Раздел 1. Теория случайных чисел.

Все события делятся на детерминированные, случайные и неопределенные.

Если событие наступает в эксперименте всегда, оно называется *достоверным*, если никогда – *невозможным*. Это детерминированные события.

Статистическое определение вероятности: Если в опыте, повторяющемся n раз, событие появляется m_A раз, тогда относительная частота наступления события: $\lim_{n\to\infty} hn(A) = \frac{m_A}{n} = P(A)$.

P(A) – вероятность наступления события A.

Для ***достоверного*** события Ω: P(Ω)=1. Для ***невозможного*** события ∅: P(∅)=0.

$0 \le P(A) \le 1$, т.к. $0 \le m_A \le n \rightarrow 0 \le hn(A) \le 1$

Ω $m_A=n$ $hn(A)=1$

∅ $m_A=0$ $hn(A)=0$

Все мыслимые взаимоисключающие исходы опыта называются ***элементарными событиями***. Наряду с ними можно наблюдать более сложные события – комбинации элементарных.

Несколько событий в данном опыте называются ***равновозможными***, если появление одного из них не более возможно, чем другого.

Классическое определение вероятности: Если n-общее число элементарных событий и все они равновозможные, то вероятность события A:

$$P(A) = \frac{m_A}{n},$$

где m_A- число исходов, благоприятствующих появлению события A.

Раздел 2. Сложные события.

Теория сложных событий позволяет по вероятностям простых событий определять вероятности сложных. Она базируется на теоремах сложения и умножения вероятностей.

1) ***Суммой (объединением)*** двух событий A и B называется новое событие A+B, заключающееся в проявлении хотя бы одного из этих событий.

2) ***Произведением (пересечением)*** двух событий A и B называется новое событие AB, заключающееся в одновременном проявлении обоих событий. A*B=AB, AA=A, ABA=AB.

3) Событие A влечет за собой появление события B, если в результате наступления события A всякий раз наступает событие B. $A \subset B$

A=B: $A \subset B$, $B \subset A$

Два события называются ***несовместными***, если появление одного из них исключает возможность появления другого.

Если события несовместны, то AB=∅.

События $A_1, A_2, \ldots A_n$ образуют полную группу событий в данном опыте, если они являются несовместными и одно из них обязательно происходит:

$$\mathbf{A_iA_j=\varnothing\ (i\neq j,\ i,j=1,2\ldots n)}$$

$$\mathbf{A_1+A_2+\ldots+A_n=\Omega}$$

$\overline{A}$ - событие противоположное событию A, если оно состоит в не появлении события A.

A и $\overline{A}$ - полная группа событий, т.к. $A+\overline{A}=\Omega$, $A\overline{A}=\varnothing$.

Теорема сложения вероятностей.

Вероятность суммы ***несовместных событий*** *равна сумме вероятностей событий:*

$$\mathbf{P(A+B+C+\ldots) = P(A) + P(B) + P(C) +\ldots}$$

Следствие. *Если события $A_1+A_2+\ldots+A_n$ - полная группа событий, то сумма их вероятностей равна 1.*

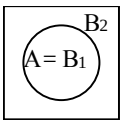
$$\left.\begin{array}{l} A_i A_j = \varnothing \\ \sum_{i=1}^{n} A_i = \Omega \end{array}\right\} \Rightarrow P(A_1 + A_2 + ... + A_n) = 1$$

$$P(A+\overline{A}) = P(A) + P(\overline{A}) = 1$$

Вероятность наступления двух ***<u>совместных событий</u>*** *равна:*

P(A+B) = P(A) + P(B) - P(AB)

P(A+B+C) = P(A) + P(B) + P(C) – P(AB) – P(AC) – P(BC) – P(ABC)



<u>Теорема.</u> Если A⊂B, то P(A) ≤ P(B).
$B=B_1+B_2$ ($B_1=A$) $P(B)=P(B_1) + P(B_2)= P(A) + P(B_2)$

<u>Теорема умножения вероятностей. Условные вероятности.</u>

Опыт повторяется n раз, m_B раз наступает событие B, m_{AB} раз наряду с событием B наступает событие A.

$$hn(B) = \frac{m_B}{n} \qquad hn(AB) = \frac{m_{AB}}{n}$$

Рассмотрим относительную частоту наступления события A, когда событие B уже наступило:

$$hn\left(A/B\right) = \frac{m_{AB}}{m_B} = \frac{m_{AB}}{n} \div \frac{m_B}{n} = \frac{hn(AB)}{hn(B)}$$

$\boxed{P\left(A/B\right) = \frac{P(AB)}{P(B)}}$ - ***условная вероятность события A по событию B*** – вероятность события A, когда событие B уже наступило.

<u>Свойства условных вероятностей.</u>

Свойства условных вероятностей аналогичны свойствам безусловных вероятностей.

1. $0 \le P(A/B) \le 1$, т.к. $P\left(A/B\right) = \frac{P(AB)}{P(B)}$; $AB \subset B$, $P(AB) \le P(B)$
2. P(A/A)=1
3. B⊂A, ➔ P(A/B)=1
4. $P\left(\Omega/B\right) = 1$ $\quad B\Omega = B$
   $P\left(\phi/B\right) = 0$ $\quad B\phi = \phi$
5. **P[(A+C)/B] = P(A/B) + P(C/B)** – Если события A и C несовместны
   P[(A+C)/B] = P(A/B) + P(C/B) - P(AC/B) – Если события A и C совместны

$$P\left(AC/B\right) = \frac{P((A+C)B)}{P(B)} = \frac{P(AB+CB)}{P(B)} = \frac{P(AB)}{P(B)} + \frac{P(CB)}{P(B)} = P\left(A/B\right) + P\left(C/B\right)$$

<u>Теорема</u>. *Вероятность произведения двух событий равна произведению вероятности одного события на условную вероятность другого.*

$$P(AB) = P(A) \cdot P\left(B/A\right) = P(B) \cdot P\left(A/B\right)$$

$$P(A_1A_2A_3...A_N) = P(A_1)\cdot P\left(A_2/A_1\right)\cdot P\left(A_3/A_1A_2\right)\cdot ...\cdot P\left(A_N/A_1A_2A_3...A_{N-1}\right)$$

Свойства независимых событий.

Если события А и В независимы, то независимы и каждая из пар: А и В, А и $\overline{B}$, $\overline{A}$ и В, $\overline{A}$ и $\overline{B}$.

Если события H_1, H_2, …H_n независимы, то заменяя любые из них на противоположные, вновь получаем независимые события.

Формула полной вероятности.

Вероятность события В, которое может произойти совместно только с одним из событий H_1, H_2, …H_n , образующих полную группу событий, вычисляется по формуле:

$$P(A) = \sum_{i=1}^{N} P(H_i)\cdot P\left(A/H_i\right)$$

События A_1, A_2, …A_n называют **гипотезами**.

Теорема гипотез (формула Байеса).

Если до опыта вероятности гипотез были $P(H_1)$, $P(H_2)$…$P(H_N)$, а в результате опыта произошло событие А, то условные вероятности гипотез находятся по формуле:

$$P\left(H_i/A\right) = \frac{P(H_i)P(A)}{\sum_{i=1}^{N} P(H_i)P\left(A/H_i\right)}$$

Пример. На трех технологических линиях изготавливаются микросхемы. Найти: 1) вероятность того, что случайно выбранное изделие оказывается бракованным; 2) вероятность того, что если изделие дефектно, то оно изготовлено на 1 линии.

№ линии	Количество изготавливаемых микросхем	Вероятность брака
1	25%	5%;
2	35%	4%
3	40%	2%

Рассмотрим события: H_1, H_2,…H_i,…,H_N (полная группа событий)– изделие изготавливается i линией; А{изделие с браком}.

$$P(A) = \sum_{i=1}^{3} P(H_i)P\left(A/H_i\right)$$

$$P(H_1) = 0.25 \qquad P(H_2) = 0.35 \qquad P(H_3) = 0.4$$

$$P\left(A/H_1\right) = 0.05 \qquad P\left(A/H_2\right) = 0.04 \qquad P\left(A/H_3\right) = 0.02$$

1) P(A)=0,25*0,05+0,35*0,04+0,4*002=0,0345=3,45%

2) $P\left(H_1/A\right) = \frac{0.25*0.05}{0.0345} = 0.36 = 36\%$

Схема последовательных испытаний Бернулли.

Проводится серия из n испытаний, в каждом из которых с вероятностью p может произойти событие А, с вероятностью q=1-p событие $\overline{A}$.

Вероятность наступления события А не зависит от числа испытаний n и результатов других испытаний.

Такая схема испытаний с двумя исходами (событие А наступило либо не наступило) называется **схемой последовательных испытаний Бернулли.**

Пусть при n испытаниях событие А наступило k раз, (n-k) раз событие $\overline{A}$.

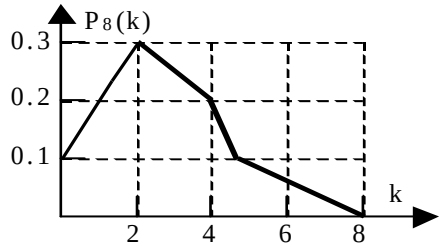
$\boxed{C_n^k = \frac{n!}{k!(n-k)!}}$ - число различных комбинаций события А

Вероятность каждой отдельной комбинации: $\boxed{p^k q^{k-1}}$

Вероятность того, что в серии из n испытаний событие А, вероятность которого равна p, появится k раз: $\boxed{P_n(k) = C_n^k p^k q^{n-k}}$

$\sum_{k=0}^{n} P_n(k) = 1$ - условие нормировки.

Пример. Вероятность изготовления нестандартной детали равна p=0,25, q=0.75. Построить многоугольник распределения вероятностей числа нестандартных деталей среди 8 изготовленных.



N=8 p=0.25 q=0.75

$P_8(k) = C_8^k \cdot 0.25^k \cdot 0.75^{8-k}$

Если **k_0 – наивероятнейшее число,** то оно находится в пределах:

$$np-q \le k_0 \le np+q$$

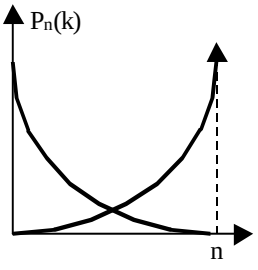
Если число (np+q) нецелое, то k_0 – единственное

Если число (np+q) целое, то существует 2 числа k_0.

$$\frac{P_n(k)}{P_n(k-1)} = \frac{p}{q} = \frac{n-k+1}{k} \begin{cases} > 1, k < np + p \text{ - ломанная линия на } [k-1, k] \text{ возрастает} \\ < 1, k > np + p \text{ - ломанная линия на } [k-1, k] \text{ убывает} \\ = 1, k = np + p \text{ ломанная линия постоянна} \end{cases}$$

<u>Предельные теоремы в схеме Бернулли.</u>

1. <u>Предельная теорема Пуассона</u>. При p≈0, n-велико, np= λ ≤ 10.



$$\boxed{P_n^k = \frac{\lambda^k}{k!} e^{-\lambda} + |R_n|, \quad R_n \le \frac{\lambda^2}{n}}$$

Формула дает распределение Пуасона, описывает редкие события.

2. <u>Предельная теорема Муавра-Лапласа.</u>

0 ≤ p ≤ 1, n –велико, np>10

$$\boxed{P_n(k) = \frac{1}{\sqrt{npq}} f\left(\frac{k-np}{\sqrt{npq}}\right)}$$

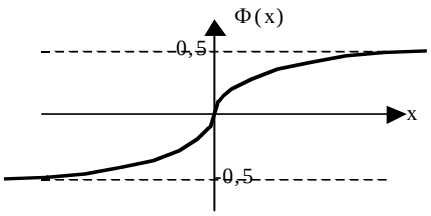
$f(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{m^2}{2}}$ - стандартное нормальное распределение

3. <u>Предельная интегральная теорема Муавра-Лапласа.</u>

В условиях предыдущей теоремы вероятность того, что событие А в серии из n испытаний наступит не менее k_1 раз и не более k_2 раз:

$$\boxed{P_n(k_1 \le k \le k_2) = \Phi(x_2) - \Phi(x_1)}$$

$$\Phi(x)=\frac{1}{2\pi}\int_0^x e^{-\frac{t^2}{2}}dt$$ - функция Лапласа



$$x_1=\frac{k_1-np}{\sqrt{npq}};x_2=\frac{k_2-np}{\sqrt{npq}}$$

Следствие:

$$P_n\left(|k-np|\le\varepsilon\right)=2\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)$$

$$\underbrace{np-\varepsilon}_{k_1}\le k\le\underbrace{np+\varepsilon}_{k_2}$$

$$x_1=\frac{-\varepsilon}{\sqrt{npq}};x_2=\frac{\varepsilon}{\sqrt{npq}}$$

Пример. ОТК проверяет на стандартность 1000 деталей. Выбранная деталь с вероятностью p=0,975 является стандартной.

1) Найти наивероятнейшее число стандартных деталей:

$K_0=np=975$

2) Найти вероятность того, что число стандартных деталей среди проверенных отличается от k_0 не более чем на 10.

$$P_n\left(|k-np|\le 10\right)=2\Phi\left(\frac{10}{\sqrt{npq}}\right)=2\Phi(2{,}026)=0{,}95=95\%$$

3) С вероятностью 0,95 найти максимальное отклонение числа стандартных деталей среди проверенных.

$$P_n\left(|k-np|\le\varepsilon\right)=0{,}95$$

$$2\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)=0{,}95 \qquad 2\Phi\left(\frac{\varepsilon}{4{,}937}\right)=0{,}475 \qquad \frac{\varepsilon}{4{,}937}=1{,}96\rightarrow\varepsilon=9{,}67$$

4) Найти число проверяемых деталей n, среди которых с вероятностью 0,9999 стандартные детали составят не менее 95%.

$0{,}95n\le k\le n$

$P(0{,}95n\le k\le n)=0.9999=\Phi(x_2)-\Phi(x_1)=2\Phi\left(\sqrt{n}\sqrt{\frac{0.025}{0.975}}\right)=0.9999$

$$x_2=\frac{n-np}{\sqrt{npq}}=\frac{\sqrt{n}(1-p)}{\sqrt{pq}}=\sqrt{n}\sqrt{\frac{0.025}{0.975}}$$

$$x_1=\frac{0{,}95n-np}{\sqrt{npq}}=-\sqrt{n}\sqrt{\frac{0.025}{0.975}}$$

$$\Phi\left(\sqrt{\frac{n}{39}}\right)=0{,}49995 \qquad \sqrt{\frac{n}{39}}=3.9 \quad n=3.9^2*39=594$$

при p=0,9999 n=594
при p=0,999 n=428
при p=0,99 n=260

Раздел 3. Случайные величины и распределение вероятностей.

Случайная – величина, которая в ходе опыта принимает то или иное значение из возможных своих значений, меняющееся от опыта к опыту и зависящее от множества непредсказуемых факторов.

Если случайные события характеризуют процесс качественно, то случайная величина – количественно.

Случайная величина – численная функция, задаваемая на множестве элементарных событий. На одном множестве может быть несколько случайных величин.

Дискретная случайная величина (ДСК) – величина, принимающая счетное (конечное или бесконечное) множество значений.

Непрерывная случайная величина (НСВ) – случайная величина, значения которой образуют несчетные множества. (Например, расход бензина на 100 км у автомобиля Жигули в Нижнем Новгороде).

Задать св – значит указать все множество ее значений и соответствующие этим значениям вероятности. Говорят, что задан закон распределения случайной величины.

Случайная величина может быть задана несколькими способами:

1. *Табличный.*

X	a_1	a_2	…	a_n
P	p_1	p_2	…	p_n

$$\sum p_i = 1$$

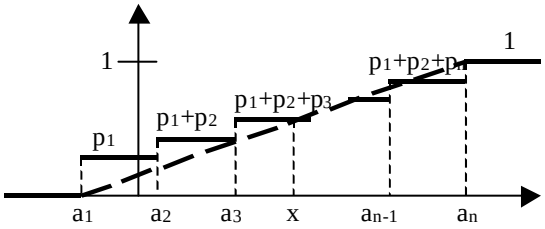
Значения случайных величин в таблице ранжируются, т.е. указываются в порядке возрастания.

Недостпаток табличного способа в том, что он пригоден только для случайных величин, принимающих небольшое количество значений.

2. *Функция распределения* ***F(x) = P(X<x)*** *или интегральный закон распределения.*

Указывается вероятность того, что случайная величина принимает значение < x.

X	a_1	a_2	a_3	…	a_{n-1}
P	p_1	p_2	p_3	…	p_{n-1}
F(x)	p_1	p_1+p_2	$p_1+p_2+p_3$	…	$p_1+p_2+\ldots+p_{n-1}$



При увеличении значения случайной величины, количество ступенек функции F(x) возрастает, уменьшается их высота и в пределе при $n \to \infty$ получаем гладкую непрерывную функцию F(x).

Свойства функции F(x).

1. Неотрицательна. $0 \le F(x) \le 1$
2. Неубывающая $F(x_2) > F(x_1)$ при $x_2 > x_1$
3. $\lim\limits_{x \to -\infty} F(x) = 0 \quad \lim\limits_{x \to +\infty} F(x) = 1$
4. **P(a<x<b) = F(a) – F(b)** Вероятность того, что значение x попадет в интервал (a,b) определяется разностью значений функции на концах интервала.

Наряду с F(x) вводится *f(x)* - функция плотности вероятности или дифференциальный закон распределения:

$$f(x) = \frac{P(x < X < x + \Delta x)}{\Delta x} = \frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{dF(x)}{dx}$$

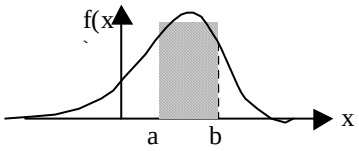
$$F(x) = \int_{-\infty}^{\infty} f(x)dx$$

Свойства функции *f(x):*

1. Неотрицательна. (т.к. F(x) неубывающая, $f(x) \ge 0$)

2. Площадь фигуры под кривой на интервале (a,b) равна:

$$\boldsymbol{P(a < x < b) = \int_a^b f(x)dx}$$



$$P(-\infty < x < \infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$ - условие нормировки функции *f(x).*

Основные дискретные и непрерывные случайные величины.

<u>Дискретные случайные величины (ДСВ).</u>

1. Биноминальная случайная величина x{0,1,2,3…n}

$P_n(m) = C_n^m p^m q^{n-m}$, p+q=1, 0<p<1

2. Пуассоновская случайная величина x{0,1,2,3…}

$$P_n(m) = \frac{\lambda^m}{m!} e^{-\lambda}, \lambda > 0$$

3. Бернуллиевая случайная величина $x\begin{Bmatrix} 1 & p \\ 0 & q \end{Bmatrix} p + q = 1$

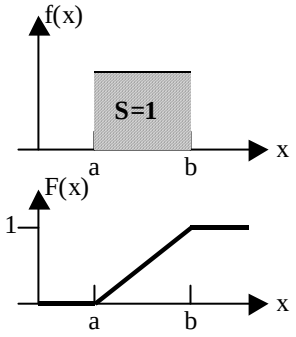
$P_n(k) = C_n^k p^k q^{n-k}$

4. Равномерное распределение $p\begin{Bmatrix} a_1, & a_2, & \dots & , a_n \\ p_1 = p_2 = \dots = p_n = \dfrac{1}{n} \end{Bmatrix}$

<u>Непрерывные случайные величины (НСВ).</u>

1. Равномерное распределение

$$f(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a,b] \\ 0, \quad x \notin [a,b] \end{cases}$$

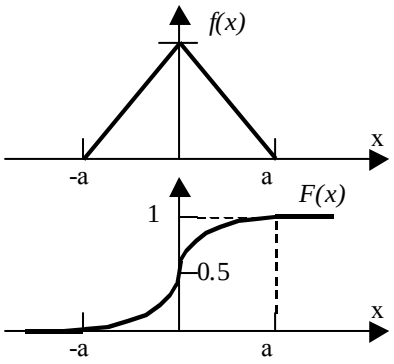


$$P(c < x < d) = \int_c^d f(x)dx = \frac{d-c}{b-a}$$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & x \in [a,b] \\ 1, & x > b \end{cases}$$

2. Треугольное распределение Симпсона

$$f(x) = \begin{cases} \dfrac{1}{a}\left(1 - \dfrac{|x|}{a}\right), & x \in (-a, a) \\ 0 & x \notin (-a, a) \end{cases}$$



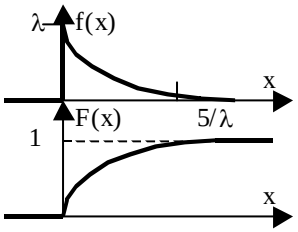
$$F(x)=\begin{cases}\frac{1}{2}\left(1+\frac{x}{a}\right)^2, & x\in(-a,0)\\ 1-\frac{1}{2}\left(1-\frac{x}{a}\right)^2, & x\in(0,a)\end{cases}$$

3. Экспоненциальное (показательное) распределение. Имеет важное значение в теории массового обслуживания и теории надежности.

$$f(x)=\begin{cases}0, & x\le 0\\ \lambda e^{-\lambda x}, & x>0\end{cases}$$

$$F(x)=\begin{cases}0, & x\le 0\\ 1-e^{-\lambda x}, & x>0\end{cases}$$

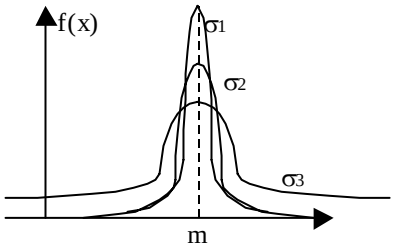
λ - интенсивность.



3. Нормальный закон распределения.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}},\ \sigma>0$$

σ=1, m=0 – нормальное стандартное распределение (m-мат. ожидание)



$t=\frac{x-m}{\sigma}$ - такой подстановкой любое нормальное распределение приводится к стандартному.

При фиксированном σ и изменяющемся m, кривая двигается вдоль Ox, не изменяя формы.

При фиксированном m и изменяющемся σ ($\sigma_1<\sigma_2<\sigma_3$), кривая вытягивается вдоль оси ординат, но площадь фигуры под каждой кривой = 1.

$$P(a<x<b)=\int_a^b f(x)dx=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)$$

Функция Лапласа: $\boxed{F(x)=\frac{1}{2}+\Phi\left(\frac{x-m}{\sigma}\right)}$

Операции со случайными величинами

Со случайными величинами, рассмотренными на одном и том же интервале исходов опыта, можно обращаться как с обычными числами и функциями.

X:

	a_1	a_2	…	a_n
p	p_1	p_2	…	p_n

Y=φ(x)

Нужно найти закон распределения СВ Y. $y_k=\varphi(a_k)$, где k=1,2,…,n.

$$P(y=y_k)=P(x=a_k)=P_k$$

Если все значения СВ Y различны, то их надо проранжировать и указать соответствующие вероятности.

Если СВ Y принимает совпадающие значения, то их надо объединить под общей вероятностью, равной сумме соответствующих вероятностей, а после в ранжированном виде привести в таблице.

$X=\{0,1,2,\ldots,9\}$, $P(x=k)=0.1$, $k=0,1,\ldots,9$, $Y=x^2$, $Z=(x-5)^2$.

X	0	1	2	3	4	5	6	7	8	9
P	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Y	0	1	4	9	16	25	36	49	64	81
$\mathbf{P_y}$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Z	25	16	9	4	1	0	1	4	9	16
$\mathbf{P_z}$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Закон распределения СВ Z:

	0	1	4	9	16	25
$\mathbf{P_z}$	0.1	0.2	0.2	0.2	0.2	0.1

Бинарные операции (с несколькими величинами)

СВ X,Y заданы в 1 опыте.

Исход опыта	E_1	E_2	…	E_n
Вероятность исхода	P_1	P_2	…	P_n
X	X_1	X_2	…	X_n
Y	Y_1	Y_2	…	Y_n
Z=φ(XY)	Z_1	Z_2	…	Z_n

Сложнее, если СВ задана только своим распределением:

	a_1	a_2	…	a_n
P	p_1	p_2	…	p_n

	b_1	b_2	…	b_n
P	g_1	g_2	…	G_n

Z=X+Y

СВ Z принимает значения a_k+b_s, где $a_k=a_1,a_2,\ldots,a_n$; $b_s=b_1,b_2,\ldots b_m$.

Общее количество возможных значений СВ = m·n.

$P(Z=a_k+b_s)=P(X=a_k, Y=b_s)$

Для нахождения такой вероятности необходимо знать закон совместного распределения СВ X и Y.

Набор точек (a_k,b_s) вместе с вероятностями $P(X=a_k, Y=b_s)$ называется ***совместным распределением СВ X и Y***. Обычно такое распределение задается таблицей.

Определение закона распределения суммы СВ по законам распределения слагаемых называется ***композицией законов распределения***.

X \Y	b_1	b_{12}	…	b_s	…	b_m	P_x
a_1	P_{11}	P_{12}	…	P_{1s}	…	P_{1m}	P_1
a_2	P_{21}	P_{22}	…	P_{2s}	…	P_{2m}	P_2
…	…	…	…	…	…	…	…
a_k	P_{k1}	…	…	P_{ks}	…	P_{km}	P_k
…	…	…	…	…	…	…	…
a_n	P_{n1}	P_{n2}	…	P_{ns}	…	P_{nm}	P_n
P_y	g_1	g_2	…	g_s	…	g_m	1

$$\sum_{k,s} P_{k,s} = 1$$

Наиболее просто вероятности P_{ks} находятся в случае независимости СВ X и Y. Две СВ X и Y называются ***независимыми*** тогда и только тогда, когда

$$P(X=a_k, Y=b_s)=P(X=a_k)\cdot P(Y=b_s)$$

$$P_{ks}=P_k\cdot P_s$$

По известному закону распределения совместного распределения СВ X и Y могут быть найдены одномерные законы распределения СВ X и Y.

$$P(Y=b_1)=g_1=\sum_{k=1}^{n}P_{k1}\;\; P(X=a_1)=P_1=\sum_{s=1}^{m}P_{1s}$$

$$P(X=a_k)=P_k=\sum_{s=1}^{m}P_{ks}$$

Теорема . Если СВ X,Y являются независимыми, то любые функции φ(X) и ψ(У) от этих величин также являются независимыми.

Распределение функции от случайной величины

X – непрерывная СВ $F_X(x),\ f_x(x)$

$Y=\varphi(x)$. По закону распределения СВ X. Найти закон распределения СВ Y.

Если СВ $X\in[x_0,x_1]$, то $Y=\varphi(x)\in[y_0,y_1]$.

Предполагается, что функция φ(x) является однозначной и имеет обратную функцию q(y).

$$P\left[x<X<x+dx\right]=P\left(y<Y<y+dy\right)$$

Воспользовавшись элементами вероятности:

$$f_x(x)dx=f_y(y)dy \qquad f_y(y)=f_x(x)\frac{dx}{dy}$$

получим $f_y(y)=f_x(q(y))\dfrac{dq(y)}{dy}$.

Закон распределения не меняется, если q(y) является линейной.

$$f_y(y)=f_x(x).$$

Многомерные законы распределения СВ

Часто при решении практических задач мы имеем дело не с одной, а с совокупностью нескольких случайных величин, которые взаимосвязаны.

n $x_1,x_2,\ldots,x_n$ ***n-мерная случайная величина*** – совокупность n взаимосвязанных случайных величин. Для ее описания используются многомерные законы распределения.

Двумерные функции распределения

X,Y F(x,y)=P(X<x,Y<Y)

Функция *F(x,y)* обладает свойствами, аналогичными свойствам одномерной функции:

– не убывающая
– не отрицательная
$0\le F(x,y)\le 1$

1. $x_2\ge x_1 \Rightarrow F(x_2,y)\ge F(x_1,y)$
   $y_2\ge y_1 \Rightarrow F(x,y_2)\ge F(x,y_1)$
2. $F(\infty,\infty)=1 \quad F(-\infty,-\infty)=0$
3. $F_x(x)=P(X<x+=P(X<x,Y<\infty)=F(x,\infty)$
   $F_y(y)=P(Y<y)=P(X<\infty,Y<y)=F(\infty,y)$

f(x,y) – функция плотности вероятности совместного распределения величин *x* и *y*.

$$f(x,y)=\frac{d^2F(x,y)}{dxdy}=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\frac{P(x<X<x+\Delta x,y<Y<y+\Delta y)}{\Delta x,\Delta y}$$

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(x,y)dxdy$$

1. $f(x,y)\ge 0$
2. $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x,y)dxdy=1$ – условие нормировки
3. По известным двумерным находятся соответствующие одномерные

$$f_x(x)=\int_{-\infty}^{+\infty} f(x,y)dy \qquad P(\alpha_x<X<\beta_x,\alpha_y<Y<\beta_y)=\int_{\alpha_x}^{\beta_x}\int_{\alpha_y}^{\beta_y} f(x,y)dxdy$$

$$f_y(y)=\int_{-\infty}^{+\infty} f(x,y)dx$$

В случае статистической независимости СВ X и У

$F(x,y)=F_x(x)\cdot F_y(y)$

$f(x,y)=f_x(x)\cdot f_y(y)$

$F(x,y)=F_x(x)\cdot F_y(y/x)=F_x(x/y)$ – для условных

$f(x,y)=f_x(x)\cdot f(y/x)=f_y(y)\cdot f(x/y)$

Раздел 4. Числовые характеристики СВ

Исчерпывающие представления о СВ дает закон её распределения.

Во многих задачах, особенно на заключительной стадии, возникает необходимость получить о величине некоторое суммарное представление: центры группирования СВ – среднее значение или математическое ожидание, разброс СВ относительно её центра группирования.

Эти числовые характеристики в сжатой форме отражают существенные особенности изучаемого распределения.

Математическое ожидание (МО)

M(x), MO(x), m_x, m

$$M(x)=\begin{cases}\sum_{i=1}^{n} x_i p_i & \sum_{i=1}^{n} p_i=1 \quad \text{ДСВ}\\ \int_{-\infty}^{\infty} xf(x)dx & \text{НСВ}\end{cases}$$

Основные свойства МО:

1. M(x) СВ X $\Rightarrow$ $X_{min}\le M(x)\le X_{max}$
2. M(C)=C МО постоянной величины есть величина постоянная
3. $M(X\pm У)=M(X)\pm M(У)$
4. $M(X\cdot У)=M(x)\cdot M(y)$ $\Rightarrow$ $M(Cx)=CM(x)$ – МО произведения двух независимых СВ
5. $M(aX+вУ)=aM(X)+вM(У)$
6. $M(X-m)=0$ – МО СВ X от её МО.

МО основных СВ

Дискретные Случайные Величины

1. Биноминальные СВ MO(X)=np
2. Пуассоновские СВ MO(X)=λ
3. Бернуллиевы СВ MO(X)=p
4. Равномерно распред. СВ $MO(X)=\frac{a_1+a_2+...+a_n}{n}$

Непрерывные Случайные Величины

1. Равномерно распределенная СВ $MO(X)=\frac{b+a}{2}$
2. Нормально распределенная СВ MO(X)=m

3. Экспоненциально распределенная СВ $MO(X)=\frac{1}{\lambda}$

Дисперсия СВ

1. $R=X_{max}-X_{min}$ – размах СВ
2. $M(|X-m|)$ – среднее абсолютное отклонение СВ от центра группирования
3. $M(X-m)^2$ – дисперсия – МО квадрата отклонения СВ от центра группирования

$M(X-m)^2=D(X)=\sigma^2=\sigma_x{}^2=\sigma^2(X)$

$\sqrt{D(X)}=\sigma=CKO$ – среднеквадратическое отклонение (стандартное отклонение).

$$D(X)=\begin{cases}\sum_{i=1}^{n}(x_i-m)^2\,p_i \qquad \sum p_i=1 & \text{ДСВ}\\ \int_{-\infty}^{\infty}(x-m)^2\,f(x)dx & \text{НСВ}\end{cases}$$

Основные свойства дисперсии:

1. Для любой СВ X: $D(X)\geq 0$. При X=const $D(X)\equiv 0$.
2. $D(X)=M(X^2)-M^2(X)=M(X^2-2mX-m^2)$
3. $D(cX)=c^2D(X)$
4. $D(X+c)=D(X)$
5. $D(X+Y)=D(X)+D(Y)$, $D(X-Y)=D(X)+D(Y)$

В общем случае:

$D(X+Y)=M(X+Y-m_{x+y})^2=M((X-m_x)+(Y-m_y))^2=M((X=m_x)^2+2(X-m_x)(Y-m_y)+(Y-m_y)^2)=$

$=D(X)+\boxed{2M((X-m_x)(Y-m_y))}+D(Y)$. Второй член этого выражения называется ***<u>корреляционным моментом</u>***. $m_{x+y}=M(X)+M(Y)=m_x+m_y$. $D(X)=M(X-m_x)^2$.

$M((X-m_x)(Y-m_y))=K(X,Y)=K_{xy}=cov(x,y)$ – ковариация

$K_{xy}/\sigma_x\sigma_y=\rho_{xy}$ – коэффициент корреляции

6. Независимые СВ: $D(XY)=D(X)D(Y)+M^2(X)D(Y)+M^2(Y)D(X)$

<u>Дисперсия основных СВ</u>

ДСВ

1. Биноминальные	$D(X)=npq$
2. Пуассоновские	$D(X)=\lambda$
3. Бернуллиевы	$D(X)=pq$

НСВ

1. Равномерно распределенные	$D(X)=(b-a)^2/12$
2. Нормально распределенные	$D(X)=\sigma^2$
3. Экспоненциально распределенные	$D(X)=1/\lambda^2$

Математическое ожидание и дисперсия суммы случайных величин

$X_1,X_2,\ldots,X_n$ – независимые СВ с одинаковым законом распределения.

$M(Xk)=a$ $D(Xk)=\sigma^2$

$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_n$ – среднее арифметическое

$$M(\overline{X})=a \qquad D(\overline{X})=\frac{\sigma^2}{n} \qquad CKO=\sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}$$

<u>Другие числовые характеристики СВ</u>

Моменты распределения делятся на начальные моменты, центральные и смешанные.

1. Начальные моменты q^{ro} порядка (q=1,2,…): $M(X^1)=MO$

$$M(X^q)=\begin{cases}\sum_{i=1}^{n} x_i^q p_i & \text{ДСВ}\\ \int_{-\infty}^{\infty} x^q f(x)dx & \text{НСВ}\end{cases}$$

2. Центральные моменты $q^{го}$ порядка: $M((X-m)^2)=D$

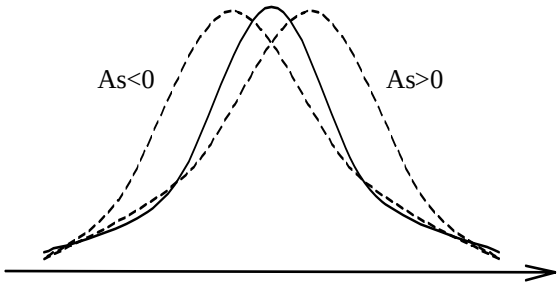
$$M((X-m)^q)=\begin{cases}\sum_{i=1}^{n} (x_i-m)^q p_i & \text{ДСВ}\\ \int_{-\infty}^{\infty} (x-m)^q f(x)dx & \text{НСВ}\end{cases}$$

$M(x-m)^q=M(x)^q-C_q^1 mM(x)^{q-1}+ C_q^2 mM(x)^{q-2}+\ldots+(-1)^q m^q$

$M(x-m)^3= M(x)^3-3mM(x)^2+2m^3$

$M(x-m)^2= M(x)^2-m^2=D(x)$

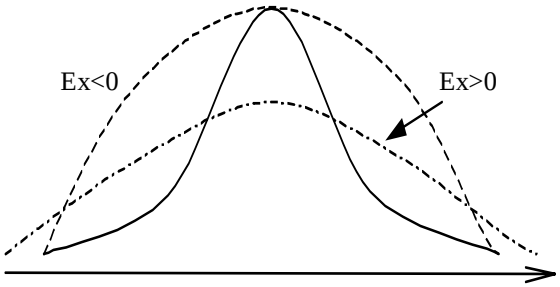
Центральные моменты $3^{го}$ и $4^{го}$ порядков используются для получения коэффициентов асимметрии и эксцесса (As, Ex), характеризующих особенности конкретного распределения.



$$As=\frac{M(x-m)^3}{\sigma^3}=M\left(\frac{x-m}{\sigma}\right)^3$$

Для нормального закона распределения As=0.

Если As>0, то распределение имеет ***правостороннюю скошенность***. При As<0 – ***левосторонняя скошенность***.



$$Ex=\frac{M(x-m)^4}{\sigma^4}-3=M\left(\frac{x-m}{\sigma}\right)^4-3$$

Эксцесс характеризует остро- или плосковершинность исследуемого распределения по сравнению с нормальным распределением.

НСВ:

1. Нормальное распределение: Ex=As=0
2. Равномерное распределение: As=0, Ex=-1,2
3. Экспоненциальное распределение: As=2, Ex=9.

Биноминальное: $As=\frac{q-p}{\sqrt{npq}}$ $\quad Ex=\frac{1-6pq}{npq}$

3. Смешанные моменты:

Начальный смешанный момент порядка (k+s) системы 2^x СВ (X+Y):

$$M(X^kY^s)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^s f(x,y)dxdy$$

Центральный моменты порядка (k+s):

$$M((X-m_x)^k(Y-m_y)^s)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-m_x)^k(y-m_y)^s f(x,y)dxdy$$

Центральный смешанный момент второго порядка:

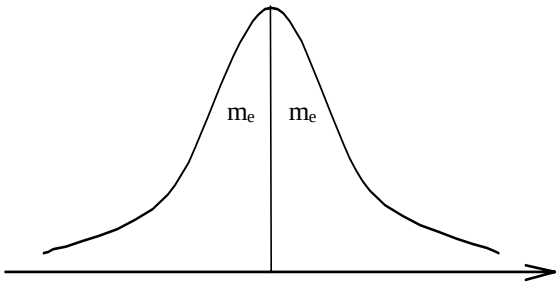
$K_{xy}=M((X-m_x)(Y-m_y))$ – корреляционный момент

$\frac{K_{XY}}{\sqrt{D(x)D(y)}}=\rho_{x,y}$ – коэффициент корреляции

Мода ДСВ – значение СВ, имеющее максимальную вероятность.

Мода НСВ – значение СВ, соответствующее максимуму функции плотности вероятности f(x).

Обозначение моды: m_0, $M_0(x)$, mod(x).



Медиана СВ X (m_e, $M_e(x)$, med(x)) – значение СВ, для которого выполняется равенство:

$$P(X<m_e)=P(X>m_e)$$

$$F(m_e)=0,5.$$

Медиана – это площадь, получаемая делением фигуры пополам.

В симметричном распределении $m=m_0=m_e$. В несимметричном они не равны.

Так как мода и медиана зависят от структуры распределения, их называют ***структурными средними***.

Медиана – это значение признака, который делит ранжированный ряд значений СВ на две равных по объему группы. В свою очередь, внутри каждой группы могут быть найдены те значения признака, которые делят группы на 4 равные части – ***квартиль***.

Ранжированный ряд значений СВ может быть поделен на 10 равных частей – децилей, на 100 – центилей.

Такие величины, делящие ранжированный ряд значений СВ на несколько равных частей, называются ***квантилями***.

Под p% квантилями понимаются такие значения признака в ранжированном ряду, которые не больше p% наблюдений.

Предельные теоремы теории вероятностей

Делятся на две группы: Закон Больших Чисел (ЗБЧ) и Центральная Предельная Теорема (ЦПТ).

Закон Больших Чисел устанавливает связь между абстрактными моделями теории вероятностей и основными ее понятиями и средними значениями, полученными при статистической обработке выборки ограниченного объема из генеральной совокупности. P, F(x), M(x), D(x).

ЗБЧ доказывает, что средние выборочные значения при $n\to\infty$ стремятся к соответствующим значениям генеральной совокупности: $h_n(A)\to P$, $X_{ср}\to M(X)$, $\sigma_{ср}^2\to D(X)$, $F^*(X)\to F(X)$.

Лемма Маркова. Если Y – СВ, принимающая не отрицательные значения, то для любого положительного ε:

$$P(Y\geq\varepsilon)\leq M(x)/\varepsilon,\quad P(Y<\varepsilon)\geq 1-M(x)/\varepsilon.$$

Доказательство. Рассмотрим Y и $Y_\varepsilon=\begin{cases}0, Y<\varepsilon\\ \varepsilon, Y\geq\varepsilon\end{cases}$: $Y_\varepsilon\leq Y$, $M(Y_\varepsilon)\leq M(Y)$

$$M(Y_\varepsilon)=0\cdot P(Y<\varepsilon)+\varepsilon\cdot P(Y\geq\varepsilon)=\varepsilon\cdot P(Y\geq\varepsilon)$$

$$M(Y)\geq M(Y_\varepsilon)=\varepsilon\cdot P(Y\geq\varepsilon).$$

Лемма позволяет сделать оценку вероятности наступления события по математическому ожиданию этой СВ.

Неравенство Чебышева. Для любой СВ с ограниченными первыми двумя моментами (есть MO и D) и для любого ε>0:

$$P(|X-m|\geq\varepsilon)\leq\frac{D(x)}{\varepsilon^2};\qquad P(|X-m|<\varepsilon)\geq 1-\frac{D(x)}{\varepsilon^2}$$

Доказательство. По лемме Маркова: рассмотрим не отрицательную СВ Y

$$Y=(X-m)^2\qquad M(Y)=M(X-m)^2=D(x)$$

$$P(|X-m|\geq\varepsilon)=P((X-m)^2\geq\varepsilon^2)=P(Y\geq\varepsilon^2)\leq M(Y)/\varepsilon^2=D(x)/\varepsilon^2.$$

Требуется только знание дисперсии СВ при любом законе распределения.

ЗБЧ в форме Чебышева. $X_1, X_2, \ldots, X_n$ – последовательность независимых СВ. Для любого ε>0 и $n\to\infty$:

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|>\varepsilon\right)<\delta$$

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-m\right|>\varepsilon\right)<\delta \qquad P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-m\right|<\varepsilon\right)<1-\delta$$

ЗБЧ в форме Бернулли. m – число успехов в серии из n последовательных испытаний Бернулли. P – вероятность успеха в каждом отдельном испытании. ε>0:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$$

ЗБЧ носит чисто качественный характер. В тех же условиях неравенство Чебышева позволяет получить количественную характеристику оценки вероятности.

Пример. Для определения вероятности события проведено 40000 опытов. События наблюдалось в m=16042 случаях. За вероятность события принимается относительная частота наступления события: m/n≈0,4. Применяя неравенство Чебышева, оценить, с какой вероятностью можно гарантировать, что число 0,4, принятое за вероятность, отличается от истинной вероятности не больше, чем на 0,05.

$$P\left(\left|\frac{m}{n}-p\right|\le 0{,}05\right)\ge 1-\frac{pq}{n\varepsilon^2}=0{,}9975$$

Неизвестные p и q находим из системы уравнений:

$$\begin{cases}(p+q)^2=1\\(p-q)^2\ge 0\end{cases} \Rightarrow\ pq\le\frac{1}{4}$$

Центральная предельная теорема Ляпунова.

Предмет внимания этой теоремы – распределение суммы большого числа СВ.

$$X=(x_1+x_2+\ldots+x_n)/n$$

Распределение суммы n независимых СВ в независимости от их законов распределения асимптотически сходятся к нормальному закону при неограниченном числе слагаемых и ограниченных двух первых моментах (МО и D).

$$P(x<b)=\frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{b}\exp\left(\frac{(x-m)^2}{2\sigma_x^2}\right)\cdot dx$$

Если $\sigma_i^2=\sigma^2$, то $\sigma_x^2=\sigma^2/n$, $\sigma_x=\frac{\sigma}{\sqrt{n}}$.

$$D(x)=\sigma_x^2=(\sigma_1^2+\sigma_2^2+\ldots\sigma_n^2)/n^2$$

ЦПТ универсальны и справедливы как для НСВ, так и для ДСВ.

$$P(a<X<b)=\Phi(t_2)-\Phi(t_1).$$

$$t_2=(b-m_x)/\sigma_x \qquad t_2=(a-m_x)/\sigma_x$$

$$S_n=(X_1+X_2+\ldots+X_n)/n$$

$$P(|S_n-m|<z\sigma)=2\Phi(z)$$

$$M(xk)=m \qquad D(xk)=\sigma^2$$

$$D(\bar{x})=\sigma^2/n \qquad \sigma_{\bar{x}}=\sigma/\sqrt{n}$$

$$P\left(|\bar{x}-m|<z\frac{\sigma}{\sqrt{n}}\right)=2\Phi(z)$$

$$P\left(\left|\frac{\bar{x}-m}{\sigma/\sqrt{n}}\right|\le z\right)=2\Phi(z)$$

ЦПТ в интегральной форме Муавра-Лапласа.

$$P\left(a \le \frac{\bar{x} - np}{\sqrt{npq}} \le b\right) = \Phi(b) - \Phi(a)$$

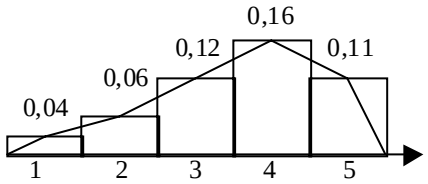
$$P(k_1 \le k \le k_2) = \Phi\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1 - np}{\sqrt{npq}}\right)$$

Статистическое оценивание параметров распределения

Мы анализируем только выборки из генеральной совокупности. По средне выборочным параметрам находим параметры самой генеральной совокупности.

Задачи такого рода решаются методами проверки статистических гипотез и статистической оценки параметров распределения.

Прежде нужно получить и провести первичную обработку исходных экспериментальных данных.



Статистические ряды часто изображают графически в виде полигона, гистограммы, кумулятивной кривой $F^*(x)$.

Полигон – ломаная линия, соединяющая в декартовой системе координат точки (x_i,n_i), (x_i,m_{xi}).

Кумулятивная кривая строится по точкам $(x_i,F^*(x_i))$.

Гистограмма – на оси абсцисс – отрезки интервалов t, на этих интервалах строятся прямоугольники с высотой, равной относительной частоте признака. По гистограмме легко строится полигон.

И полигон, и гистограмма характеризуют функцию $f^*(x)$ – плотность вероятности.

НСВ – проблема выбора интервала варьирования h.

h выбирается, исходя из необходимости выявления характерных черт рассматриваемого распределения.

Правило ***Старджесса***: $h = \dfrac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n}$

Как только характерные особенности распределения проявились, ставится вопрос об условиях, при которых сформировалось данное распределение – вопрос об однородности статистических данных.

Если функция $f^*(x)$ – бимодальная (имеет два максимума), то статистическое данные неоднородные.

Методы математической статистики должны позволить сделать обоснованные выводы о числовых параметрах и законе распределения генеральной совокупности по ограниченному числу выборок из этой совокупности.

Состав выборок случаен и выводы могут быть ложными. С увеличением объема выборки вероятность правильных выводов растет. Всякому решению, принимаемому при статистической оценке параметров, ставится в соответствие некоторая вероятность, характеризующая степень достоверности принимаемого решения.

Задачи оценки параметров распределения ставятся следующим образом:

Есть СВ X, характеризуемая функцией $F(X, \theta)$.

θ – параметр, подлежащий оценке.

Делаем m независимых выборок объемом n элементов x_{ij} (i – номер выборки, j – номер элемента в выборке).

1	$x_{11}, x_{12}, \ldots, x_{1n}$	X_1
2	$x_{21}, x_{22}, \ldots, x_{2n}$	X_2
…		
m	$x_{m1}, x_{m2}, \ldots, x_{mn}$	X_m

Случайные величины $X_1, X_2, \ldots X_m$ мы рассматриваем как m независимых СВ, каждая из которых распределена по закону $F(X, \theta)$.

Всякую однозначную функцию наблюдений над СВ x, с помощью которой судят о значении параметра θ, называют $\tilde{\Theta}_n$ ***– оценкой параметра θ***.

$$\tilde{\Theta}_n = \varphi(x_1, x_2, ..., x_m)$$

Выбор оценки, позволяющей получить хорошее приближение к оцениваемому параметру – задача исследования.

Основные свойства оценок

Несмещенность, эффективность и состоятельность.

Оценка $\tilde{\Theta}_n$ параметра θ называется ***несмещенной***, если $M(\tilde{\Theta}_n)=\theta$.

Если $\begin{cases} M(\tilde{\Theta}_n) > \Theta \\ M(\tilde{\Theta}_n) < \Theta \end{cases}$ – в оценке параметра θ имеется систематическая ошибка.

Несмещенность оценки гарантирует отсутствие систематической ошибки в оценке параметра.

Несмещенных оценок может быть несколько.

$\tilde{T}_n$ – несмещенная оценка θ.

Разброс параметров или рассеяние величины относительно математического ожидания θ характеризует дисперсия $D(\tilde{\Theta}_n)$, $D(\tilde{T}_n)$.

Из двух или более несмещенных оценок предпочтение отдается оценке, обладающей меньшим рассеянием относительно оцениваемого параметра.

Оценка $\tilde{\Theta}_n$ называется ***состоятельной***, если она подчиняется закону больших чисел:

$$\lim_{n\to\infty} P(|\tilde{\Theta}_n - \Theta| < \varepsilon) = 1$$

На практике не всегда удается удовлетворить одновременно всем трем требованиям.

Оценка математического ожидания по выборке

Теорема 1. Среднее арифметическое $\overline{X}$ по n независимым наблюдениям над СВ x с МО m является несмещенной оценкой этого параметра.

Доказательство: $x_1, x_2, \ldots, x_n$ $M(x)=m$ $M(x_1)=M(x_2)=\ldots=M(x_n)=m$

$$\overline{X} = \sum_{i=1}^{n} x_i / n \qquad \begin{matrix} M(\overline{X}) = m \\ M(\tilde{\Theta}_n) = \Theta \end{matrix}$$

Теорема 2. Среднее арифметическое $\overline{X}$ по n независимым наблюдениям над СВ x с МО m и дисперсией $D(x)=\sigma^2$ является состоятельной оценкой МО.

Доказательство: $D(x)=\sigma^2$ $D(x_1)=D(x_2)=\ldots=D(x_n)=\sigma^2$

$$D(\overline{X}) = \sigma^2 / n$$

$$P(|\overline{X} - m| \geq \varepsilon) \leq \frac{D(\overline{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\lim_{n\to\infty} P(|\overline{X} - m| > \varepsilon) = 0$$

$$\lim_{n\to\infty} P(|\overline{X} - m| \leq \varepsilon) = 1$$

Теорема 3. Если СВ X распределена по нормальному закону с параметрами (m,σ^2), то несмещенная и состоятельная оценка $\overline{X}$ МО m имеет минимальную дисперсию σ^2/n => $\overline{X}$ является и эффективной.

Оценки дисперсии по выборке

Если случайная выборка состоит из n независимых наблюдений над СВ X с $M(X)=m$ и $D(X)=\sigma^2$, то выборочная дисперсия не является несмещенной оценкой дисперсии генеральной совокупности.

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

$$M(\sigma^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Несмещенной оценкой D(x) является $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $M(\tilde{S}^2) = \sigma^2$.

Легко доказать по формуле Чебышева, что оценки S2 и $\tilde{S}^2$ являются состоятельными оценками дисперсии.

Несмещенная, состоятельная и эффективная оценка дисперсии:

$$S_*{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Если МО генеральной совокупности неизвестно, то используют $\tilde{S}^2$.

Существуют регулярные методы получения оценок параметров генеральной совокупности по данным выборок.

Методы оценки параметров генеральной совокупности

<u>Метод наибольшего (максимального) правдоподобия (МНП)(ММП)</u> обладает следующими достоинствами:

1. Всегда приводит к состоятельным оценкам (иногда смещенным)
2. Получаемые оценки распределены асимптотически нормально и имеют минимально возможную дисперсию по сравнению с другими асимптотически нормальными оценками.

Недостаток: требуется решать громоздкие системы уравнений.

Имеется СВ X, $f(x,\theta)$ – функция ее плотности вероятности, выражение которой известно.

θ – неизвестный параметр, подлежащий оценке.

$x_1, x_2,\ldots,x_n$ – n независимых наблюдений над СВ x.

В основе МНП лежит функция $L(\theta)$ – функция правдоподобия, формирующаяся с учетом свойств многомерной функции распределения наблюдений над СВ x.

$$f(x_1, x_2,\ldots,x_n,\theta)=f(x_1, \theta)\cdot f(x_2,\theta)\cdot\ldots\cdot f(x_n,\theta)$$

В указанное равенство подставляются данные и получаем функцию $L(\theta)$:

$$L(\theta)=f(x_1, \theta)\cdot f(x_2,\theta)\cdot\ldots\cdot f(x_n,\theta)$$

За максимальное правдоподобное значение параметра θ принимаем $\tilde{\Theta}_n$, при которой $L(\theta)$ максимально.

$$L'(\theta)=0 \Rightarrow \theta_{max}=\tilde{\Theta}_n$$

<u>Метод моментов(Метод Пирсона).</u>

Метод обладает следующими достоинствами:

1. Оценки получаемые этим методом всегда являются состоятельными.
2. Метод моментов мало зависит от закона распределения случайной величины.
3. Сложность вычисления незначительна.

Известна случайная величина X, которая характеризуется $f(x, \theta_1, \theta_2\ldots\theta_q)$, аналитический вид этой функции известен.

По выборке объёмом n $x_1,x_2,x_3,\ldots x_n$ – значения случайной величины в выборке вычисляем эмпирические начальные моменты случайной величины:

$$\tilde{M}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \tilde{M}_1(x_1, x_2, ...x_n)$$

$$\tilde{M}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i = \tilde{M}_2(x_1, x_2, ...x_n)$$

$$\tilde{M}_q = \frac{1}{n}\sum_{i=1}^{n} x_i = \tilde{M}_q(x_1, x_2, ...x_n)$$

Находим теоретические моменты:

$$M_1 = \int_{-\infty}^{\infty} x f(x, \theta_1, \theta_2, ..., \theta_q) dx = M_1(\theta_1, \theta_2, ..., \theta_q)$$

$$M_2 = \int_{-\infty}^{\infty} x^2 f(x, \theta_1, \theta_2, ..., \theta_q) dx = M_2(\theta_1, \theta_2, ..., \theta_q)$$

$$M_q = \int_{-\infty}^{\infty} x^q f(x, \theta_1, \theta_2, ..., \theta_q) dx = M_q(\theta_1, \theta_2, ..., \theta_q)$$

Основная идея метода моментов заключается в приравнивании значения эмпирических значений моментов теоретическим.

$$\tilde{M}_1(x_1, x_2, ..., x_n) = M_1(\theta_1, \theta_2, ..., \theta_q)$$

$$\tilde{M}_2(x_1, x_2, ..., x_n) = M_2(\theta_1, \theta_2, ..., \theta_q)$$

$$\tilde{M}_q(x_1, x_2, ..., x_n) = M_q(\theta_1, \theta_2, ..., \theta_q)$$

Решим систему q-уравнений с q-неизвестными:

$$\tilde{\theta}_1 = \tilde{\theta}_1(x_1, x_2, ..., x_n)$$

$$\tilde{\theta}_2 = \tilde{\theta}_2(x_1, x_2, ..., x_n)$$ состоятельные оценки.

$$\tilde{\theta}_q = \tilde{\theta}_q(x_1, x_2, ..., x_n)$$

Состоятельность этих оценок основана на том, что эмпирические моменты при достаточно большом n ($n \to \infty$) стремится к теоретическим. Выполняется закон больших чисел.

$$\underset{n\to\infty}{P}\left\{\left|\tilde{M}_q - M_q\right| > \varepsilon\right\} = 0$$

Распределение средней арифметической для выборки из нормальной совокупности. Распределение Стьюдента.

Выборочное среднее рассчитанное по конкретной выборке, есть конкретное число. Состав выборки случаен и среднее арифметическое вычисленное по элементам другой выборки того же объёма, будет число отличное от первого.

$\overline{X}$ - средняя арифметическая величина меняющаяся от выборки к выборке.

Теорема: Если случайная величина X подчиняется нормальному закону с параметрами m и σ^2 $X(m, \sigma^2)$, а $x_1, x_2, x_3, \ldots, x_n$ – это выборка из генеральной совокупности, то средняя арифметическая:

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

так же является случайной величиной подчиняющаяся нормальному закону с параметрами m и σ^2/n, а нормированная случайная величина:

$$t = \frac{\overline{X} - m}{\sigma / \sqrt{n}}$$

так же подчиняется нормальному закону с параметрами (0;1).

Предполагается при использовании таблиц интеграла вероятности, что объём выборки n достаточно велик($n \geq 30$).
Существует достаточно большое количество технических задач в которых не удаётся собрать выборку такого объёма. Тем не менее анализу такой выборки необходимо дать вероятностную оценку.

В 1908 году английский математик Вильям Госсет дал решение задачи малых выборок (псевдоним Стьюдент). Стьюдент показал, что в условиях малых выборок надо рассматривать не распределение самих средних, а их нормированных отклонений от средних генеральных.

Надо рассматривать:

$$t = \frac{\overline{X} - m}{\sigma / \sqrt{n}}$$

$S(t,n) = S(-t,n) = B_n\left(1 + \frac{t^2}{n-1}\right)^{-n/2}$ - это чётное распределение.

Оно зависит только от объёма выборки n и не зависит ни от математического ожидания, ни от дисперсии случайной величины X. При $n \to \infty$ t – распределение Стьюдента переходит в нормальное распределение.

Поскольку в большинстве случаев σ генеральной совокупности неизвестно, то работает с такой величиной:

- состоятельная и несмещённая оценка.

Существуют t таблицы распределения Стьюдента.

Величина доверительной вероятности, её выбор находятся за пределами прикладной статистики. Они задаются самим исследователем. Величина доверительной вероятности определяется тяжестью тех последствий, которые могут произойти в случае, если произойдёт нежелательное событие.

Величина $t_{n,p}$ показывает предельную случайную ошибку расхождения средневыборочного и математического ожидания.

Распределение дисперсии в выборках нормальной совокупности.

Распределение χ^2 Пирсона.

Выборочная дисперсия так же является случайной величиной меняющейся от выборки к выборки.

1) M(X) – известно;
2) M(X) – не известно.

1) Имеется случайная величина X, которая подчиняется нормальному закону с параметрами (m, σ^2),
где: $x_i (i = 1, 2, \ldots, n)$ – независимые наблюдения над случайной величиной.

Для дисперсии мы выбираем вот такую оценку:

$S_*^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - m)^2\right)$ - несмещённая, состоятельная и эффективная оценка дисперсию генеральной совокупности.

$$\frac{x_i - m}{\sigma} = U_i$$

Величина U_i является случайной величиной с параметрами (0;1).

$$\chi^2 = \frac{nS_*^2}{\sigma^2} = \sum_{i=1}^{n} U_i^2$$

Случайная величина представляющая собой сумму квадратов n независимых случайных величин, каждая из которых подчиняется нормальному закону распределения с параметрами (0;1) и независимых случайных величин с распределением χ^2 с к = n – степенями свободы.

Сама функция плотности вероятности $f(\chi^2)$ имеет вид:

$$f\left(\chi^2\right) = L_n \chi^{n-2} e^{-\chi^2/2}$$

Эта функция зависит только от объёма выборки и не зависит ни от математического ожидания, ни от дисперсии, ни от x.

Имеются таблицы распределения χ^2 позволяющие вычислить вероятность события $\left(\chi^2 > \chi^2_{к,\alpha}\right)$

$$P\left(\chi^2, \chi^2_{к,\alpha}\right),$$

где: к – число степеней свободы;
α – доверительная вероятность, которая задаётся самим исследователем.

2) *Математическое ожидание неизвестно*.

Когда случайная величина X с параметрами (m, σ^2) – неизвестны.

Для оценки дисперсии генеральной совокупности используется величина:

$$\widetilde{S}^2 = \frac{1}{n+1} \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2$$

Случайная величина $\frac{n\widetilde{S}^2}{\sigma^2}$ имеет распределение χ^2 с к = n – 1 степенями свободы.

Уменьшение степени свободы использована для получения среднего выборочного.

Доверительный интервал.

Рассмотренные ранее оценки получили название точечных оценок. На практике широко используются интервальные оценки, для получения которых используется метод доверительных интервалов.

В методе доверительных интервалов указывает не одно(точечное) значение интересующего нас параметра, а целый интервал. Он строится на основе неравенства Чебышева:

$$P\left\{\left|\frac{m}{n} - p\right| \le \varepsilon\right\} \ge 1 \frac{Д\left(\frac{m}{n}\right)}{\varepsilon^2} = 1 - \frac{p(1-p)}{n\varepsilon^2} = 1 - \frac{1}{4n\varepsilon^2}$$

Задаётся некоторое число 0 < α < 1 близкое к нулю, которое называется ***уровень значимости***.

Параметр ε находится из неравенства:

$1 - \frac{1}{4n\varepsilon^2} = 1 - \alpha \;\Rightarrow \varepsilon = \frac{1}{2\sqrt{n\alpha}}$, тогда:

$$P\left\{\left|\frac{m}{n}-p\right|\le\frac{1}{2\sqrt{n\alpha}}\right\}\ge 1-\alpha;$$

$$P\left\{\left|\frac{m}{n}-p\right|>\frac{1}{2\sqrt{n\alpha}}\right\}<\alpha;$$

$$P\left\{\frac{m}{n}-\frac{1}{2\sqrt{n\alpha}}\le p\le\frac{m}{n}+\frac{1}{2\sqrt{n\alpha}}\right\}\ge 1-\alpha$$

Интервал $\left(P_*, P^*\right)=\left(\frac{m}{n}-\frac{1}{2\sqrt{n\alpha}}, \frac{m}{n}+\frac{1}{2\sqrt{n\alpha}}\right)$ называется ***доверительным интервалом с уровнем значимости α***.

Доверяясь расчёту мы утверждаем, что неизвестная вероятность принадлежит указанному интервалу, а вероятность возможной ошибки имеющей место тогда, когда этот интервал не накрывает истинное значение α не превосходит уровня значимости α.

n = 1000, m/n = 0,6

При α = 0,1 (0,550; 0,650)

При α = 0,01 (0,442; 0,758)

Истинное значение вероятности P мы незнаем, но можем утверждать, что первый интервал накрывает это значение с вероятностью не менее чем 0,9 , а второй – 0,99.

Пример. Имеется некоторое предположение, гипотеза, о том, что неизвестная вероятность P равна заданному число P_0:

$$H_0: p = p_0; (P_0 = 0,5).$$

Эту гипотезу можно принять, а можно и отклонить посчитав её противоречащей известным статистическим данным.

Для принятия решения(проверки гипотезы) мы проделаем следующую процедуру:

Если $P_0\in(P_*, P^*)$ с α, то гипотезу принимаем(возможно здесь и ошибка, мы можем принять ложную гипотезу – такая ошибка первого рода).

Если $P_0\notin (P_*, P^*)$ с α, то гипотеза отвергается(здесь тоже можем совершить ошибку отклонить верную гипотезу – такая ошибка второго рода, вероятность такой ошибки заранее задаётся нами при построении доверительного интервала).

При наших предположениях, когда уровень значимости равен 0,1 в общем мы имеем $P_0\notin$ (0,550; 0,650). Эта гипотеза отвергается, при этом мы ошибаемся не более чем в 1 случае из 10.

Построение доверительного интервала для математического ожидания.

Случайная величина X распределённая с параметрами (m, σ^2).

Математическое ожидание неизвестно и требуется построить для него доверительный интервал.

1. Известно σ^2.
2. Неизвестно σ^2.

1. σ^2 известно.

Проводится выборка из генеральной совокупности и в качестве несмещённой, состоятельной и эффективной оценки математического ожидания выбирается $\overline{X}$. Оно тоже подчиняется нормальному закону с параметрами:

$\left(m, \sigma^2/n\right)$, где: n – объём выборки.

Нормированная величина:

$$\frac{\overline{X}-m}{\sigma/\sqrt{n}}$$

подчиняется нормальному закону распределения с параметрами (0; 1), тогда вероятность:

$$P\left\{\frac{|\overline{X}-m|}{\sigma/\sqrt{n}} < Z_p\right\} = \Phi(Z_p) = P = 1-\alpha$$

Вероятность задаётся уровнем α, величина P – доверительная вероятность. По таблице находим величину Z_p.

При известном Z_p получим:

$$P\left\{\overline{X} - Z_p\frac{\sigma}{\sqrt{n}} < m < \overline{X} + Z_p\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Интервал для математического ожидания (m_*; m^*) получим:

– ***<u>доверительный интервал для математического ожидания с уровнем значимости α</u>***.

<u>2. σ^2 неизвестно.</u>

Точно так же проводится выборка объёмом n, формируется случайная величина t

$$t = \frac{\overline{X}-m}{\widetilde{S}}\sqrt{n}$$

Случайная величина t имеет распределение Стьюдента.

Зная объём выборки n, задаваясь уровнем значимости α или задаваясь доверительной вероятностью p=1-α.

По распределению Стьюдента находим $t_{n,p}$ – максимальное отклонение m и $\overline{X}$.

$$P\left\{\frac{|\overline{X}-m|}{\widetilde{S}}\sqrt{n} < t_{n,p}\right\} = P = 1-\alpha$$

где: P – доверительная вероятность.

Отсюда легко строится доверительный интервал.

$$P\left\{\overline{X} - \frac{t_{n,p}\widetilde{S}}{\sqrt{n}} < m < \overline{X} + \frac{t_{n,p}\widetilde{S}}{\sqrt{n}}\right\} = 1-\alpha$$

$$(m_*; m^*) = \left(\overline{X} - \frac{t_{n,p}\widetilde{S}}{\sqrt{n}}; \overline{X} + \frac{t_{n,p}\widetilde{S}}{\sqrt{n}}\right)$$

Несмотря на кажущиеся совпадения двух формул они существенно отличаются друг от друга.

Во втором случае величина доверительного интервала зависит не только от доверительной вероятности, но и от объёма выборки.

Это различие наиболее существенно проявляется при малых выборках.

<u>Построение доверительного интервала для дисперсии.</u>

Случайная величина X распределена по нормальному закону с параметрами (m, σ^2).

Требуется построить доверительный интервал для дисперсии по выборочным дисперсия.

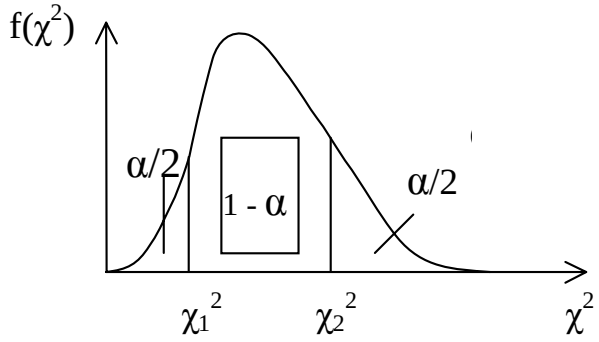
$$\underbrace{S_*^2}_{A} \text{ или } \underbrace{\widetilde{S}^2}_{Б}$$

Построение доверительного интервала для дисперсии основывается на том, что случайные величины:

$\frac{nS_*^2}{\sigma^2}, \frac{n\widetilde{S}^2}{\sigma^2}$ – имеют распределение χ^2 с

к = n, к = n – 1 – степенями свободы.

При заданной доверительной вероятности 1 – α мы записываем:



$$P\left\{\chi_1^2 < \frac{n\widetilde{S}^2}{\sigma^2} < \chi_2^2\right\} = 1-\alpha$$

По таблице распределения χ^2 мы должны выбрать такие два числа χ_1^2 и χ_2^2, чтобы площадь заштрихованная была равна 1-α.

Обычно величины χ_1^2 и χ_2^2 выбирают таким образом, чтобы выполнялось неравенство:

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \alpha/2$$

$$P\left\{\chi_1^2 < \frac{n\widetilde{S}^2}{\sigma^2} < \chi_2^2\right\} = 1 - \overbrace{P\left(\chi^2 < \chi_1^2\right)}^{\alpha/2} - \overbrace{P\left(\chi^2 > \chi_2^2\right)}^{\alpha/2}$$

В таблице распределения χ^2 имеется только вероятность вида:

$$P\left(\chi^2 > \chi_{\kappa,\alpha}^2\right)$$

$$P\left(\chi^2 < \chi_1^2\right) = 1 - P\left(\chi^2 > \chi_2^2\right)$$

Тогда:

$$P\left\{\chi_1^2 < \frac{n\widetilde{S}^2}{\sigma^2} < \chi_2^2\right\} = P\left(\chi^2 > \chi_1^2\right) - P\left(\chi^2 > \chi_2^2\right) = 1-\alpha$$

Преобразуя это неравенство получим:

$$P\left\{\chi_1^2 < \frac{n\widetilde{S}^2}{\sigma^2} < \chi_2^2\right\} = \left(\frac{1}{\chi_2^2} < \frac{\sigma^2}{n\widetilde{S}^2} < \frac{1}{\chi_1^2}\right) = P\left\{\frac{n\widetilde{S}^2}{\chi_2^2} < \sigma^2 < \frac{n\widetilde{S}^2}{\chi_1^2}\right\} = 1-\alpha$$

$$\left(\sigma_*^2, \sigma^{*2}\right) = \left(\frac{n\widetilde{S}^2}{\chi_2^2}; \frac{n\widetilde{S}^2}{\chi_1^2}\right)$$

$$\left(\sigma_*, \sigma^*\right) = \left(\frac{\sqrt{n}\widetilde{S}^2}{\chi_2^2}; \frac{\sqrt{n}\widetilde{S}^2}{\chi_1^2}\right)$$

- доверительный интервал с уровнем значимости α.

Проверка статистических гипотез.

Наряду с оценкой параметров распределения по выборочным данным большой интерес представляет вид (закон) распределения неизвестный на практике. Такие задачи решаются методами статических гипотез.

Относительно неизвестного теоретического распределения формируется некоторое предположение, которое формируется в виде гипотез.

Например, теоретическое распределение подчиняется нормальному, экспоненциальному закону.

При проверки гипотез используется принцип значимости основывающийся на принципе практической невозможности.

Согласно принципу практической невозможности события с очень малыми вероятностями в практических приложениях считаются невозможными.

Максимум таких вероятностей определяет уровень значимости α, который задаётся.

В свою очередь согласно принципу значимости отвергается случайность появления практически невозможного события.

Поскольку теоретическое распределение задано гипотезой, то легко рассчитать вероятность появления некоторого события при проведении испытаний или взятии выборки и пусть такая расчётная вероятность не превышает ε, т.е. событие является практически невозможным.

Если же такое событие происходит, то возникает противоречие между выдвинутой гипотезой и выборкой. Гипотезу следует отвергнуть в этом и заключается содержание принципа значимости.

Проверяемая гипотеза называется нулевой или основной $\mathbf{H_0}$.

Если гипотеза отвергается, то принимается противопоставляемая ей гипотеза H_1, которая называется конкурирующей ил альтернативной.

Про проверки гипотезы $\mathbf{H_0}$ возможны ошибки.

Можно отвергнуть гипотезу $\mathbf{H_0}$ в условиях когда она верна и совершить ошибку I-го рода и можно принять гипотезу, когда она не верна и совершить ошибку II-го рода.

Решение поставленной задачи по сути дела состоит в разделении всего множества выборочных данных на 2-а не пересекающихся подмножества O и W. Таких, что решение принимается в пользу гипотезы H_0, если выборка принадлежит области O и в пользу гипотезы H_1, если выборка принадлежит подмножеству W. Область W называется критической областью выборочного пространства. Здесь гипотеза H_0 отвергается, а область O является областью допустимых значений. Здесь гипотеза H_0 принимается.

Проверка гипотезы о равенстве центров распределения математического ожидания 2-х нормальных генеральных совокупностей.

Задача имеет большой практический интерес. Достаточно часто наблюдается такая ситуация, что средний результат в одной серии эксперимента отличается от среднего результата в другой серии эксперимента.

Возникает вопрос: можно ли объяснить отличительное расхождение случайными ошибками эксперимента и относительно малыми объёмами выборки или это отклонение вызвано какими-либо неизвестными, незамеченными закономерностями.

Имеется две случайных величин X и Y с нормальным законом распределения.

Получим 2-е независимых выборки объёмом n_1 и n_2 из указанных генеральных совокупностей.

Необходимо проверить: H_0: M(X) = M(Y)

H_1: |M(X) – M(Y)| > 0

Рассмотрим два случая:

1. – известны дисперсия генеральной совокупности σ_X^2;
2. – дисперсия неизвестна σ_X^2.

1 - σ_X^2, σ_Y^2 M(X) и M(Y) - неизвестны, для их оценки мы используем средние выборочные $\overline{X}$ и $\overline{Y}$.

Относительно $\overline{X}$ и $\overline{Y}$ известно, что они подчиняются нормальному закону распределения с параметрами:

$$\overline{X}\left(m_X, \sigma_x^2 / n_1\right)$$

$$\overline{Y}\left(m_X, \sigma_y^2 / n_2\right)$$

Рассмотрим случайную величину $\overline{X} - \overline{Y}$. В силу независимости выборок эта случайная величина подчиняется нормальному закону распределения.

Её дисперсия:

$$D\left(\overline{X} - \overline{Y}\right) = D\left(\overline{X}\right) + D\left(\overline{Y}\right) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}$$

Если гипотез H_0 верна(справедлива), то тогда: $M\left(\overline{X} - \overline{Y}\right) = 0$.

Величина:

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}}$$ с параметрами (0, 1)

Выбирая уровень значимости α или доверительную вероятность P = 1- α можем записать:

$$P\left\{\frac{\left|\overline{X} - \overline{Y}\right|}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} \le Z_P\right\} = P = 1 - \alpha = \Phi(Z_P);\ \Phi(Z) = \frac{2}{\sqrt{2\pi}}\int_0^Z e^{-\frac{t^2}{2}}dt;$$

Выбирая по величине интеграла вероятности значения Z_P мы тем самым делим выборочных данных на область допустимых значений и критическую область.

Для области, где выполняется неравенство $|Z| \le Z_P$ – область допустимых значений(ОДЗ) H_0 – принимается.

А, если $|Z| > Z_P$ – критическая область(КО) H_0 – отвергается, H_1 – принимается.

Чем меньше α, тем меньше вероятность отклонить проверяемую гипотезе, если она верна. Но в этом случае увеличивается вероятность совершения ошибки II-го рода.

Чем меньше α, тем больше ОДЗ и тем больше вероятность принять проверяемую гипотезу, если она не верна, т.е. совершить ошибку II-го рода.

Методы проверки гипотез позволяют только отвергнуть проверяемую гипотезу, но они не могут доказать её справедливость.

2 -Дисперсия неизвестна.

Есть 2-е случайных величины X и Y, $\left(m_x, \sigma_x^2\right)\left(m_y, \sigma_y^2\right)$.

$\sigma_x^2 = \sigma_y^2 = \sigma^2 = ?$ m_x и m_y неизвестны берутся независимые выборки $(n_1;n_2)$ и рассматривается гипотеза: H_0: M(X) = M(Y)

H_1: |M(X) – M(Y)| > 0.

Для оценки математического ожидания M(X) и M(Y) используем среднее выборочное $\overline{X}, \overline{Y}$. Для оценки дисперсий используем:

$$\widetilde{S}_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \overline{X}\right)^2$$

$$\widetilde{S}_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \overline{Y}\right)^2$$

- несмещённые, состоятельные оценки дисперсии.

Поскольку генеральные совокупности X и Y имеют одинаковые дисперсии, то для оценки дисперсии σ^2 целесообразно использовать результаты обеих выборок.

Наиболее целесообразной оценкой дисперсии является средняя взвешенная этих двух оценок.

$$\widetilde{S}^2 = \frac{\widetilde{S}_X^2\left(n_1 - 1\right) + \widetilde{S}_Y^2\left(n_2 - 1\right)}{n_1 + n_2 - 2}$$

Если гипотеза H_0 справедлива, то тогда случайная величина $\overline{X} - \overline{Y}$ подчиняется нормальному закону распределения с $M\left(\overline{X} - \overline{Y}\right) = 0$ и с дисперсией $\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$$\left[m_{\overline{X} - \overline{Y}} = 0; \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

$$\tilde{S}_{\bar{X}-\bar{Y}} = \tilde{S}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

$$M\left[\tilde{S}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right] = \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

Если построить случайную величину:

$$t = \frac{\left(\bar{X}-\bar{Y}\right)-M\left(\bar{X}-\bar{Y}\right)}{\tilde{S}_{\bar{X}-\bar{Y}}}$$

, то она будет подчиняться нормальному закону с параметрами (0; 1).

Т.к. σ^2 неизвестна, то такая величина подчиняется t-распределению Стьюдента(со степенями свободы $n_1 + n_2 - 2$).

Для α(P = 1– α) подсчитывается критическое значение $t_{n_1+n_2-2,\alpha}$

$$P\left(|t| > t_{n_1+n_2-2,\alpha}\right) = \alpha$$

Если вычисленные значения $|t| > t_{n_1+n_2-2,\alpha}$, то гипотеза H_0 отвергается и наоборот:

$|t| \le t_{n_1+n_2-2,\alpha}$ H_0 принимается.

Проверка гипотезы о совпадении 2-х дисперсий.

Задача имеет важное практическое значение. Возникает при наладке какого-либо оборудования при сравнении точности приборов, инструментов, методов измерений.

По 2-м независимым выборкам вычислены оценки дисперсий:

$$S_{y_1}^2 \text{ и } S_{y_2}^2$$

$$H_0 : \sigma_{y_1}^2 = \sigma_{y_2}^2$$

$$H_1 : \left|\sigma_{y_1}^2 - \sigma_{y_2}^2\right| > 0$$

Для проверки гипотезы H_0 используется критерий Фишера(F–критерий, F–распределение).

Вычисляется коэффициент:

$$F_н = S_{y_2}^2 \Big/ S_{y_1}^2 \ , \ S_{y_2}^2 > S_{y_1}^2$$

Вычисляется критическое значение $F_{кр}$(α (или P = 1 - α))

$$\nu_2 = n_2 - 1$$

$$\nu_1 = n_1 - 1$$

,где: ν – число степеней свободы числителя и знаменателя.

Если $F_н > F_{кр}$, то H_0 отвергается,

$F_н < F_{кр}$, то H_0 принимается.

Анализ однородности дисперсий.

Понятие однородности является обобщением понятия равенства дисперсий в случае, если число выборок превосходит 2(N > 2).

Для проверки гипотезы H_0:

H_0: $\sigma^2_{y_1} = \sigma^2_{y_2} = ... = \sigma^2_{y_N}$

H_1: дисперсия неоднородна.

Объёмы выборок $n_1, n_2, \dots, n_N$ различны.

Когда объёмы выборок различны для решения задачи является χ^2 с (N-1) степенями свободы.

На практике наиболее частым является когда объёмы выборок одинаковы.

При равных объёмах выборок используется критерий Кохрана для проверки H_0.

Есть соответствующее распределение, но оно громоздко.

В начале вычисляется фактическое значение критерия:

$$G_н = \frac{S^2_{y_{max}}}{\sum_{i=1}^{N} S^2_{y_i}};$$

Отношение максимальной оценки дисперсии к сумме всех оценок дисперсий вычисленных по табличным данным.

Для $P = 1 - \alpha$ вычисляется критическое значение критерия Кохрана $G_{кр}$.

При $G_н \le G_{кр}$ - H_0 принимается;

$G_н > G_{кр}$ - H_0 отвергается.

Проверка гипотез о законе распределения.

Имеется случайная величина X, требуется проверить гипотезу H_0:

H_0: эта случайная величина подчиняется некоторому закону распределения F(x).

Для проверки гипотезы делается выборка состоящая из n независимых наблюдений над случайной величиной X. По выборке строится эмпирическая функция распределения F*(x). Сравнивая эти распределения с помощью некоторого критерия(критерий согласия) делается вывод о том, что эти два распределения согласуются, т.е. H_0 – принимается.

Существует несколько критериев согласия: χ^2 Пирсона, критерий Колмогорова и т.д.

Критерий согласия χ^2 Пирсона.

Имеется случайная величина X, выдвигается гипотеза H_0: F(x), делается выборка.

Диапазон $X_{min} - X_{max}$ разбивается на ℓ интервалов. Размер интервала определяется по правилу Старджесса. $\Delta_1;\Delta_2;\Delta_3;\dots;\Delta_\ell$.

Интервал Δ_i	Δ_1	Δ_2	Δ_3	…	Δ_ℓ
Эмпирическая частота m_i	m_1	m_2	m_3	…	m_ℓ
Теоретическая частота np_i	np_1	np_2	np_2	…	np_ℓ

$$\sum_{i=1}^{\ell} m_i = n;$$

$m_i > 3$(в среднем 5 - 7).

При $m_i < 3$ укрупнить интервал.

Находим частоту попадания случайной величины внутрь каждого интервала.

Поскольку теоретическое распределение задано в гипотезе H_0 всегда можно найти вероятность p_i попадания случайной величины внутрь каждого интервала.

$$\sum_{i=1}^{\ell} p_i = 1;$$

χ^2 Пирсона предполагает, что надо построить:

$$\chi^2 = \sum_{i=1}^{\ell} \frac{(m_i - np_i)^2}{np_i}$$

(имеет распределение χ^2 только при относительно больших n (n > 50)).

Порядок применения χ^2 Пирсона:

1. Рассчитывается эмпирическое значение критерия χ^2;
2. Выбирается уровень значимости α (при P = 1 - α);
3. По таблице подсчитывается $\chi^2_{к,\alpha}$,

где: α – уровень значимости;
к – число степеней свободы.

В общем случае к = ℓ - r – 1,

где: ℓ - количество интервалов разбиения;
r – количество параметров распределения подсчитанных по выборке;

Здесь к = r – 1.

Если $\chi^2 > \chi^2_{к,\alpha}$ гипотеза H_0 отвергается

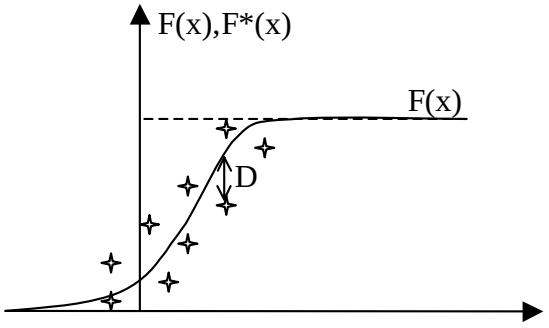
$\chi^2 < \chi^2_{к,\alpha}$ гипотеза H_0 принимается

Критерий Колмогорова.

По результатам выборки объёмом n строится эмпирическая функция распределения F(x). Принимается гипотеза H_0: случайная величина X подчиняется распределению описанному функцией F(x).

За меру расхождения функций принимается величина:

$$D = \max\left|F^*(x) - F(x)\right|$$



Существуют таблицы распределения Колмогорова в которых можно найти:

$\overset{\circ}{D}_n$ - критическое значение. Оно зависит от уровня значимости α(P = 1 - α), величины D и величины выборки n.

Если полученные из опыта значения коэффициента D оказывается больше критического $\overset{\circ}{D}_n$, то H_0 отвергается.

Если $D > \overset{\circ}{D}_n$ гипотеза H_0 отвергается

$D < \overset{\circ}{D}_n$ гипотеза H_0 принимается

С помощью величины $\overset{\circ}{D}_n$ можно построить доверительные границы для неизвестной функции F(x):

$$F^*(x) - \overset{\circ}{D}_n < F(x) < F^*(x) + \overset{\circ}{D}_n$$

Колмогоров показал, что при $n \to \infty$ величина:

$$\lambda = D\sqrt{n}$$

подчиняется распределению Колмогорова.

$$F(\lambda) = \sum_{\kappa=-\infty}^{+\infty} (-1)^{\kappa} e^{-2\kappa^2\lambda^2}$$

Критерий Колмогорова так же может быть использован для статистической проверки принадлежности двух выборок объёмом n_1 и n_2 к одной и той же генеральной совокупности. Вычисляется параметр λ:

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max\left|F_1^*(x) - F_2^*(x)\right|$$

где: F_1^* и F_2^* - эмпирические функции распределения соответственно первой и второй выборки.

По величине λ судят о согласии.

Раздел 6. Основы дисперсионного анализа.

Дисперсионный анализ – это статистический метод анализа результатов наблюдений зависящий от различных одновременно действующих факторов и позволяющий выбрать из ряда факторов наиболее важные, оценивать их влияние.

Основными предпосылками дисперсионного анализа является как правило нормальное распределение результатов наблюдений и отсутствие влияния исследуемых факторов на дисперсию результатов наблюдения.

Обязательным здесь является возможность управляемого изменения фактора в рамках его разновидностей называется *уровнями фактора*. Эти эксперименты могут быть пассивными, когда существование уровней и их смена является естественными для исследуемого объекта и активными, когда эти изменения искусственно вносятся экспериментатором по заранее составленному плану.

Идея дисперсионного анализа в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждый из которых характеризует влияние того или иного фактора, или их взаимодействие. Последующие сравнения этих дисперсий позволяют оценить сущность влияния факторов на исследуемую величину.

Пусть X – это некоторая случайная величина зависящая от 2^x действующих на неё факторов A и B.

$\overline{X}$ - среднее значение исследуемой величины.

Отклонение: $X - \overline{X} = \alpha + \beta + \gamma$

где: α – отклонение вызванное фактором A;
β – отклонение вызванное фактором B;
γ - отклонение вызванное другими факторами.

α, β, γ – случайные величины независимы.

Дисперсию случайной величины X, α, β, γ обозначим:

$$\sigma_x^2, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$$

где: величина σ_γ^2 - остаточная дисперсия учитывающая влияние случайных и прочих неучтённых факторов.

Для независимых и случайных величин имеет место равенство:

$$\sigma_x^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2$$

Сравнивая σ_α^2 или σ_β^2 с величиной σ_γ^2 можно установить степень влияния факторов A и B на величину X по сравнению с неучтёнными и случайными факторами.

Сравнивая между собой σ_α^2 и σ_β^2 мы можем оценить сравнительную степень влияния факторов A и B на величину X.

Дисперсионный анализ позволяет на основании выборочных данных найти все значения дисперсии $\sigma_x^2, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$. Далее используя соответствующие критерии можно оценить степень влияния параметров A и B на исследуемую случайную величину.

Если речь идёт о влиянии одного фактора на исследуемую случайную величину, то речь идёт об однофакторном дисперсионном анализе. Если же речь идёт о многих факторах, то говорят о многофакторном дисперсионном анализе.

Однофакторный дисперсионный анализ.

Большое количество практических задач приводится к задачам однофакторного дисперсионного анализа.

Типичным примером является работа технологической линии в составе которой имеется несколько параллельных рабочих агрегатов.

На выходе имеют место какие-то детали. Эти детали по какому-то параметру можем контролировать.

Ясно, что среднее значения контролируемых параметров после каждого станка будут несколько отличаться.

Вопрос: Обусловлены ли эти отличия действием случайных факторов или имеет место влияние конкретного станка агрегата.

В данном случае фактор только один – станок.

Совокупность размеров деталей подчиняется нормальному закону распределения, и все эти совокупности имеют равные дисперсии.

Имеется m станков, т.о. имеется m совокупностей. Из этих совокупностей мы проводим выборки объёмом n. Так, что значение параметров i-той совокупности i: $x_{i_1}, x_{i_2}, ..., x_{i_n}$.

Все выборки можно записать в виде таблицы, которая называется матрицей наблюдения.

i \ j	1	2	.	j	.	n	Ср. выборочное $\bar{x}_i$
1	x_{11}	x_{12}	.	x_{1j}	.	x_{1n}	$\bar{x}_1$
2	x_{21}	x_{22}	.	x_{2j}	.	x_{2n}	$\bar{x}_2$
.	.	.	.	.	.	.	.
i	x_{i1}	x_{i2}	.	x_{ij}	.	x_{in}	$\bar{x}_i$
.	.	.	.	.	.	.	.
m	x_{m1}	x_{m2}	.	x_{mj}	.	x_{mn}	$\bar{x}_m$

Выдвигаем гипотезу H_0 заключающуюся в равенстве средних выборочных.

$$H_0 : \bar{x}_1 = \bar{x}_2 = ... = \bar{x}_m$$

$$H_1 : \bar{x}_1 \neq \bar{x}_2 \neq ... \neq \bar{x}_m$$ – влияние станков значимо

Гипотеза H_0 проверяется сравнением внутригрупповых и межгрупповых дисперсий по F критерию Фишера.

Если расхождение незначительно, то принимается гипотеза H_0, в противном случае гипотеза H_0 отвергается.

$$\bar{x}_1 = \frac{\sum_{j=1}^{n} x_1 j}{n}; \quad \bar{x}_i = \frac{\sum_{j=1}^{n} x_i j}{n};$$

$$\bar{x} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} = \frac{1}{m} \sum_{i=1}^{m} \bar{x}_i;$$

Далее находят сумму квадратов отклонений от общего среднего:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x}\right)^2=$$

$$=\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}_{Q}+\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\bar{x}_i-\bar{x}\right)^2}_{Q_1}+\underbrace{2\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2\left(\bar{x}_i-\bar{x}\right)^2}_{0}$$

$$Q_1=n\sum_{i=1}^{m}\left(\bar{x}_i-\bar{x}\right)^2$$

Ноль потому, что стоит сумма от $\underbrace{\left(x_{ij}-\bar{x}_i\right)}_{0}\left(\bar{x}_i-\bar{x}\right)$ - сумма отклонений переменных одной совокупности от средней арифметической той же совокупности.

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2}_{Q}=\underbrace{n\sum_{i=1}^{m}\left(\bar{x}_i-\bar{x}\right)^2}_{Q_1}+\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)2}_{Q_2}$$

Слагаемое Q_1 является суммой квадратов разностей между средними отдельных совокупностей и общей средней всех совокупностей. Эта сумма называется суммой квадратов отклонений между группами. Она характеризует систематическое отклонение между совокупностями наблюдений.

Величину Q_1 – рассеяние по фактору.

Слагаемое Q_2 – представляет собой сумма квадратов разностей между отдельными и средней соответствующей совокупности. Эта сумма называется суммой квадратов отклонений внутри группы.

Она характеризует остаточное рассеяние случайных погрешностей совокупностей.

Величина Q называется общей или полной суммой квадратов отклонений отдельных отклонений от общей средней.

Получим оценки дисперсий: S^2, S_1^2, S_2^2.

- дисперсия обусловленная влиянием фактора;

$S_2^2=\dfrac{Q_2}{m(n-1)}=S_{ост}^2$ - остаточная дисперсия – влиянием случайных и других неучтённых факторов.

$S^2=\dfrac{Q}{mn-1}$ - полная дисперсия.

Далее формируем оценку различия между оценками S_1^2 и S_2^2 :

$\dfrac{S_1^2}{S_2^2}=\dfrac{S_\phi^2}{S_{ост}^2}=\dfrac{Q_1/(m-1)}{Q_2/[m(n-1)]}=F_н$ подчиняется распределению f^2 Фишера.

Выбираем уровень значимости α, или доверительной вероятности 1– α = P и по таблице F-распределения с числом степеней свободы: $к_1 = m–1$; $к_2 = m(n–1)$ находим критическое значение $F_{кр,\alpha}$ Фишера.

$$P\{F_н > F_{кр,\alpha}\}=\alpha \qquad P\{F_н \le F_{кр,\alpha}\}=P=1-\alpha$$

Сравнивая между собой $F_н$ и $F_{кр,\alpha}$ мы делаем вывод насколько сильно влияние интересующего нас фактора на исследуемую случайную величину.

В этом и состоит идея дисперсионного анализа.

Однофакторный дисперсионный анализ обычно представляют в виде таблицы.

	Компоненты дисперсии	Оценки дисперсии	Число степеней свободы
Основной фактор	Межгрупповая дисперсия	$S_1^2 = \frac{Q_1}{m-1} = S_\phi^2$	m - 1
Случайные, неучтенные факторы	Внутригрупповая дисперсия	$S_2^2 = \frac{Q_2}{m(n-1)} = S_{ост}^2$	m(n - 1)
	Общая дисперсия	$S^2 = \frac{Q}{mn-1}$	mn - 1

Основы регрессионного и корреляционного анализа.

Связи между различными явлениями в природе сложны и многообразны. В технике чаще всего речь идет о функциональной зависимости. В большинстве случаев интересующие нас явления протекают в условиях воздействия на них множества неконтролируемых факторов. Воздействие каждого из этих факторов в целом невелико, при этом связь теряет строгую функциональность и система переходит не в строго определенное состояние, а в одно из множества возможных. Речь идет о стохастической связи.

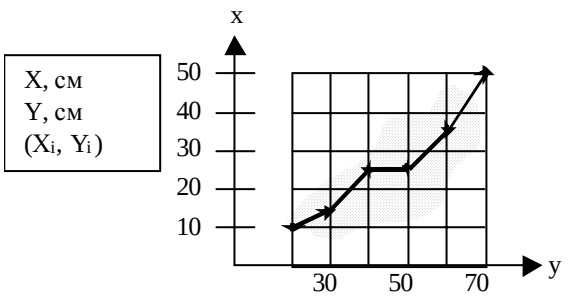
Под ***стохастической*** мы понимаем такую связь, когда одна случайная переменная реагирует на изменения другой случайной переменной изменением своего закона распределения.

Наиболее широко в технике используется частный случай стохастической связи, называемый ***статистической связью***, при которой условное МО некоторой случайной величины Y является функцией от значения, которое принимает другая случайная величина X:

$$M\left(y/x\right) = f(x)$$

Как правило исследуются такие виды статистической связи, при которых значение некоторой случайной переменной зависит в среднем от значений, принимаемых другой случайной переменной:

$$M\left(y/x\right) = f(x) = \overline{Y}(x)$$



Такое представление зависимости между переменными X и Y называется полем корреляции. Можно также построить таблицу корреляции.

Проделывая операцию усреднения для всех тех значений X, по которым есть экспериментальный материал, приходим к тому, что облако исчезает и получается набор точек, представляющих средние значения. Соединяя эти точки, получаем ломанную, называемую ***эмпирической линией регрессии***.

Связь между СВ характеризуется формой и теснотой связи.

Определение формы связи и понятие регрессии.

Определить форму связи между СВ – значит выявить механизм получения зависимой случайной величины. При изучении статистических связей, форму связей характеризует функция регрессии:

$M\left(Y/X = x\right) = f(x)$ - зависимость условного МО

Если св X и Y зависимы, то МО их произведения:

$$M(xy) = M(x)M\left(y/x\right) = M(y)M\left(x/y\right)$$

Регрессия св Y относительно X определяется как:

$$M\left(Y/X = x\right) = \int_{-\infty}^{+\infty} yf\left(y/x\right)dy,$$

где $f\left(y/x\right)dy$ - условная плотность вероятности по формуле Байеса:

$$f\left(y/x\right) = \frac{f(x,y)}{f(x)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy}$$

$M\left(X/Y = y\right) = \int_{-\infty}^{+\infty} xf(x,y)dx$ - регрессия X по Y.

Функция регрессии имеет важное практическое значение. Она может быть использована для прогноза значений, которые может принимать известная случайная величина при ставших известными значениях другой случайной величины.

Точность прогноза определяется дисперсией условного распределения:

$$\sigma^2\left(Y/X = x\right) = M\left\{Y/X = x - M\left(Y/X = x\right)\right\}^2 = M\left(Y^2/X = x\right) - M^2\left(Y/X = x\right)$$

учитывая: $\sigma^2(x) = M(x - m_x)^2 = M(x)^2 - M^2(x)$

Несмотря на важность функции регрессии, возможности ее практического использования ограничены, т.к. для ее вычисления необходимо знать аналитический вид двумерной функции {x,y}. Мы же, как правило, имеем выборку ограниченного объема.

Традиционный путь приводи к большим ошибкам, т.к. одну и ту же совокупность точек на плоскости можно описать с помощью различных функций.

Другой характеристикой формы связи, используемой на практике, стала ***кривая регрессии*** – зависимость условного среднего случайной величины от значения, которое принимает случайная величина X: $\overline{Y}(x) = f(x)$.

Определение кривой регрессии инвариантно закона совместного распределения св X и Y. Важное значение в практике *имеет двумерный нормальный закон распределения*. Особенностью этого распределения является то, что условные МО совпадают с условными средними. При этом функция регрессии совпадает с кривой регрессии.

Линейная регрессия (ЛР). Метод наименьших квадратов.

Линейная регрессия занимает в технике и теории корреляции особое место. Она обусловлена двумерным нормальным законом распределения СВ X и Y:

$\overline{Y}(x) = a_0 + a_1 x$, где

a_0 и a_1 – коэффициенты регрессии,

x – независимая случайная величина

Параметры уравнения регрессии определяются методом наименьших квадратов, предложенным Лагранжем и Гауссом, который сводится к следующему.

Строятся квадратичные формы:

$$Q = \sum_{i=1}^{n} (x_i - \varepsilon)^2 \to \min$$

x_i – измеренное значение переменной,

ε - истинное или теоретическое значение этой величины.

Требуется, чтобы сумма квадратов отклонений измеренных значений относительно истинных была минимальна.

В случае линейной регрессии за теоретическое значение принимается значение $\overline{Y}(x)$, т.е. ищется такая прямая линия с коэффициентами a_0 и a_1, чтобы сумма квадратов отклонений от этой линии была минимальна.

$$Q = \sum_{i=1}^{n} (y_i - a_0 - a_1 x)^2 ,$$

y_i – измеренное значение переменной Y.

Минимальные квадратичные формы получают, приравнивая к нулю ее производные по a_0 и a_1:

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = -2\sum (y - a_0 - a_1 x) = 0 & a_0, a_1 = const \\ \dfrac{\partial Q}{\partial a_1} = -2\sum (y - a_0 - a_1 x) x = 0 & \sum_{i=0}^{n} a_0 = n a_0 \quad \sum_{i=0}^{n} a_1 x = a_1 \sum_{i=0}^{n} x \end{cases}$$

$$\begin{cases} n a_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum yx \end{cases}$$

$$a_0 = \frac{\sum y \sum x^2 - \sum x \sum yx}{n\sum x^2 - \left(\sum x\right)^2} \qquad a_1 = \frac{n\sum yx - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

<u>Нелинейная регрессия (НР).</u>

Форма связи между условными средними определяется уравнениями регрессии. В зависимости от вида уравнений можно говорить о ЛР или НР.

В общем случае эта зависимость может быть представлена в виде полинома степени k:

$$\overline{Y}(x) = a_0 + a_1 x + a_1 x^2 + ... + a_k x^k$$

Определение коэффициентов регресии производится по методу наименьших квадратов:

$$Q = \sum \left(y_i - \overline{Y}(x)\right)^2 \to \min$$

$$\frac{\partial Q}{\partial a_0}, \frac{\partial Q}{\partial a_1}, \frac{\partial Q}{\partial a_2}, ..., \frac{\partial Q}{\partial a_k}$$

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = -2\sum \left(y - a_0 - a_1 x - a_2 x^2 - ... - a_k x^k\right) = 0 \\ \dfrac{\partial Q}{\partial a_1} = -2\sum \left(y - a_0 - a_1 x - a_2 x^2 - ... - a_k x^k\right) x = 0 \\ ..... \\ \dfrac{\partial Q}{\partial a_k} = -2\sum \left(y - a_0 - a_1 x - a_2 x^2 - ... - a_k x^k\right) x^k = 0 \end{cases}$$

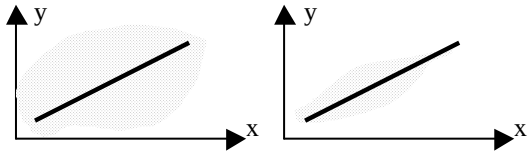
В результате получаем систему нормированных уравнений:

$$\begin{cases} a_0 n + a_1 \sum \mathrm{x} + \mathrm{a}_2 \sum \mathrm{x}^2 + \ldots + \mathrm{a}_\mathrm{k} \sum \mathrm{x}^\mathrm{k} = \sum y \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum \mathrm{x}^3 + \ldots + \mathrm{a}_\mathrm{k} \sum \mathrm{x}^{\mathrm{k}+1} = \sum yx \\ \ldots\ldots \\ a_0 \sum x^k + a_1 \sum x^{k+1} + a_2 \sum \mathrm{x}^{\mathrm{k}+2} + \ldots + \mathrm{a}_\mathrm{k} \sum \mathrm{x}^{2\mathrm{k}} = \sum yx^{\mathrm{k}} \end{cases}$$

Решая полученную систему известным способом, находим коэффициенты регрессии.

<u>Измерение тесноты связи.</u>

Если бы величина Y полностью определялась аргументом X, все точки лежали бы на линии регрессии. Чем сильнее влияние прочих факторов, тем дальше отстоят точки от линии регрессии. В случае в) связь между X и Y является более тесной.



а) в)

За основу показателя, характеризующего тесноту связи, берется общий показатель изменчивости дисперсии:

$$\sigma_y^2 = M\left[\left(Y - m_y\right)\right]^2 = M\left[\left(Y + \overline{Y}(x) - \overline{Y}(x) - m_y\right)\right]^2 =$$

$$= \underbrace{M\left[\left(Y - \overline{Y}(x)\right)\right]^2}_{\sigma^2 y/x} + \underbrace{M\left[\overline{Y}(x) - m_y\right]^2}_{\delta^2 y/x} + 2\underbrace{M\left[\left(Y - \overline{Y}(x)\right)\left(\overline{Y}(x) - m_y\right)\right]}_{0}$$

$$\sum_y \left(Y - \overline{Y}(x)\right) p\left(y/x\right) = \sum p\left(y/x\right) - \overline{Y}(x) \sum p\left(y/x\right) = 0$$

$$\boxed{\sigma_y^2 = \sigma_{y/x}^2 + \delta_{y/x}^2} \qquad (*)$$

$\boxed{\sigma_{y/x}^2}$ - дисперсия переменной Y относительно теоретической линии дисперсии, определяющей влияние прочих факторов на величину Y.

$\boxed{\delta_{y/x}^2}$ - условная дисперсия, характеризует дисперсию теоретической линии регрессии относительно условной генеральной средней m_y. Именно она определяет влияние данного фактора (X) на величину Y и может быть использована для оценки тесноты связи между величинами X и Y.

$\boxed{\eta_{T,y/x} = \dfrac{M\left(Y(x) - m_y\right)^2}{\sigma_y^2} = \dfrac{\delta_{y/x}^2}{\sigma_y^2}}$ - теоретическое корреляционное отношение.

Изменяется от 0 до 1, что легко доказать, поделив (*) на σ_y^2:

$$1 = \frac{\sigma_{y/x}^2}{\sigma_y^2} + \frac{\delta_{y/x}^2}{\sigma_y^2}$$

$$\eta_{T,y/x} = 1 - \frac{\sigma_{y/x}^2}{\sigma_y^2}; \qquad \sigma_{y/x}^2 \le \sigma_y^2$$

1) Если $\eta_{T,y/x} = 1$, то $\sigma_{y/x}^2 = 0$

Влияние прочих факторов отсутствует. Все распределение будет сконцентрировано на линии регрессии. В этом случае между X и Y существует простая функциональная зависимость.

2) Если $\eta_{T,y/x}$=0, когда $\overline{Y}(x) = m_y$.

В этом случае линия регрессии Y по X будет горизонтальной прямой, проходящей через центр распределения.

В случае, когда вид зависимости (форма связи) случайных величин X и Y не установлен, часто бывает необходимо убедиться в наличии какой-либо связи вообще. Может оказаться, что связь несущественна и вычисление коэффициентов регрессии неоправданно.

Для объяснения такого вопроса вычисляется эмпирическое корреляционное отношение, определяемое на основе выборочных данных. При выводе формул для ЭКО пользуются эмпирической линией регрессии и оценкой дисперсии по выборке.

Определение эмпирического корреляционного соотношения.

$$S_Y^2 = \frac{1}{n}\sum\left[Y - \overline{Y_i}(x)\right]^2 + \frac{1}{k}\sum\left[\overline{Y_i}(x) - \overline{Y}\right]^2 = S_{y(x)}^2 + \delta_{y(x)}^2$$

y – измеряемое значение зависимой переменной

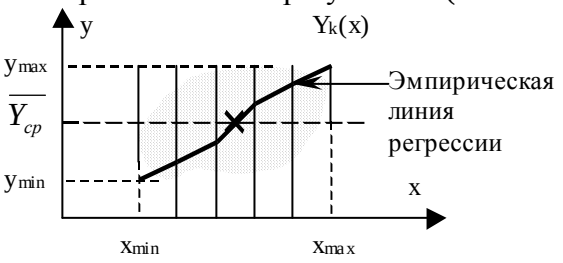
n – общее количество измерений

$\overline{Y_i}(x)$ - условное среднее (среднее значение зависимой переменной y в i-ом интервале св X)

k – общее количество интервалов

$\overline{Y}$ - среднее всей совокупности измерений

В пределах каждого интервала, для всех тех значений X, для которых есть экспериментальные результаты (значения Y), находим средние значения.



$\mathbf{S_{y(x)}^2}$ – составляющая полной дисперсии, характеризует дисперсию результатов измерений относительно эмпирической линии регрессии, т.е. влияние прочих факторов на зависимую переменную Y.

$\delta_{y(x)}^2$ – характеризует дисперсию эмпирической линии регрессии относительно среднего всей совокупности, т.е. влияние исследуемого фактора на зависимую переменную Y.

$$\eta^2 = \frac{\delta_{y(x)}^2}{S_y^2} = 1 - \frac{S_{y(x)}^2}{S_y^2}$$ - Эмпирическое корреляционное соотношение

Из сравнения с формулой для теоретического корреляционного соотношения видно: при расчете теоретического корреляционного соотношения необходимо знать форму связи между переменными.

При вычислении эмпирического корреляционного соотношения никакие предположения о форме связи не используются, нужна только эмпирическая линия регрессии.

Свойства:

1. $0 \le \eta^2 \le 1$
2. если η^2=1, все точки корреляционного поля лежат на линии регрессии – функциональная связь между X и Y.
3. Если η^2=0 (когда $\sum\left[\overline{Y_i}(x) - \overline{Y}\right]^2 = 0$), отсутствует изменчивость условных средних $\left[\overline{Y_i}(x)\right]$, эмпирическая линия регрессии проходит параллельно оси абсцисс – связи между X и Y нет.

Эмпирическое корреляционное соотношение η^2 завышает тесноту связи между переменными и случайными величинами, причем тем сильнее, чем меньше число измерений, поэтому η^2 рекомендуется использовать для предварительной оценки тесноты связи, а для окончательной оценки – теоретическое корреляционное соотношение.

Коэфициент корреляции.

Рассмотрим случай вычисления теоретического корреляционного соотношения η_m^2, когда связь между случайными величинами X и Y является ***линейной.***

$$\overline{Y}(x) = a_0 + a_1 x$$

Такая форма связи между X и Y имеет место в случае, когда случайные величины подчиняются двумерному нормальному закону распределения.

$$\eta_m^2 = \frac{M\left[\overline{Y}(x) - m_y\right]^2}{\sigma_y^2} = \frac{\sum\left[\overline{Y}(x) - \overline{Y}\right]^2}{n\sigma_y^2}$$

Подставив вместо Y и $\overline{Y}$ их значения для случая линейной зависимости:

$$\overline{Y} = \frac{\sum y}{n} = \frac{na_0 + a_1\sum x}{n} = a_0 + \overline{a_1 x}$$

$\overline{Y}$ (x)=a_0 + a_1x

$$\eta_m^2 = \frac{\sum\left(a_0 + a_1 x - a_0 - a_1\overline{x}\right)^2}{n\sigma_Y^2} = \frac{\sum a_1^2\left(x - \overline{x}\right)^2}{n\sigma_Y^2} = a_1^2\frac{n\sigma_X^2}{n\sigma_Y^2}$$

Заменим a_1 ее значением, полученным из решения нормальных уравнений:

$$\frac{a_1^2\sigma_x^2}{\sigma_y^2} = \left[\frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}\right]^2\frac{\sigma_x^2}{\sigma_y^2} = \left[\frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}\right]^2\frac{\sigma_x^2}{\sigma_y^2} = \frac{\left[\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}\right]^2}{\sigma_x^2\sigma_{Yx}^2}$$

$$\boxed{r = a_1\frac{\sigma_x}{\sigma_Y} = \frac{\frac{\sum xy}{n} - \frac{\sum x\sum y}{n}}{\sigma_x\sigma_Y} = \frac{\sum\left(x - \overline{x}\right)\left(y - \overline{y}\right)}{n\sigma_x\sigma_Y}}$$

Коэфициент корреляции является частным случаем теоретического корреляционного соотношения η_m^2, когда связь между СВ является линейной. В этом случае r является показателем тесноты связи.

$k_{x,y} = \frac{\sum\left(x - \overline{x}\right)\left(y - \overline{y}\right)}{n}$ - выборочный корреляционный момент

$$\rho_{x,y} = \frac{\sum\left(x - \overline{x}\right)\left(y - \overline{y}\right)}{n\sigma_x^2} \qquad \boxed{r = \sqrt{\rho(x,y)\rho(y,x)}}$$

Выборочный коэфициент корреляции обладает свойствами:

1. r=0, если св X и Y независимы
2. $|r| < 1$ - Для любых св X и Y
3. $|r| = 1$ - Для случая линейной зависимости св X и Y. $r = \begin{cases} 1 & a_1 > 0 \\ -1 & a_1 < 0 \end{cases}$

Коэфициент корреляции используется для оценки тесноты связи и в случае нелинейной зависимости между случайными величинами.

Если предварительный графический анализ поля корреляции указывает на какую либо тесноту связи, полезно вычислить коэфициент корреляции.

Если модуль коэфициента корреляции $|r| = 0.8 \div 0.9$, то независимо от вида связи можно считать, что она достаточно тесна, чтобы исследоват ее форму.

Двумерное нормальное распределение.

Его возникновение объясняется центральной предельной теоремой Ляпунова:

$$f_{xy}(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{1}{2(1-\rho^2)}\cdot\left(\left(\frac{x-m_x}{\sigma_x}\right)^2+\left(\frac{y-m_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)\right)\right)$$

ρ – коэффициент корреляции. X и У по отдельности распределены нормально (m_x,σ_x) и (m_y,σ_y).

В частном случае независимых СВ X и У ρ=0:

$$f_{xy}(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\cdot\exp\left(-\frac{1}{2}\cdot\left(\left(\frac{x-m_x}{\sigma_x}\right)^2+\left(\frac{y-m_y}{\sigma_y}\right)^2\right)\right)$$

Исходные плотности одномерных нормальных распределений X и У:

$$f_x(x)=\int_{-\infty}^{+\infty}f_{xy}(x,y)dx=\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(\frac{-(x-m_x)^2}{2\sigma_x^2}\right)$$

$$f(y/x)=\frac{f_{xy}(x,y)}{f_x(x)}=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{1}{2}\cdot\left(\frac{y-m_y-\rho\cdot\frac{\sigma_y}{\sigma_x}\cdot(x-m_x)^2}{\sigma_y\sqrt{1-\rho^2}}\right)\right)$$

<u>Условное распределение</u> – нормальное с условиями:

$$M(y/x)=m_y+\rho\frac{\sigma_y}{\sigma_x}(x-m_x)\ \text{ и }\ \sigma_{y/x}=\sigma_y\sqrt{1-\rho^2}\ .$$

Первое условие является уравнением функции регрессии.

$$M(x/y)=m_x+\rho\frac{\sigma_x}{\sigma_y}(y-m_y)\ \text{ и }\ \sigma_{x/y}=\sigma_x\sqrt{1-\rho^2}\ .$$

Нормальная регрессия прямолинейна. Точность оценки у/х одинакова для всех х. В качестве меры тесноты связи используется коэффициент корреляции, а форму связи при этом характеризует коэффициент регрессии.

Z=f_{xy}(x,y) – трехмерная поверхность, сечения которой плоскостями XZ и YZ представляют собой графики плотности одномерных распределений.

Коэффициент множественной корреляции

$$R=\sqrt{\beta_1 r_{yx1}+\beta_2 r_{yx2}+\ldots+\beta_n r_{yxn}}=\sqrt{\Delta^*/\Delta}$$

$$\Delta=(-1)^{n+1}\begin{vmatrix}1 & r_{12} & \ldots & r_{1n}\\ r_{21} & 1 & \ldots & r_{2n}\\ \ldots & \ldots & \ldots & \ldots\\ r_{n1} & r_{n2} & \ldots & 1\end{vmatrix}$$

Δ* – это Δ с добавочными верхней строкой и правым столбцом, состоящих из свободных членов уравнений.

<u>Пример</u>: Вычислить КМК:

$$\Delta=\begin{vmatrix}1 & 0{,}57\\ 0{,}57 & 1\end{vmatrix}=0{,}675\qquad \Delta^*=\begin{vmatrix}-0{,}875 & -0{,}69 & 0\\ 1 & 0{,}57 & -0{,}69\\ 0{,}57 & 1 & -0{,}825\end{vmatrix}=0{,}5\qquad R=\sqrt{\Delta^*/\Delta}=0{,}86$$

<u>Коэффициент корреляции рангов (объединенные ранги)</u>

Анализ информации неподдающейся количественной оценке.

На экзаменах разные экзаменаторы ставят одним и тем же студентам разные оценки. Чтобы исключить элемент субъективизма, всех учащихся располагают в соответствии со степенью их

способностей и ранжируют. Корреляция между рангами значительно точнее отражает взаимосвязь.

Есть n учащихся и ранги по некоторому фактору A: $X_1 \dots X_n$ и по фактору B: $Y_1 \dots Y_n$.

X_i, Y_i – перестановки n первых натуральных чисел.

X_k-Y_k=d_k – мера тесноты связи A и B. Если все d_k=0, то A и B полностью соответствуют.

$$\sum X_k = \frac{n(n+1)}{n}, \quad \overline{X} = \frac{n+1}{2}$$

$$\rho = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{n\sigma_X \sigma_Y} = \frac{\sum \left(X_k - \frac{n+1}{2}\right)\left(Y_k - \frac{n+1}{2}\right)}{\sqrt{\sum \left(X_k - \frac{n+1}{2}\right)^2 \sum \left(Y_k - \frac{n+1}{2}\right)^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{12\sum xy}{n^3-n} = 1 - \frac{6\sum d^2}{n^3-n}$$

$$\rho = 1 - \frac{6\sum d^2}{n^3-n}$$

Последнее выражение – ***коэффициент корреляции рангов Спирмена***.

Существуют и другие показатели тесноты связи:

ККР Кендела: удобен для углубленных исследований, когда невозможно установит ранговые различия. Строятся объединенные усредненные ранги и

$$\rho = \frac{\frac{n^3-n}{6} - (T_x + T_y) - \sum d^2}{\sqrt{\left(\frac{n^3-n}{6} - 2T_x\right)\left(\frac{n^3-n}{6} - 2T_y\right)}}$$

$$T_x = T_y = \sum_{i=1}^{l} \frac{t_i^3 - t_i}{12}$$

t_i – число объединенных рангов.

Метод ранговой корреляции

Позволяет анализировать множество факторов и выделять доминирующие.

Для построения математической модели процесса необходимо выделить из множества факторов доминирующие. На первом этапе это делается с помощью экспертных оценок: максимальному кругу специалистов предлагается расположить факторы в порядке убывания степени влияния. При этом предлагается максимально полный список факторов, хотя каждый может включать в этот список дополнительные факторы.

Результат – матрица рангов, которая строится с учетом квалификации опрашиваемого: показания специалистов умножаются на коэффициент квалификации. Чем меньше сумма рангов фактора, тем более важное место он занимает, тем большее влияние он оказывает на выходной параметр.

Если распределение на диаграмме близко к равномерному, то все факторы должны учитываться. Обычно отмечается, что опрос не дал желаемого результата.

Если не равномерно, но изменение рангов не велико, значит специалисты делают различия между факторами, но неуверенно. Таким образом, надо учитывать все факторы.

Наиболее благоприятен случай быстрого экспоненциального спада суммы рангов. Малозначащие факторы отсеиваются. Для оценки степени согласованности мнений специалистов вычисляется ***коэффициент конкордации***:

$$W = \frac{12\sum d^2}{m^2(n^3-n)}$$

m – число специалистов

n – число факторов.

Чем больше W, тем больше степень согласованности. Если W=0, то согласованность отсутствует. При W=1 – полная согласованность.

Планирование эксперимента

Классический регрессионный и корреляционный анализ базируются на пассивном эксперименте, который сводится к сбору и обработке данных, полученных в результате наблюдения за процессом или явлением.

Привлекательность пассивного эксперимента в том, что он избавляет от необходимости тратить время и средства на постановку опытов. Полученные результаты в виде уравнения регрессии можно затем использовать для управления процессом. Однако пассивный эксперимент имеет ряд недостатков:

1. При сборе экспериментальных данных на реальном действующем промышленном объекте во избежание появления брака возможны лишь незначительные изменения параметров процесса. При этом интервалы варьирования параметров оказываются столь малыми, что изменение выходной величины будет в значительной степени обусловлено воздействием случайных факторов.

2. Часто упускают из вида важные факторы из-за невозможности их измерения или регистрации.

3. При пассивном эксперименте нельзя произвольно варьировать параметры. В результате этого экспериментальные точки часто располагаются неудачно и при большом количестве опытов затрудняют точное описание процесса.

Активный эксперимент

Ставится по плану. Достоинства:

1. Появляется четкая логическая схема всего исследования.

2. Повышается эффективность исследования. Оказывается возможным извлечь максимальное количество информации.

3. Обработка результатов эксперимента осуществляется стандартными приемами.

4. Планирование эксперимента позволяет обеспечить случайный порядок проведения опытов (рандомизация).

Отпадает необходимость в жесткой стабилизации мешающих факторов.

Активный эксперимент эффективен в лабораторной практике, а пассивный – в производстве.

С помощью методов планирования эксперимента можно получить математическую модель изучаемого процесса в аналитическом виде при отсутствии сведений о механизме процесса.

Математическая модель процесса задается полиномом:

$$y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i \neq j} b_{ij} x_i x_j + \sum_{i=1}^{k} b_{ii} x_i^2 + \ldots$$

Чаще всего используется линейная модель:

$$y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i \neq j} b_{ij} x_i x_j$$

План эксперимента определяет расположение точек в к-мерном факторном пространстве.

Матрица планирования: каждая строчка – условие проведения опыта, а столбец – значения переменной в различных опытах.

При выборе линейной модели достаточно варьировать каждый фактор на двух уровнях. Если при этом осуществляются все возможные комбинации из k факторов, то реализация эксперимента по такому плану называется полным факторным экспериментом типа 2^k (ПФЭ 2^k).

Построение математической модели методом ПФЭ проводится в следующем порядке:

1. Планирование эксперимента
2. Проведение эксперимента
3. Проверка воспроизводимости
4. Построение математической модели с проверкой статистической значимости всех коэффициентов
5. Проверка адекватности математической модели.

Центр плана (точка, вокруг которой ставится серия опытов) выбирается на основании априорных сведений о процессе.

Если эти сведения отсутствуют, то в качестве центра плана выбирается центр исследуемой области.