1]E[\xi_2]=¿$$

$$¿ E[\xi_1\xi_2]-E[\xi_1]E[\xi_2].$$

Что и требовалось доказать.

17. Теорема о математическом ожидании произведения и дисперсии суммы двух независимых случайных величин (с доказательством).

Теорема. Математическое ожидание произведения двух случайных величин равно сумме произведений математических ожиданий и ковариации этих величин:

$$E[\xi_1\xi_2]=E[\xi_1]E[\xi_2]+V_{\xi_1\xi_2}.$$

Доказательство

Из определения ковариации двух случайных величин и свойств математического ожидания:

$$V_{\xi_1\xi_2}=E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]=¿$$
$$¿E\left[\xi_1\xi_2-\xi_1E\left[\xi_2\right]-E\left[\xi_1\right]\xi_2+E\left[\xi_1\right]E\left[\xi_2\right]\right]=E\left[\xi_1\xi_2\right]-E\left[\xi_1E\left[\xi_2\right]\right]-¿$$
$$-E\left[\left[\xi_1\right]\xi_2\right]+E\left[E\left[\xi_1\right]E\left[\xi_2\right]\right]=E\left[\xi_1\xi_2\right]-E\left[\xi_1\right]E\left[\xi_2\right]-E\left[\xi_1\right]E\left[\xi_2\right]\pm E\left[\xi_1\right]E\left[\xi_2\right]=¿$$
$$¿E\left[\xi_1\xi_2\right]-E\left[\xi_1\right]E\left[\xi_2\right].$$

Что и требовалось доказать.

Следствие. Если случайные величины независимы, их ковариация равна нулю и: $E\left[\xi_1\xi_2\right]=E\left[\xi_1\right]E\left[\xi_2\right]$.

Теорема. Дисперсия суммы (разности) двух случайных величин равна сумме их дисперсий плюс (минус) удвоенная ковариация этих случайных величин, т. е.:

$$V\left[\xi_1\pm\xi_2\right]=V\left[\xi_1\right]+V\left[\xi_2\right]\pm2V_{\xi_1\xi_1}.$$

Доказательство

$$V\left[\xi_1\pm\xi_2\right]=E\left[\left(\xi_1\pm\xi_2-E\left[\xi_1\pm\xi_2\right]\right)^2\right]=E\left[\left(\xi_1-E\left[\xi_1\right]\pm\left(\xi_2-E\left[\xi_2\right]\right)\right)^2\right]=¿$$
$$¿E\left[\left(\xi_1-E\left[\xi_1\right]\right)^2\pm2\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)+\left(\xi_2-E\left[\xi_2\right]\right)^2\right]=¿$$
$$¿E\left(\xi_1-E\left[\xi_1\right]\right)^2\pm2E\left[\left(\xi_1-E\left[\xi_1\right]\right)\left(\xi_2-E\left[\xi_2\right]\right)\right]+E\left(\xi_2-E\left[\xi_2\right]\right)^2=¿$$
$$¿V\left[\xi_1\right]\pm2V_{\xi_1\xi_1}+V\left[\xi_2\right].$$

Что и требовалось доказать.

18. Коэффициент корреляции и его свойства (с доказательством).

Коэффициентом корреляции двух случайных величин называется отношение их момента корреляции к произведению средних квадратических отклонений:

$$\rho_{\xi_1\xi_2}=\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}}.$$

Коэффициент корреляции – безразмерная величина. Коэффициент корреляции – симметричная величина: $\rho_{\xi_1\xi_2}=\rho_{\xi_2\xi_1}$.

Свойства коэффициента корреляции:

1. $-1\le\rho_{\xi_1\xi_2}\le1$.

Доказательство

Согласно свойству 7 ковариации

$$\left|V_{\xi_1\xi_2}\right|\le\sigma_{\xi_1}\cdot\sigma_{\xi_2}\Rightarrow\frac{\left|V_{\xi_1\xi_2}\right|}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}}\le1\Rightarrow-1\le\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}}\le1.$$

По определению:

$$\rho_{\xi_1\xi_2}=\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}},\ следовательно,\ -1\le\rho_{\xi_1\xi_2}\le1,\ что\ и\ требовалось\ доказать.$$

2. Коэффициентом корреляции двух независимых случайных величин равен нулю: $\rho_{\xi_1\xi_2}=0$.

Доказательство

Согласно свойству 2 ковариации: момент корреляции независимых случайных величин равен нулю. По определению:

$\rho_{\xi_1\xi_2}=\frac{V_{\xi_1\xi_2}}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}}=\frac{0}{\sigma_{\xi_1}\cdot\sigma_{\xi_2}}=0$, *что и требовалось доказать*.

Обратное утверждение, вообще говоря, неверно: из некоррелированности двух случайных величин еще не следует их независимость.

Если коэффициентом корреляции двух случайных величин равен по модулю 1, то между этими случайными величинами существует линейная функциональная зависимость, при чем, если $\rho_{\xi_1\xi_2}=1-¿$ связь прямая, если $\rho_{\xi_1\xi_2}=-1-¿$ связь обратная.

19. Ковариационная и корреляционная матрицы: формулы для вычисления.

Ковариационной матрицей называют симметричную и неотрицательно определенную матрицу вида:

$$V_\eta=\begin{bmatrix} V[\xi_1] & V_{\xi_1\xi_2} \\ V_{\xi_2\xi_1} & V[\xi_2] \end{bmatrix}.$$

Корреляционной матрицей называют симметричную и неотрицательно определенную матрицу вида:

$$R_\eta=\begin{bmatrix} 1 & \rho_{\xi_1\xi_2} \\ \rho_{\xi_2\xi_1} & 1 \end{bmatrix}.$$

20. Понятие генеральной совокупности и выборки.

Генеральной совокупностью называют всю совокупность объектов (наблюдений), подлежащую изучению.

Выборкой или выборкой значений случайной величины (выборочной совокупностью) называется конечный набор значений, извлеченных из генеральной совокупности, с учетом того, что

1. Выбор производят случайным образом.
2. Выбор производят независимо друг от друга.
3. Выбор производят из одной и той же генеральной совокупности.

n – объём выборки.

Вариационным рядом называют выборку, все элементы которой расположены в порядке возрастания.

Выборки бывают: повторные и бесповторные,

Если случайно отбираемые объекты не возвращаются в генеральную совокупность, то это бесповторная выборка. Если же выбранный объект возвращается обратно (перед выбором следующего), то это повторная выборка.

Выборки: случайная (вероятностный отбор) и неслучайные (целенаправленный или неслучайный отбор).

21. Понятие статистического ряда. Сгруппированный статистический ряд, интервальный статистический ряд.

Статистический ряд – это последовательность различных элементов вариационного ряда x_i с указанием числа повторений (частот). Обычно эти значения дополняются относительными частотами.

Сгруппированный статистический ряд – это представление выборки случайной величины X, если есть основания предполагать, что изучаемая случайная величина является дискретной, т.е. может принимать значения $x_i, i=1,\dots n$ .x_i еще называют вариантами выборки. Например:

x_i	3	4	5	6	7	8	9	$\sum$
n_i	3	5	2	1	3	2	4	20
ω_i	0,15	0,25	0,1	0,05	0,15	0,1	0,2	1

$n_i - частоты$¿

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

Интервальный статистический ряд – это представление выборки случайной величины, если есть основания предполагать, что изучаемая случайная величина является непрерывной.

$R = x_{max} - x_{min} = 9 - 3 = 6$ – размах выборки.

Составим вариационный ряд, разделив интервал значений X на 4 равных части. Количество интервалов можно определить по формуле Стерджеса:

$r = [1 + 3{,}32 \log n]$.

Ширина интервала: $\delta = \frac{x_{max} - x_{min}}{4} = 1{,}5$.

Интервалы	Частоты, n_i	Накопленные частоты
$(3;4{,}5)$	8	8
$(4{,}5;6)$	3	11
$(6;7{,}5)$	3	14
$(7{,}5;9)$	6	20

22. Эмпирические функции распределения и плотности распределения: аналитические выражения и графики.

Эмпирическая функция распределения – это оценка теоретической функции распределения исследуемой случайной величины; строится на основе полученных статистических рядов распределения.

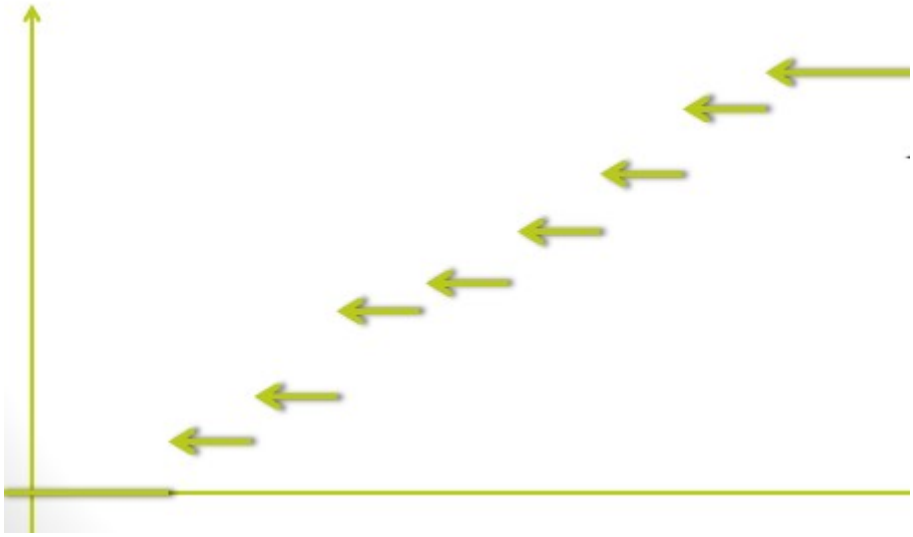
При дискретном распределении выборки функция $F^{¿}_n(x)$ показывает суммарную относительную частоту элементов выборки $p^{¿}_i$, меньших указанного числа x. Функция принимает значения от 0 (при x меньше минимального элемента выборки), до 1 (при x больше максимального элемента выборки). При дискретном распределении выборки функция имеет вид:

$$F^{¿}_n(x) = \sum_{i=1}^{k-1} p^{¿}_i = \begin{cases} 0, x \le x_1 \\ p^{¿}_i, x_1 < x \le x_2 \\ \dots \\ \sum_{i=1}^{k-1} p^{¿}_i, x_{k-1} < x \le x_k \\ 1, x > x_k \end{cases}.$$

График функции $F^{¿}_n(x)$ является ступенчатой линией с высотой очередной ступени равной относительной частоте пройденной варианты x_i (слева

 vk.com/id446425943

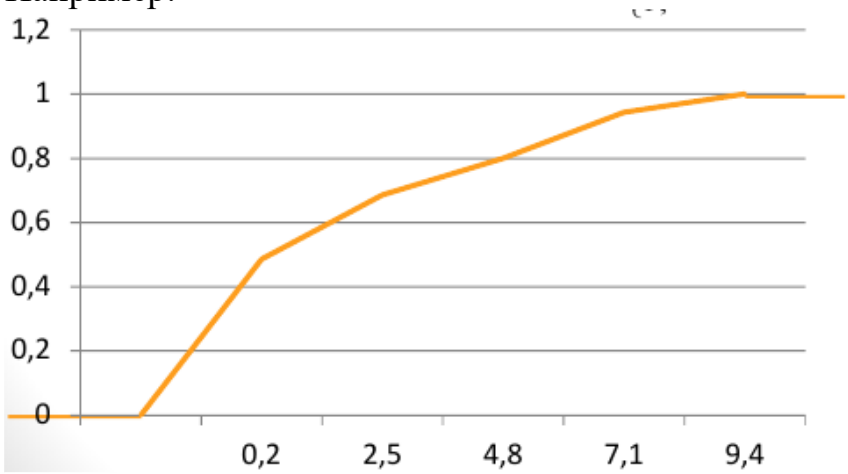
направо). График эмпирической функции распределения – называется кумулятой или кумулятивной кривой. Например:



Для непрерывной случайной величины:

$$F^{\prime}_{n}(x)=\frac{\tilde{p}^{\prime}_{r}\left(x-\tilde{x}_{r}\right)}{\delta}+\sum_{k=1}^{r-1}\tilde{p}^{\prime}_{k}=\begin{cases}0, x\le\tilde{x}_{1}\\ \dfrac{\tilde{p}^{\prime}_{1}\left(x-\tilde{x}_{1}\right)}{\delta}, x\in J_{1}\\ \dfrac{\tilde{p}^{\prime}_{2}\left(x-\tilde{x}_{2}\right)}{\delta}+\tilde{p}^{\prime}_{1}, x\in J_{2}\\ \dots\\ \dfrac{\tilde{p}^{\prime}_{r}\left(x-\tilde{x}_{r}\right)}{\delta}+\sum_{k=1}^{r-1}\tilde{p}^{\prime}_{k}, x\in J_{r}\\ 1, x>\tilde{x}_{r+1}\end{cases}.$$

График кумуляты представляет собой непрерывную ломанную. Например:

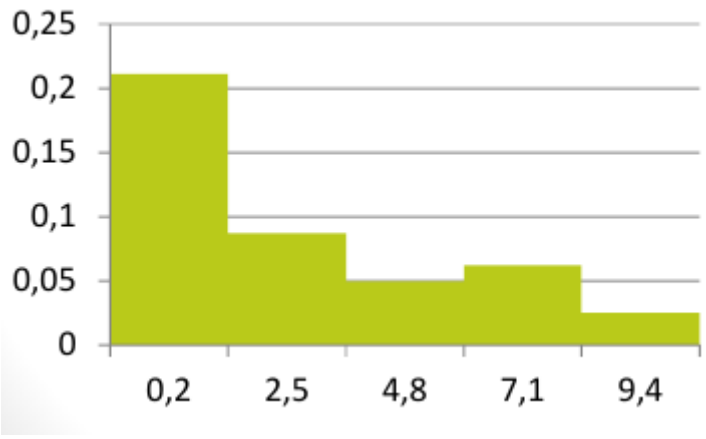


Эмпирическая плотность распределения – это оценка теоретической функции плотности распределения исследуемой случайной величины; строится на основе полученных статистических рядов распределения $f^{\prime}_{n}(x)$:

$$f_n^{\prime}(x)=\frac{\tilde{p}_r^{\prime}\left(x-\tilde{x}_r\right)}{\delta}+\sum_{k=1}^{r-1}\tilde{p}_k^{\prime}=\begin{cases}0, x\le\tilde{x}_1\\ \frac{\tilde{p}_1^{\prime}}{\delta}, x\in J_1\\ \ldots.\\ \frac{\tilde{p}_r^{\prime}}{\delta}, x\in J_r\\ 1, x>\tilde{x}_{r+1}\end{cases}.$$

График эмпирической плотность распределения называется ***гистограммой***.

Гистограмма строится только для выборки значений непрерывной случайной величины. В зависимости от количества интервалов гистограмма может принимать разный вид. Например:



23. Точечные оценки основных числовых характеристик для дискретных и непрерывных случайных величин.

Оценкой параметра θ называют всякую функцию результатов наблюдений за случайной величиной, с помощью которой судят о значении параметра:

$\theta_n=f\left(x_1, x_2, \ldots x_n\right)$.

Оценка параметра является случайной величиной, зависящей от закона распределения случайной величины и объем выборки.

Различают точечные и интервальные оценки параметров.

Оценку математического ожидания называют выборочным средним, обозначают $\acute{x}$ и в случае не сгруппированной выборки вычисляют по формуле:

$$\acute{x}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

Оценку дисперсии называют выборочной дисперсией, обозначают S^2 и в случае не сгруппированной выборки вычисляют по формуле:

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2.$$

Оценка дисперсии, найденная по формуле:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2$$

является смещённой оценкой со смещением $\frac{n-1}{n}$.

Для дискретной случайной величины (сгруппированный ряд):

$$\acute{x}=\frac{1}{n}\sum_{i=1}^{q} x_i^{\prime} k_i; S^2=\frac{1}{n-1}\sum_{i=1}^{q}\left(x_i^{\prime}-\acute{x}\right)^2 k_i.$$

Для непрерывной случайной величины (интервальный ряд):

$$\acute{x}=\frac{1}{n}\sum_{i=1}^{r} a_i l_i; S^2=\frac{1}{n-1}\sum_{i=1}^{r}\left(a_i-\acute{x}\right)^2 l_i, \text{где } a_i=\frac{\tilde{x}_{i+1}+\tilde{x}_i}{2}.$$

Величина $S=\sqrt{S^2}$ называется выборочным среднеквадратическим отклонением и служит ***статистической оценкой (несмещенной) среднеквадратического отклонения*** $\sigma=\sqrt{V[\xi]}$.

Оценка неизвестной вероятности $p=P(A)$ – относительная частота появления события A в n независимых испытаниях:

$$p^{\prime}=\frac{n_A}{n}.$$

24. Свойства точечных оценок (несмещенность, состоятельность, эффективность). Доказательство несмещенности выборочного среднего.

Оценка $\acute{x}$ неизвестного математического ожидания $E[\xi]$ является:

- несмещенной;
- состоятельной;
- асимптотически эффективной.

Оценка S^2 неизвестной дисперсии $V[\xi]=\sigma^2$ является:

- несмещенной;
- состоятельной;
- не является эффективной.

Оценка неизвестной вероятности $p=P(A)$ – относительная частота появления события А в n независимых испытаниях $p^{\prime}=\frac{n_A}{n}$ является:

- несмещенной;
- состоятельной;
- эффективной.

Оценка неизвестной функции распределения $F_{\delta}(x)$ - эмпирическая функция распределения выборки $F^{\prime}_n(x)$ является:

- несмещенной;
- состоятельной;
- не является эффективной.

Оценка θ_n параметра θ называется несмещенной, если ее математическое ожидание равно оцениваемому параметру:

$$E[\theta_n]=\theta.$$

Докажем несмещенности выборочного среднего: $E[\acute{x}]=E[\xi]$

$$E[\acute{x}]=E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]=\frac{1}{n}\sum_{i=1}^{n} E[x_i]=\frac{1}{n}\sum_{i=1}^{n} E[x_i]=\frac{1}{n}\cdot n\cdot E[\xi]=E[\xi].$$

Что и требовалось доказать.

25. Теоремы о свойствах известных оценок $\left(\acute{x}, S^2, p^{\prime}, F^{\prime}_n(x)\right)$.

Теорема. Пусть $x_1, x_2, \ldots x_n$ – выборка из генеральной совокупности X и $E(x_i)=E(X)=\mu$, $V(x_i)=V(X)=\sigma$.

Тогда выборочная средняя арифметическая $\acute{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$ является несмещенной и состоятельной оценкой математического ожидания $E(X)$.

Теорема. Пусть $x_1, x_2, \ldots x_n$ – выборка из генеральной совокупности X и $E(x_i)=E(X)=\mu$, $V(x_i)=V(X)=\sigma$.

Тогда величина $S_H{}^2=\frac{n}{n-1}S^2$,где $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\acute{x})^2$ – выборочная дисперсия, является несмещенной и состоятельной оценкой дисперсии $V(X)$.

Теорема. Пусть $x_1, x_2, \ldots x_n$ – выборка из генеральной совокупности X и $E(x_i)=E(X)=\mu$, $V(x_i)=V(X)=\sigma$.

Тогда величина $S=\sqrt{S^2}$ является несмещенной оценкой среднеквадратического отклонения $\sigma=\sqrt{V[\xi]}$.

Теорема. Пусть $x_1, x_2, \ldots x_n$ – выборка из генеральной совокупности X и оценка неизвестной вероятности $p=P(A)$ – относительная частота появления события А в n независимых испытаниях $p^{\prime}=\frac{n_A}{n}$ является:

- несмещенной;
- состоятельной;
- эффективной.

Теорема. Пусть $x_1, x_2, \ldots x_n$ – выборка из генеральной совокупности X и оценка неизвестной функции распределения $F_\delta(x)$ - эмпирическая функция распределения выборки $F^{\prime}_n(x)$ является:

- несмещенной;
- состоятельной;
- не является эффективной.

26. Понятие доверительного интервала. Доверительная вероятность.

Недостаток точечных оценок состоит в том, что при небольшом объеме выборки (как оно часто бывает), мы можем получать выборочные значения, которые далеки от истины. И в этих случаях логично потребовать, чтобы выборочная характеристика θ (средняя, дисперсия или какая-то другая) отличалась от генерального значения $\theta_г$ не более чем на некоторое положительное значение ε.

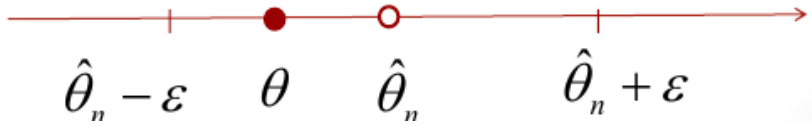
Пусть α некоторая достаточно большая вероятность, а $\theta-$¿ неизвестный параметр. Интервал $(\theta_1;\theta_2)=(\theta-\varepsilon;\theta+\varepsilon)$ со случайными границами, для которого выполняется следующее равенство:
$P\{\theta_1\le\theta\le\theta_2\}=\alpha$ называется ***доверительным интервалом*** и представляет собой интервальную оценку генерального значения по найденному выборочному значению, а вероятность $\alpha-$¿ ***доверительной вероятностью*** или надежностью интервальной оценки.

В качестве α выбирают обычно 0,90; 0,95; 0,997.

Часто ищут доверительный интервал симметрично относительно

несмещённой точечной оценки параметра $\tilde{\theta}_n$, т.е. ищут интервал вида: $(\tilde{\theta}_n-\varepsilon;\tilde{\theta}_n+\varepsilon)$ для которого выполняется:

$P\{\theta\in(\tilde{\theta}_n-\varepsilon;\tilde{\theta}_n+\varepsilon)\}=P\{|\theta-\tilde{\theta}_n|<\varepsilon\}=\alpha$.



Алгоритм построения доверительного интервала

1. Найти оценку $\tilde{\theta}_n$неизвестного параметра θ.
2. Ввести вспомогательную случайную величину Z (статистику), которая связана и с оцениваемым параметром и вычисленной оценкой: $Z=f(\tilde{\theta}_n,\theta)$. Закон распределения такой статистики должен быть известен.
3. Выбирают доверительную вероятность α и по специальным статистическим таблицам находят такие два числа v_1 и v_2, для которых выполняется неравенство:

$P\{v_1<Z=f(\tilde{\theta}_n,\theta)<v_2\}$.

5. Переходим к неравенству вида:

$P\{\varphi(v_1;\tilde{\theta}_n,)<\theta<\psi(v_2;\tilde{\theta}_n,)\}=\alpha$.

27. Доверительный интервал для неизвестного математического ожидания нормально распределённой случайной величины в случае известной дисперсии (с выводом).

Пусть случайная величина ξ имеет нормальное распределение: $N(m;\sigma)$.

Известно значение σ и задана доверительная вероятность (надежность) α. Требуется построить доверительный интервал для параметра m по выборочному среднему $\acute{x}$:

$\acute{x}=\frac{1}{n}\sum_{i=1}^{n}x_i-$*точечная оценка* m.

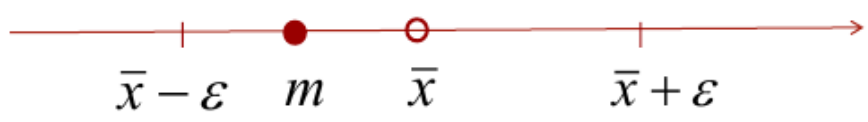
Примем без доказательства, что если случайная величина ξ распределена нормально, то и выборочное среднее $\acute{x}$, найденное по независимым наблюдениям, также распределено нормально - $\acute{x}\in N(m_1;\sigma_1)$.

$E[\acute{x}]=E[\xi]=m$, так как выборочное среднее является несмещенной оценкой математического ожидания. Следовательно, $m_1=m$.

$V[\acute{x}]=V\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n^2}\cdot n\cdot V[\xi]=\frac{\sigma^2}{n}$.

Получим: $\acute{x}\in N\left(m;\frac{\sigma}{\sqrt{n}}\right)$.

Выберем доверительную вероятность α:



$P\{|\acute{x}-m|<\varepsilon\}=\alpha$.

По свойству нормально распределённых случайных величин:

$$P\{|\xi-m|<\varepsilon\}=P\{m-\varepsilon<\xi<m+\varepsilon\}=\Phi\left(\frac{m+\varepsilon-m}{\sigma}\right)-\Phi\left(\frac{m-\varepsilon-m}{\sigma}\right)=¿$$

$$¿\,2\,\Phi\left(\frac{\varepsilon}{\sigma}\right)\Rightarrow P\{|\acute{x}-m|<\varepsilon\}=2\,\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\Rightarrow 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\alpha\Rightarrow$$

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\frac{\alpha}{2}=\Phi(C_\alpha),\text{где}\,C_\alpha=\frac{\varepsilon\sqrt{n}}{\sigma}\Rightarrow\varepsilon=\frac{\sigma C_\alpha}{\sqrt{n}}.$$

Окончательно получим:

$$P\left\{\acute{x}-\frac{\sigma C_\alpha}{\sqrt{n}}<\xi<\acute{x}+\frac{\sigma C_\alpha}{\sqrt{n}}\right\}=\alpha.$$

28. Доверительный интервал для неизвестного математического ожидания нормально распределённой случайной величины в случае неизвестной дисперсии (без вывода).

$$\acute{x}-\frac{s}{\sqrt{n}}t_\gamma<m_x<\acute{x}+\frac{s}{\sqrt{n}}t_\gamma,\text{где}\,s-\textit{выборочная дисперсия},$$

t_γ находим из распределения Стьюдента по выбранной доверительной вероятности и числу степеней свободы $k=n-1$. Распределение Стьюдента не зависит от дисперсии генеральной совокупности.

Или t_γ можно найти в Excel СТЬЮДЕНТ.ОБР.2Х(1-α;n-1).

29. Доверительный интервал для неизвестной дисперсии нормально распределённой случайной величины (с выводом).

Пусть случайная величина ξ имеет нормальное распределение: $N(m;\sigma)$.

Задана доверительная вероятность (надежность) α. Требуется построить доверительный интервал для параметра σ^2 по выборочной дисперсии $\tilde{S}^2$:

$$\tilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2-\textit{точечная оценка}\,S^2.$$

Рассмотрим случайную величину:

$$U_¿{}^2=\frac{n\tilde{S}^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(x_i-m)^2}{\sigma^2}=\frac{(x_1-m)^2}{\sigma^2}+\frac{(x_2-m)^2}{\sigma^2}+\ldots\frac{(x_n-m)^2}{\sigma^2},\text{где}$$

$$\frac{x_i-m}{\sigma^2}\in N(0;1)\Rightarrow U_¿{}^2\in\chi_n^2.$$

$$P\{u<U_¿{}^2<v\}=\alpha\Rightarrow P\left\{u<\frac{n\tilde{S}^2}{\sigma^2}<v\right\}=\alpha.$$

Тогда:

$$P\left\{\frac{n\tilde{S}^2}{v}<\sigma^2<\frac{n\tilde{S}^2}{u}\right\}=\alpha\Rightarrow P\left\{\frac{(n-1)S^2}{v}<\sigma^2<\frac{(n-1)S^2}{u}\right\}=\alpha$$

где $u=\chi^2_{1,n-1}$- меньшее, а $v=\chi^2_{2,n-1}$ - большее значение с.в., распределённой по закону χ^2 с $n-1$ степенью свободы. Значения находятся по таблицам критических точек распределения χ^2_n по доверительной вероятности α и числу степеней свободы $n-1$ из условий:

$$P\left\{\chi^2 > \chi^2_{1,n-1}\right\} = \frac{1+\alpha}{2};$$

$$P\left\{\chi^2 > \chi^2_{2,n-1}\right\} = \frac{1-\alpha}{2}.$$

EXCEL: $\chi^2_{1,2;n-1} = ХИ2.ОБР\left(\frac{1\pm\alpha}{2}; n-1\right)$.

30. Оценка объёма выборки, необходимого для обеспечения заданной точности при построении доверительных интервалов для математического ожидания и дисперсии нормально распределенной случайной величины: формулы.

Достаточный размер выборки для оценки математического ожидания при выбранной доверительной вероятности α и заданной точности ε:

$$n \geq \frac{C_\alpha^2 S^2}{\varepsilon^2} \Rightarrow n \geq \frac{C_\alpha^2 \sigma^2}{\varepsilon^2},$$

где -доверительная вероятность; ε - заданная точность; S^2- выборочная дисперсия; σ^2-дисперсия.

$$\Phi(C_\alpha^2) = \frac{\alpha}{2};$$

$$C_\alpha = НОРМ.СТ.ОБР\left(\frac{1+\alpha}{2}\right).$$

Достаточный размер выборки для оценки дисперсии при выбранной доверительной вероятности α и заданной точности ε:

$$n \geq \frac{C_\alpha^2 S^4}{\varepsilon^2} + 1.$$

31. Понятие статистической гипотезы. Параметрические и непараметрические гипотезы.

Статистической гипотезой называется любое предположение о свойствах и характеристиках исследуемых генеральных совокупностей, которое может быть проверено на основе анализа выборок.

Гипотезы можно формулировать:
- о виде распределения (параметрические),
- о параметрах распределения (непараметрические),
- об однородности двух выборок (непараметрические),
- о независимости двух выборок (непараметрические),
-о случайности выборки (непараметрические).

Основная (нулевая, H_0) гипотеза – проверяемая. Проверяется и по результатам либо принимается, либо отклоняется и принимается альтернативная (H_1).

Статистический критерий проверки гипотезы – правило, по которому проводится проверка H_0. Не отвечает, верна или нет, лишь помогает ответить противоречат или нет выборочные данные выдвинутой гипотезе (можно принять или следует отклонить).

H_0	Решение	Вероятность	Примечание

 vk.com/id446425943

Верна	Принимается	α	Доверительная вероятность
	Отклоняется	$1-\alpha$	Уровень значимости (ошибки I рода)
Неверна	Принимается	β	Вероятность ошибки II рода
	Отклоняется	$1-\beta$	Мощность критерия

Предположение, которое касается неизвестного значения параметра распределения, входящего в некоторое параметрическое семейство распределений, называется ***параметрической гипотезой***

Предположение, при котором вид распределения не рассматривается (т.е. не предполагается, что оно входит в некоторое параметрическое семейство распределений), называется ***непараметрической гипотезой***.

Непараметрические методы позволяют исследовать данные без допущений о характере распределения переменных. Так как в этих тестах обрабатывается не само измеренное значение, а его ранг, то эти тесты нечувствительны к выбросам. Непараметрические тесты могут применяться в тех случаях, когда переменные измерены при помощи порядковой или метрической шкалы. Существуют тесты, предназначенные для анализа номинальных данных.

32. Общий алгоритм проверки статистической гипотезы. Виды критической области.

Алгоритм проверки статистических гипотез:

- Формулируем основную и альтернативную гипотезы.
- Задаем уровень значимости.
- Выбираем статистику - критерий проверки гипотезы.
- Определяем критическую область.
- Вычисляем значение статистики по выборке.
- Сравниваем критическое значение с границами критической области.
- Делаем вывод.

Область, при попадании в которую критерия $K_{набл}$ отвергается основная гипотеза в пользу альтернативной, называется ***критической областью***.

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

Критические значения отделяют критическую область от области принятия гипотезы.

В зависимости от альтернативной гипотезы различают одностороннюю и двустороннюю критические области. В свою очередь односторонняя область может быть левосторонней и правосторонней.

Если значение статистики попало в область принятия гипотезы, то гипотеза H_0 принимается.

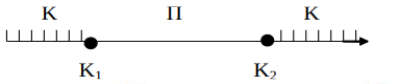
Если значение статистики попало в критическую область, то гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1.

Если значение статистики $<K_1$, то область правосторонняя

Если значение статистики $> K_2$, то область левосторонняя

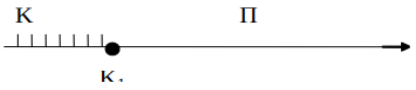
$H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$ двусторонняя область



$H_1: \theta > \theta_0$ правосторонняя область

$H_1: \theta < \theta_0$ левосторонняя область



K_1 и K_2 называются границами критической области.

33. Ошибки первого и второго рода. Уровень значимости и мощность критерия.

Ошибка первого рода происходит, если отвергаем верную основную гипотезу.

Ошибка второго рода происходит, если принимаем основную гипотезу, когда она неверна.

Например, H_0: проверяемая деталь стандартная. Ошибка I рода: стандартную деталь отбраковали; ошибка II рода – бракованную деталь посчитали стандартной.

Уровнем значимости $(1-\alpha)$ гипотезы называют вероятность совершить ошибку первого рода, то есть отклонить верную основную гипотезу (0,1; 0,05 или 0,01)

Статистика (критерий) есть специальная функция от элементов выборки, по значениям которой принимают решение о принятии или отклонении основной гипотезы.

Мощность критерия $(1-\beta)$ - это вероятность правильно отвергнуть нулевую гипотезу, то есть отвергнуть ее, когда она неверна.

α - доверительная вероятность,

$1-\alpha$ - уровень значимости (вероятность ошибки I рода),

β - вероятность ошибки II рода,

$1-\beta$ - мощность критерия.

H_0	Решение	Вероятность	Примечание
Верна	Принимается	α	Доверительная вероятность
	Отклоняется	$1-\alpha$	Уровень значимости (ошибки I рода)
Неверна	Принимается	β	Вероятность ошибки II рода
	Отклоняется	$1-\beta$	Мощность критерия

34. Проверка гипотезы о величине математического ожидания (при известной дисперсии и при неизвестной дисперсии). Определение границ критической области для различных видов альтернативной гипотезы.

Предположим, что изучаемая случайная величина распределена по нормальному закону с параметрами $m=E(X)$, $\sigma^2=V(X)$, которые не известны. По выборочному среднему можно выдвинуть гипотезу о равенстве математического ожидания некоторому числу.

Сформулируем основную гипотезу: $H_0: m=m_0$;
и альтернативную: $H_1: m\neq m_0$.

Критическая область будет двусторонней.

При известной дисперсии выборочную статистику вычисляют по формуле:

$$Z^{¿}=\frac{\acute{x}-m_0}{\frac{\sigma}{\sqrt{n}}}=\frac{\acute{x}-m_0}{\sigma}\cdot\sqrt{n}.$$

Если нулевая гипотеза верна, то эта случайная величина $Z^{¿}$ распределена по нормальному закону с параметрами $Z^{¿}\in N(0;1)\Rightarrow P\{k_1<Z^{¿}<k_2\}$.

$$k_2=\Phi_0^{-1}\left(\frac{\alpha}{2}\right); k_1=-k_2.$$

$$\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{\frac{-t^2}{2}}\,dt \text{ находим из таблиц Лапласа}.$$

$$Excel: k_1=НОРМ.СТ.ОБР\left(\frac{\alpha}{2}\right), k_2=НОРМ.СТ.ОБР\left(\frac{\alpha+1}{2}\right).$$

Для правосторонней области ($m>m_0$) необходимо $Z^{¿}<k_2$;
для левосторонней области ($m<m_0$) необходимо $Z^{¿}>k_1$.

Дисперсия неизвестна:

$$Z^{¿}=\frac{\acute{x}-m_0}{\frac{S}{\sqrt{n}}}=\frac{\acute{x}-m_0}{S}\cdot\sqrt{n}, k_1<Z^{¿}<k_2.$$

Если нулевая гипотеза верна, то эта случайная величина распределена по закону Стьюдента с $n-1$ степенью свободы.
$k_2=t_{1-\alpha,n-1}$ $k_1=-k_2-¿$ распределение Стьюдента.
$Excel: k_1=-СТБЮДЕНТ.ОБР(1-\alpha,n-1); k_2=СТБЮДЕНТ.ОБР(\alpha,n-1)$.

Для правосторонней области ($m>m_0$) необходимо $Z^{¿}<k_2$;
для левосторонней области ($m<m_0$) необходимо $Z^{¿}>k_1$.

35. Проверка гипотезы о величине дисперсии. Определение границ критической области для различных видов альтернативной гипотезы.

Предположим, что изучаемая случайная величина распределена по нормальному закону с параметрами $m=E(X)$, $\sigma^2=V(X)$, которые не известны. По выборочной дисперсии можно выдвинуть гипотезу о равенстве дисперсии некоторому числу.

Сформулируем основную гипотезу: $H_0: \sigma^2=\sigma_0^2$;
и альтернативную: $H_1: \sigma^2\neq\sigma_0^2$.

Критическая область будет двусторонней.

Выборочную статистику вычисляют по формуле:

$$Z^{*}=\frac{(n-1)S^2}{\sigma_0^{\,2}}.$$

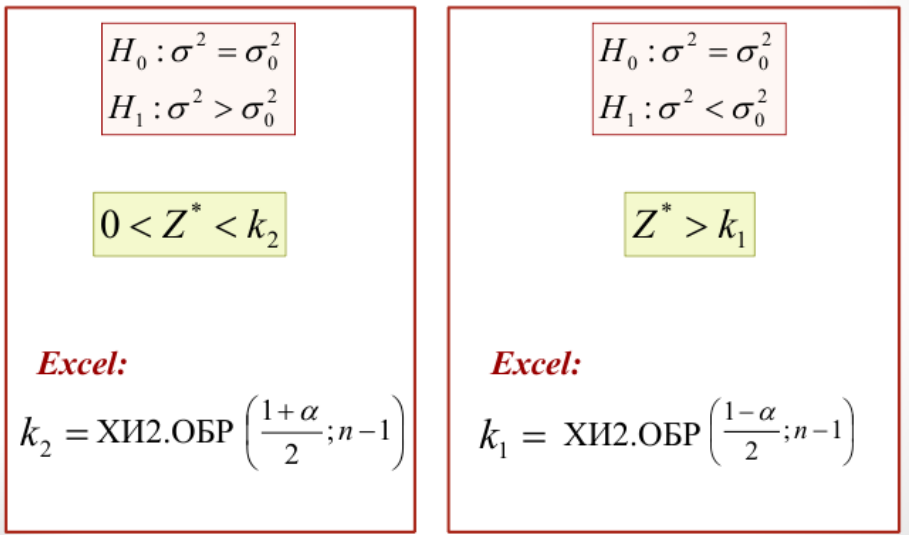
Если основная гипотеза верна, то эта случайная величина распределена по закону χ^2 с $n-1$ степенью свободы.

Двусторонняя критическая область:

$$P\{\chi^2>k_2\}=\frac{1-\alpha}{2};P\{\chi^2>k_1\}=\frac{1+\alpha}{2}.$$

$$Excel: k_{1,2}=ХИ2.ОБР\left(\frac{1\pm\alpha}{2}\right).$$

Односторонние критические области:



36. Проверка гипотезы об однородности выборок.

Две выборки называются ***однородными***, если они одинаково распределены. Другими словами: обе выборки взяты из одной и той же генеральной совокупности.

Пусть имеются две выборки нормально распределенных случайных величин с параметрами: $m_1=E(\xi_1)$, $\sigma_1^{\,2}=V(\xi_1)$, $m_2=E(\xi_2)$, $\sigma_2^{\,2}=V(\xi_2)$.

Для каждой случайной величины по имеющимся выборкам объемов n_1 и n_2 вычисляют выборочные средние и выборочные дисперсии, анализируют полученные результаты и выдвигают гипотезы. Для математического ожидания:

- основную гипотезу: $H_0: m_1=m_2$;
- и альтернативную: $H_1: m_1\neq m_2$.

Для дисперсии:

- основную гипотезу: $H_0:\sigma_1^{\,2}=\sigma_2^{\,2}$;
- и альтернативную: $H_1:\sigma_1^{\,2}\neq\sigma_2^{\,2}$.

Случай 1. Дисперсии известны

Выборочную статистику вычисляют по формуле:

$$Z^{\text{¿}} = \frac{\acute{x}_1 - \acute{x}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}.$$

Если основная гипотеза верна, то эта случайная величина распределена по нормальному закону с параметрами $Z^{\text{¿}} \in N(0;1)$.

$$Excel: k_2 = НОРМ.СТ.ОБР\left(\frac{1+\alpha}{2}\right), k_1 = -k_2.$$

Случай 2. Дисперсии неизвестны

Выборочную статистику вычисляют по формуле:

$$Z^{\text{¿}} = \frac{\acute{x}_1 - \acute{x}_2}{\sqrt{\frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Если основная гипотеза верна, то эта случайная величина имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

$$Excel: k_2 = СТЬЮДЕНТ.ОБР.2X\left(1-\alpha, n_1 + n_2 - 2\right), k_1 = -k_2.$$

Для проверке гипотезы о равенстве дисперсий выборочную статистику вычисляют по формуле:

$$Z^{\text{¿}} = \frac{S^{\,2}_{max}}{S^{\,2}_{min}}.$$

Если основная гипотеза верна, то эта случайная величина распределена по закону Фишера с $n_{max} - 1, n_{min} - 1$ степенями свободы. Критическая область будет правосторонней.

$$Excel: k_2 = F.ОБР.ПХ\left(1-\alpha, n_{max} - 1, n_{min} - 1\right), k_1 = -k_2.$$

37. Критерий Пирсона для проверки гипотезы о виде распределения для непрерывной случайной величине.

Критерием согласия называют статистический критерий проверки гипотезы о виде закона распределения (непараметрическая гипотеза).

Если рассматриваемая выборка – выборка значений дискретных случайных величин, то для проверки непараметрических гипотез применяется критерий Пирсона. Если рассматриваемая выборка – выборка значений непрерывной случайной величины, то для проверки непараметрических гипотез применяется критерий Пирсона или критерий Колмогорова.

После визуального изучения кумуляты, гистограммы и анализа полученных оценок числовых характеристик выдвигается гипотеза о распределении:

$H_0: \xi \in N(\underline{x}, S)$;

$H_1: \xi \notin N(\underline{x}, S)$.

Для непрерывной случайной величины:

- Строится интервальный статистический ряд.
- Вычисляется выборочные характеристики.
- Вычисляется выборочная статистика:

$2*¿=\sum_{i=1}^{r} ¿¿¿¿$
χ
n - объем выборки;
l_i - частота попадания выборочного значения в интервал с номером i;
p_i- это теоретические (гипотетические!!!) вероятности, которые вычисляются исходя из предположения о виде закона распределения:
$p_i = P\{\xi \in J_i\} = F_\xi(\tilde{x}_{i+1}) - F_\xi(\tilde{x}_i)$, *где*
$F_\xi(x)$ - гипотетическая функция распределения (предполагаемая нами).

В предположении, что закон распределения нормальный:

$$p_i = F_\xi(\tilde{x}_{i+1}) - F_\xi(x_i^{\sim}) = \Phi\left(\frac{\tilde{x}_{i+1} - \acute{x}}{\sigma}\right) - \Phi\left(\frac{\tilde{x}_i - \acute{x}}{\sigma}\right), \text{где } \Phi(x) - \text{функция Лапласа}.$$

Количество интервалов k рекомендуется рассчитывать по формуле Стерджеса. Длину $i-¿$го интервала принимают равной $\delta = \frac{x_n - x_1}{k}$, где $x_n - ¿$ наибольшее, а $x_1 - ¿$ наименьшее значение в вариационном ряду.

Определим границу критической области. Так как статистика Пирсона измеряет разницу между эмпирическим и теоретическим распределениями, то чем больше ее наблюдаемое значение $\chi^{2*¿¿}$, тем сильнее довод против основной гипотезы. Поэтому критическая область для этой статистики всегда правосторонняя.

Границы ищутся по таблицам распределения Хи-квадрат по заданному уровню значимости $1-\alpha$ и степенями свободы v из условия:
$P\{\chi^2 > k_2\} = 1-\alpha$.
для непрерывной случайной величины: $v = r - w - 1$, где
w - число параметров распределения, оцениваемых по выборке,
r - число интервалов.

38. Критерий Колмогорова для проверки гипотезы о виде распределения для непрерывной случайной величины.

Если закон распределения случайной величины известен (т. е. вид функции и числовые значения параметров известны), то можно применять ***критерий согласия Колмогорова***. При его использовании сравниваются эмпирическая $F^{¿}_n(x)$ и гипотетическая (предполагаемая) $F(x)$ функции распределения. Отметим, что критерий Колмогорова применяется для проверки гипотез о законах распределения только непрерывных случайных величин.

Критерий Колмогорова проверки гипотезы о виде закона распределения является наиболее простым, однако этот критерий можно применять только в том случае, когда гипотетическое распределение $F(x)$ полностью известно, т. е. когда известен не только вид функции распределения, но и все входящие в нее параметры.

Общий алгоритм применения критерия Колмогорова:

Пусть из генеральной совокупности с функцией распределения $F(x)$ произведена выборка объемом n.

1. Представить **р**езультаты выборки представить в виде интервального ряда или расположить в возрастающем порядке.

2. Найти эмпирическую функцию распределения:

$$F^{\nearrow}_n(x)=\frac{p^{\nearrow}_i}{n},$$

где $p^{\nearrow}_i$ – накопленная частота.

3. Вычислить значения предполагаемой теоретической функции распределения $F(x)$ по данным выборки.

4. Для каждого значения x_i вычислить $\left|F^{\nearrow}_n(x_i)-F(x_i)\right|$.

5. Вычислить наблюдаемое значение выборочной статистики Колмогорова $\lambda_{набл}=max\left|F^{\nearrow}_n(x_i)-F(x_i)\right|\cdot\sqrt{n}$.

6. По заданному уровню значимости α из таблицы критических значений распределения Колмогорова найти $\lambda_{кр}$.

7. Если $\lambda_{набл}<\lambda_{кр}$, то наблюдаемые данные хорошо согласуются с теоретическим распределением; иначе проверяемая гипотеза отклоняется.

39. Понятие корреляционного анализа.

Постановка проблемы

Переменные X,Y – случайные величины.

1. Существует ли связь между двумя или более переменными?
2. Какой тип имеет эта связь?
3. Насколько она сильна?
4. Какой прогноз можно сделать, основываясь на этой связи?

Корреляционная зависимость между X и Y:

Корреляционный анализ – статистический метод, позволяющий определить, существует ли зависимость между переменными и насколько она сильна.

Однако корреляционный анализ не предполагает выявления каузальных связей, поэтому при интерпретации результатов формулировки типа «переменная x влияет на переменную y» или «переменная x зависит от переменной y» недопустимы.

Различают ***парную и множественную корреляции***. Парная корреляция характеризует тип, форму и плотность связи между двумя признаками, множественная - между несколькими.

Корреляционная зависимость возникает чаще всего там, где одно явление находится под воздействием большого числа факторов, действующих с разной силой, поэтому существуют специальные меры корреляционной связи, называемые коэффициентами корреляции. Коэффициенты (в статистике их общее количество исчисляется десятками) показывают степень взаимосвязи явлений (плотность корреляционной связи, иногда исследователи говорят об интенсивности связи) и характер этой связи (направленность). Связь может быть прямой и обратной. Например, чем старше избиратель, тем более активно он участвует в выборах. Чем выше уровень доходов людей, тем в меньшей степени они склонны участвовать в выборах в качестве избирателей (обратная связь). Чем выше коэффициент корреляции между двумя переменными, тем точнее можно предсказать значения одной из них по значениям другой. Характер связи также определяется в категориях «монотонная» (направление

 vk.com/id446425943

изменения одной переменной не меняется при изменении второй переменной) и «немонотонная» связь. Помимо оценки плотности и направленности связи необходимо учитывать надежность (достоверность) связи.

Корреляционный анализ последовательно решает три практические задачи:

- определение корреляционного поля и составление корреляционной (в данном случае это комбинированная) таблицы;
- вычисление выборочных корреляционных отношений или коэффициентов корреляции;
- проверка статистической гипотезы значимости связи.

Коэффициент корреляции не содержит информации о том, является ли данная связь между ними причинно-следственной или сопутствующей (порожденной общей причиной). Этот вопрос исследователь должен решать самостоятельно на основе содержательных представлений о структуре, динамике изучаемых социальных объектов, корреляций между изучаемыми признаками, использовать иные способы статистического анализа (регрессионный, факторный, дискриминантный, путевой и т.д.). Но величина коэффициента позволяет оценить плотность связи как меньшую (незначимую) или большую. По знаку коэффициента корреляции для порядковых рядов мы можем сказать, является ли эта связь прямой или обратной (для номинальных рядов знак коэффициента не несет смысловой нагрузки).

Для установления корреляционной связи между двумя признаками необходимо доказать, что все другие переменные не оказывают воздействия на отношения двух переменных, являющихся предметом изучения. В противном случае возникает ситуация ложной корреляции. Секрет возникновения ложной корреляции заключается в том, что у двух явлений, связь которых формально подкрепляется наличием статистической связи, есть общая причина, в равной степени влияющая на каждое из них.

40. Понятие регрессионного анализа.

Регрессионный анализ – статистический метод, позволяющий описать тип связи между переменными (линейная или нелинейная).

Каждому значению одной из переменных (например, x) соответствует условное математическое ожидание другой (например, y):

$E\left[\xi_2 / \xi_1 = x_i\right] = \varphi(x) \neq const$ *или*

$E\left[\xi_1 / \xi_2 = y_i\right] = \omega(y) \neq const$ – уравнения регрессии.

Регрессионный анализ с помощью коэффициента регрессии позволяет количественно прогнозировать изменения одной переменной при изменении другой.

Для описания связи могут использоваться различные математические функции, основными из которых являются:

- линейная,
- логарифмически линейная регрессия;
- нелинейная (например, $Y = \beta_0 + \beta_1 x^2 + \beta_2 x$);
- экспоненциальная ($Y = \beta_0 e^{\beta_1 x}$);

- логистическая.

Простая линейная регрессия или множественная регрессия могут применяться для непрерывных признаков, например, давление, вес.

Логистическая регрессия применима в тех случаях, когда зависимые признаки являются бинарными (например, умер/жив, выздоровел/не выздоровел).

Математическое уравнение, которое оценивает линию простой линейной регрессии: $Y=\beta_0+\beta_1 x$, где

x – независимая или объясняющая переменная;

β_0 – свободный член (пересечение) линии оценки; это значение Y, когда $x=0$;

β_1 – угловой коэффициент или градиент оценённой линии; он представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем x на одну единицу. Коэффициент b называют коэффициентом регрессии.

Математически решение уравнения линейной регрессии сводится к вычислению параметров β_0 и β_1 таким образом, чтобы точки исходных данных корреляционного поля как можно ближе лежали к прямой регрессии.

Логарифмически линейная регрессия: $lnY=\ln\beta_0+\beta_1 lnx$.

41. Интерпретация величины выборочного коэффициента корреляции.

Коэффициент корреляции (Пирсона) измеряет силу и направление связи между переменными.

Генеральная совокупность (теоретический коэффициент корреляции)	Выборочная совокупность (эмпирический коэффициент корреляции)
ρ	$\widetilde{\rho}$

Коэффициент корреляции

Теория вероятностей:

Теория вероятностей:

$$\rho=\frac{V_{xy}}{\sigma_x\sigma_y};$$

Статистика:

$$\rho=\frac{\widetilde{V}_{xy}}{S_x S_y}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i-\acute{x}\right)\left(y_j-\acute{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2\sum_{j=1}^{n}\left(y_j-\acute{y}\right)^2}}.$$

Свойства коэффициента корреляции:

1. Коэффициент корреляции изменяется на отрезке от –1 до +1.

2. Если между переменными существует сильная прямая линейная связь, то значение будет близко к +1.

 vk.com/id446425943

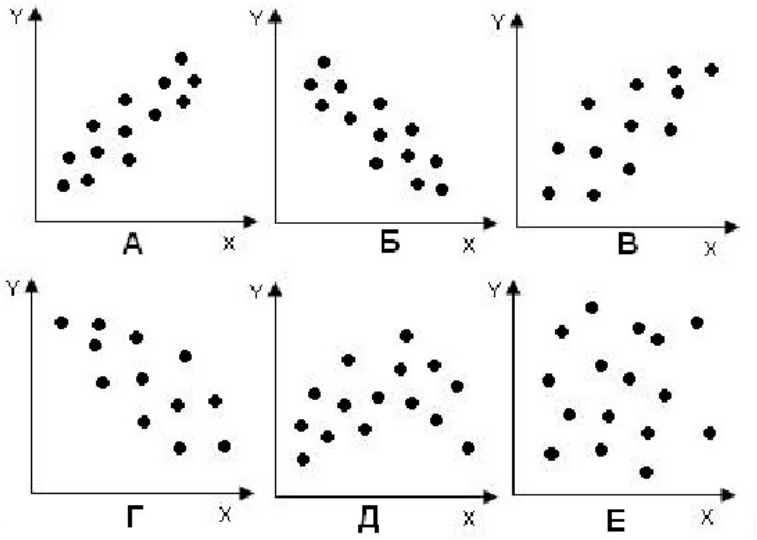
3. Если между переменными существует сильная обратная линейная связь, то значение будет близко к –1.

4. Когда между переменными нет линейной связи или она очень слабая, значение будет близко к 0.

Интерпретация коэффициента корреляции:

Значение	Уровень линейной связи между переменными
от 0,75 до 1,00	Очень высокая прямая зависимость
от 0,50 до 0,74	Высокая прямая зависимость
от 0,25 до 0,49	Средняя прямая зависимость
от 0,00 до 0,24	Слабая прямая зависимость
от -0,24 до 0,00	Слабая обратная зависимость
от -0,49 до -0,25	Средняя обратная зависимость
от -0,74 до -0,50	Высокая обратная зависимость
от -1,00 до -0,75	Очень высокая обратная зависимость

Теснота линейной зависимости между случайными величинами



А, В – прямая линейная связь, на А более сильная, Б, Г – обратная линейная связь, на Б более сильная, Д – нелинейная связь (квадратичная), Е – отсутствие связи.

42. Проверка значимости коэффициента корреляции.

Предположение:

Случайные величины X и Y имеют нормальное распределение.

Рассмотрим гипотезы:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

(H_0)

Основная гипотеза (H_0) утверждает, что не существует корреляции между признаками X и Y в генеральной совокупности.

Альтернативная гипотеза (H_1) утверждает, что корреляция между признаками X и Y в генеральной совокупности значима.

$Z^{\cdot} = \tilde{\rho}\sqrt{\frac{n-2}{1-\tilde{\rho}^2}} - статистика.$

Если нулевая гипотеза верна, то статистика $Z^{\cdot}$ имеет распределение

Стьюдента с числом степеней свободы $n-2$. Критическая область – двусторонняя.

43. Модель парной линейной регрессии.

Зависимая переменная (Y) - результирующий признак (объясняемая переменная). Независимая переменная (X) - фактор (объясняющая переменная).

Двумерная выборка $(x_i; y_i)$ - упорядоченный набор значений переменных.

Математическое уравнение, которое оценивает линию простой линейной регрессии: $Y=\beta_0+\beta_1 x$, где

x – независимая или объясняющая переменная;

β_0 – свободный член (пересечение) линии оценки; это значение Y, когда $x=0$;

β_1 – угловой коэффициент или градиент оценённой линии; он представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем x на одну единицу. Коэффициент b называют коэффициентом регрессии.

Математически решение уравнения линейной регрессии сводится к вычислению параметров β_0 и β_1 таким образом, чтобы точки исходных данных корреляционного поля как можно ближе лежали к прямой регрессии.

Коэффициенты $\beta_0 и \beta_1$ обычно находят по методу наименьших квадратов:

$$\begin{cases} \beta_0 \sum x_i+\beta_1 \sum x_i^2=\sum x_i y_i \\ n\beta_0+\beta_1 \sum x_i=\sum y_i \end{cases}$$

Решаем полученную систему уравнений любым методом (например, методом подстановки или методом Крамера) и получаем формулы для нахождения коэффициентов по методу наименьших квадратов (МНК):

$$\beta_0=\frac{\sum x_i \cdot \sum x_i y_i-\sum x_i^2 \cdot \sum y_i}{\left(\sum x_i\right)^2-n\sum x_i^2}; \beta_1=\frac{\sum x_i \cdot \sum y_i-n\sum x_i y_i}{\left(\sum x_i\right)^2-n\sum x_i^2}.$$

При данных β_0 и β_1 функция $F(a,b)=\sum\left(\tilde{y}_i-\left(\beta_0+\beta_1 x_i\right)\right)^2$ принимает наименьшее значение, то есть, при данных β_0 и β_1 сумма квадратов отклонений экспериментальных данных $\tilde{y}_i$ от найденной прямой будет наименьшей.

44. Условия Гаусса – Маркова.

Для того чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможных результаты, случайная ошибка ε_i (или неучтённые в модели факторы) должна удовлетворять определенным условиям, известным как условия Гаусса-Маркова:

1. Ошибки наблюдения являются случайными величинами, распределенными по нормальному закону.

2. Математическое ожидание всех случайных величин ошибок равно нулю: $E[\varepsilon_i]=0, i=1,\ldots,n$.

Это означает, что иногда случайная ошибка будет положительной, иногда отрицательной, но она не должна иметь систематического смещения. Фактически, если уравнение регрессии включает свободный член, то разумно предположить, что это условие выполнено автоматически, так как роль

 vk.com/id446425943

константы состоит в определении любой систематической тенденции объясняемой переменной, которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

3. Все ошибки наблюдения имеют одинаковую (но неизвестную!) дисперсию (условие гомоскедастичности): $V[\varepsilon_i]=\sigma^2, i=1,\ldots,n$.

Если условие постоянства дисперсии не выполняется, то оценки, найденные по методу наименьших квадратов, будут неэффективны.

4. Все пары случайных величин ошибок независимы друг от друга.

Например, если случайная ошибка в одном наблюдении велика и положительна, то это не должно обусловливать систематическую тенденцию к тому, что в следующем наблюдении она будет обязательно мала и отрицательна (или велика и отрицательна, или мала и положительна). Случайные ошибки должны быть абсолютно независимы друг от друга, то есть ковариация между ними должна быть равна нулю.

Это условие часто нарушается в случае, когда наши данные являются временными рядами. В случае, когда условие некоррелированности ошибок не выполняется, то говорят об автокорреляции ошибок.

Замечания:

1. Случайные величины зависимых переменных распределены по нормальному закону с одинаковыми дисперсиями σ^2 и с математическим ожиданием: $E[y_i]=\beta_0+\beta_1 x_i$.

2. Независимая переменная может быть *случайной* или *неслучайной*. Если случайна, то тогда еще одно условие: x_i и ε_i *независимые* случайные величины.

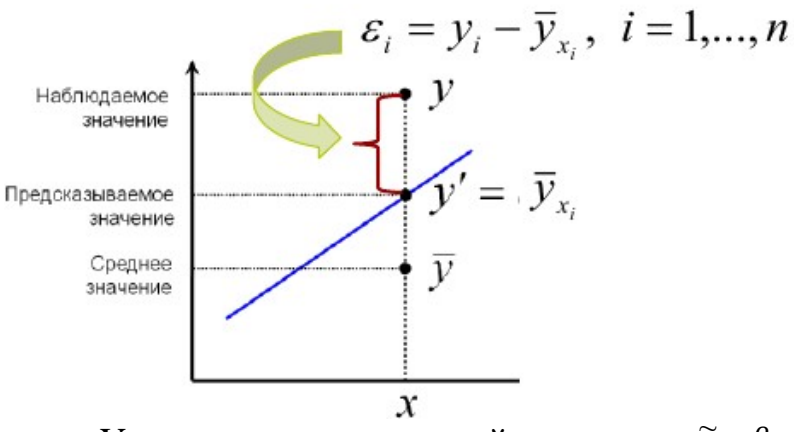
3. Математическое ожидание случайной величины y_i будет зависеть от значения x_i, и является условным математическим ожиданием: $E[y/x=x_i]=\phi(x)$.

Эмпирическая функция регрессии – функция регрессии, служащая статистической оценкой модельной функции регрессии.

45. Оценка параметров регрессионной модели с помощью метода наименьших квадратов. Система нормальных уравнений.

Метод наименьших квадратов (МНК, OLS).

Идея МНК: минимизация суммы квадратов ошибок.



Уравнение эмпирической регрессии: $\tilde{y}_i=\beta_0+\beta_1 x_i$.

Математически решение уравнения линейной регрессии сводится к вычислению параметров β_0 и β_1 таким образом, чтобы точки исходных данных корреляционного поля как можно ближе лежали к прямой регрессии.
Коэффициенты $\beta_0 и \beta_1$ обычно находят по методу наименьших квадратов (система нормальных уравнений):

$$\begin{cases} \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \\ n\beta_0 + \beta_1 \sum x_i = \sum y_i \end{cases}$$

Решаем полученную систему уравнений любым методом (например, методом подстановки или методом Крамера) и получаем формулы для нахождения коэффициентов по методу наименьших квадратов (МНК):

$$\beta_0 = \frac{\sum x_i \cdot \sum x_i y_i - \sum x_i^2 \cdot \sum y_i}{\left(\sum x_i\right)^2 - n\sum x_i^2}; \beta_1 = \frac{\sum x_i \cdot \sum y_i - n\sum x_i y_i}{\left(\sum x_i\right)^2 - n\sum x_i^2}.$$

При данных β_0 и β_1 функция $F(a,b) = \sum \left(\tilde{y}_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2$ принимает наименьшее значение, то есть, при данных β_0 и β_1 сумма квадратов отклонений экспериментальных данных $\tilde{y}_i$ от найденной прямой будет наименьшей.

В Excel задача решается с помощью встроенной функции ЛИНЕЙН, аргументами которой являются истинные значения $x_i и y_i$, а так же Анализ данных → регрессия.

46. Интерпретация коэффициентов функции регрессии.
Математическое уравнение, которое оценивает линию простой линейной регрессии: $Y = \beta_0 + \beta_1 x$, где

x – независимая или объясняющая переменная;

β_0 – свободный член (коэффициент смещения) линии оценки; это значение Y, когда $x = 0$;

β_1 – угловой коэффициент или коэффициент наклона; характеризует чувствительность показателя к изменению фактора; он представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем x на одну единицу. Коэффициент b называют коэффициентом регрессии.

47. Суммы квадратов: полная, объяснённая, остаточная. Коэффициент детерминации и его интерпретация.

Остаточная сумма квадратов (Residual Sum of Squares):

$$RSS = \sum \left(y_i - \tilde{y}_{x_i}\right)^2 = \sum \varepsilon_i^2.$$

Объясненная сумма квадратов отклонений (объясненных регрессией) (Explained Sum of Squares):

$$ESS = \sum \left(\tilde{y}_{x_i} - \acute{y}\right)^2.$$

Полная сумма квадратов (Total Sum of Squares):

$$TSS = \sum \left(y_i - \acute{y}\right)^2.$$

$$TSS = ESS + RSS.$$

Парный коэффициент детерминации – это мера вариации зависимой переменной, определяемая отношением объяснимой вариации к общей вариации:

$\widetilde{R}^2_{xy}=\frac{ESS}{TSS}$.

Интерпретация: «коэффициент детерминации показывает, какая доля дисперсии независимой переменной y определяется дисперсией соответствующей функции регрессии».

$0\le\widetilde{R}^2_{xy}\le 1$.

Чем ближе коэффициент к 1, тем больше есть основания предполагать, что уравнение регрессии статистически значимо и линейная функция фактора x оказывает сильное воздействие на результирующий признак y.

$0\le\widetilde{R}^2_{xy}\le 0,09$ – использование линейной регрессионной модели для приближенной оценки взаимосвязи x и y статистически необоснованно.

$0,09\le\widetilde{R}^2_{xy}\le 0,49$ – использование линейной регрессионной модели для приближенной оценки взаимосвязи x и y возможно, но затем следует провести анализ значимости модели.

$0,49\le\widetilde{R}^2_{xy}\le 1$ – есть все основания для использования линейной регрессионной модели для приближенной взаимосвязи x и y.

48. Проверка значимости уравнения регрессии в целом.

$H_0: R^2=0$.

$H_1: R^2\neq 0$.

Основная гипотеза утверждает, что не существует статистически значимой линейной зависимости между признаками x и y в генеральной совокупности.

Альтернативная гипотеза утверждает, что признаки x и y в генеральной совокупности связаны линейной зависимостью.

Статистика:

$F=\frac{\widetilde{R}^2_{xy}}{1-\widetilde{R}^2_{xy}}\cdot\frac{n-k-1}{k}=\frac{ESS}{RSS}\cdot\frac{n-k-1}{k}$. Если нулевая гипотеза верна, то статистика F имеет распределение Фишера с числом степеней свободы числителя: $df_1=k$, а знаменателя - $df_1=n-k-1$, где k – число факторов в уравнении, n – объём выборки. В Excel - $F.ОБР$.

Критическая область – правосторонняя.

49. Средняя ошибка аппроксимации.

$\acute{A}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widetilde{y}_{x_i}}{y_i}\right|\cdot 100\,\%$

– показывает, на сколько в среднем отклоняется предсказанное значение от наблюдаемого (в %)

Ошибка аппроксимации в пределах 5-7% свидетельствует о хорошем подборе модели к исходным данным.

Допустимый предел значений A − не более 8-10%.