ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Задачи математической статистики

Первая задача математической статистики - указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.

Вторая задача математической статистики - разработать методы анализа статистических данных в зависимости от целей исследования:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости случайной величины от одной или нескольких случайных величин и др.;

б) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид

Генеральная и выборочная совокупности

Выборочной совокупностью или просто выборкой называют совокупность случайно отобранных объектов. **Генеральной совокупностью** называют совокупность объектов, из которых производится выборка. **Объемом совокупности** (выборочной или генеральной) называют число объектов этой совокупности.

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 – n_2 раз, x_k - n_k раз и $\sum n_i = n$ - объем выборки. Наблюдаемые значения x_i - называют **вариантами**, а последовательность вариант, записанных в возрастающем порядке - **вариационным рядом**. Числа наблюдений называют частотами, а их отношения $\frac{n_i}{n} = W_i$ к объему выборки - **относительными частотами**. **Статистическим распределением выборки** называют перечень вариант и соответствующих им частот или

Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения ***x*** относительную частоту события ***X < x***.

Функцию распределения ***F*(*x*)=*P*(*X*<*x*)** генеральной совокупности называют **теоретической функцией распределения**. Из теоремы Бернулли следует, что относительная частота события ***X < x***, т.е. $F^*(x)$ стремится по вероятности к вероятности ***F(x)*** этого события. Отсюда следует целесообразность использования эмпирической функции распределения выборки для приближенного представления теоретической (интегральной) функции распределения генеральной совокупности.

Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Будем называть выборочной статистикой любую величину, полученную в результате обработки данных эксперимента. Любая выборочная статистика является случайной величиной.

Определение. *Точечной оценкой параметра* ***x*** *называется его приближенное значение* x^**, полученное в результате обработки выборки.*

Определение. *Оценка* x^* *параметра* ***x*** *называется* ***несмещенной****, если математическое ожидание оценки совпадает с оцениваемым параметром:* ***M(*** x^* ***) = x****.* ***Эффективной*** *называют статистическую оценку, которая (при заданном объеме выборки* ***n****) имеет наименьшую возможную дисперсию.*

Любая величина, определенная по выборке, называется статистикой. Очевидно, что любая статистика – случайная величина.

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Определение. *Оценка* x^* *параметра* ***x*** *называется* ***состоятельной****, если при увеличении объема выборки* ***n*** *для любого* ***ε ≥ 0*** *вероятность отклонения оценки* x^* *от параметра* ***x*** *на величину, меньшую* ***ε****, равна 1:*

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

Несмещенной и состоятельной оценкой математического ожидания ***a*** оцениваемого признака ***θ*** является выборочная средняя:

$$\bar{\theta} = \frac{\sum_{i=1}^{n} \theta_i}{n}$$

, где $\boldsymbol{\theta_i}$ - варианты выборки выходного параметра ***θ*** .

Несмещенной и состоятельной оценкой дисперсии ***D*** является **выборочная дисперсия** $\boldsymbol{s^2}$:

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

Выборочным средним квадратическим отклонением $s = \sqrt{s^2}$

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Медианой m_e называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т. е. $n = 2k+1$, $m_e = x_{k+1}$, при четном $n = 2k$ медиана $m_e = \frac{x_k + x_{k+1}}{2}$.

Размахом варьирования R называют разность между наибольшей и наименьшей вариантами: $R = x_{max} - x_{min}$.

Средним абсолютным отклонением θ называют среднее арифметическое абсолютных отклонений:

$$\theta = \frac{\sum_{i=1} |x_i - \bar{x}|}{n}$$

Коэффициентом вариации V называют выраженное в процентах отношение выборочного среднего квадратического отклонения к выборочной средней:

$$V = \frac{S}{\bar{x}} \cdot 100\%$$

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Оценкой асимметрии теоретического распределения $A_s = \frac{\mu_3}{\sigma^3}$

является ее выборочное значение: $A_в = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$

Оценкой теоретического эксцесса $E_k = \frac{\mu_4}{\sigma^4} - 3$ является его выборочное значение:

$$E_в = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Доверительным интервалом для параметра ***θ*** называется интервал ***(θ*** $_1$***,θ*** $_2$***)***, накрывающий истинное значение ***θ*** с заданной вероятностью ***γ = 1 - α***, то есть $P[\theta_1<\theta<\theta_2]=\gamma$. Число ***γ = 1 - α*** называется доверительной вероятностью, а значение ***α*** - уровнем значимости. Статистики ***θ*** $_1$ и ***θ*** $_2$, определяемые по выборке из генеральной совокупности с неизвестным параметром ***θ***, называются нижней и верхней границами доверительного интервала (или левой и правой границами).

Чтобы найти доверительный интервал для параметра ***θ***, необходимо знать закон распределения статистики $\tilde{\theta}=\tilde{\theta}(x_1,\dots,x_n)$, значение которой является точечной оценкой параметра ***θ***.

Пусть существует статистика $Y=Y(\tilde{\theta},\theta)$ такая, что:

а) закон распределения ***Y*** известен и не зависит от ***θ***,

б) функция $Y(\tilde{\theta},\theta)$ непрерывна и строго монотонна по ***θ***.

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Пусть $\gamma = 1 - \alpha$ - заданная доверительная вероятность, а $y_{\frac{\alpha}{2}} = y_{\frac{1-\gamma}{2}}$ и $y_{1-\frac{\alpha}{2}} = y_{\frac{1+\gamma}{2}}$ - квантили статистики Y порядков $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$. Тогда с вероятностью $\gamma = 1 - \alpha$ выполняется неравенство:

$$y_{\frac{\alpha}{2}} < Y\left(\tilde{\theta}, \theta\right) < y_{1-\frac{\alpha}{2}}$$

.

Решая неравенство относительно θ, найдем границы θ_1 и θ_2 доверительного интервала для θ.

Распределение некоторых статистик для НСВ X ~ N(m,σ)

1. Выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ имеет нормальное $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

Распределение некоторых статистик для НСВ X ~ N(m,σ)

2. Выборочная дисперсия $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$ связана со случайной величиной $\chi^2(n-1)$ соотношением: $s^2=\frac{\sigma^2}{n-1}\chi^2(n-1)$.

3. Статистика $\frac{\overline{x}-m}{s/\sqrt{n}}$ является случайной величиной, имеющей распределение Стьюдента с **(*n* – *1*)** степенью свободы ***T*(*n* – *1*)**.

3. Пусть s_1^2 и s_2^2 - выборочные дисперсии, вычисленные по независимым выборкам объема n_1 и n_1 из двух нормально распределенных генеральных совокупностей с дисперсиями σ_1^2 σ_2^2 и ,

тогда отношение $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ имеет распределение Фишера $F(n_1-1,n_2-1)$

Доверительный интервал для среднего при известной *σ*

Пусть $x_1, \dots, x_n$ - выборка *n* наблюдений случайной величины ***X ~ N*(*m*, *σ*)**. Рассмотрим статистику $Y = \dfrac{\bar{x} - m}{\sigma / \sqrt{n}}$ **~ *N*(*0,1*)**. Вероятность события $u_{\frac{\alpha}{2}} < Y < u_{1-\frac{\alpha}{2}}$ равна заданной доверительной вероятности ***γ***, то есть $P\left[u_{\frac{\alpha}{2}} < Y < u_{1-\frac{\alpha}{2}}\right] = \gamma$. Для квантилей стандартного (нормированного) нормального распределения $u_{\frac{\alpha}{2}} = -u_{1-\frac{\alpha}{2}}$. После решения неравенства относительно ***m*** получим доверительный интервал в виде:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} < m < \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}$$

Доверительный интервал для дисперсии

В качестве оценки дисперсии используют выборочную дисперсию $s^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}\right)$. Статистика $Y = \frac{(n-1)s^2}{\sigma^2}$ имеет распределение $\chi^2(n-1)$.

Определим квантили $\chi^2_{\frac{\alpha}{2}}(n-1)$ и $\chi^2_{1-\frac{\alpha}{2}}(n-1)$, удовлетворяющие условию

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

Решая неравенство относительно σ^2, получим доверительный интервал для дисперсии:

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

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистической гипотезой *H* называется предположение относительно параметров или вида распределения случайной величины ***X***. Если по выборке наблюдений необходимо проверить предположение о значении параметров известного распределения, то такие гипотезы называются **параметрическими**.

Проверяемая гипотеза называется **нулевой гипотезой H_0**. Вместе с гипотезой H_0 рассматривают одну из **альтернативных (конкурирующих)** гипотез H_1. Правило, по которому принимается решение принять или отклонить гипотезу H_0, называется **критерием**. Статистика Z, на основании которой принимается решение об истинности гипотезы H_0, называют **статистикой критерия**.

Перед анализом выборки назначается некоторая малая вероятность α, называемая **уровнем значимости**.

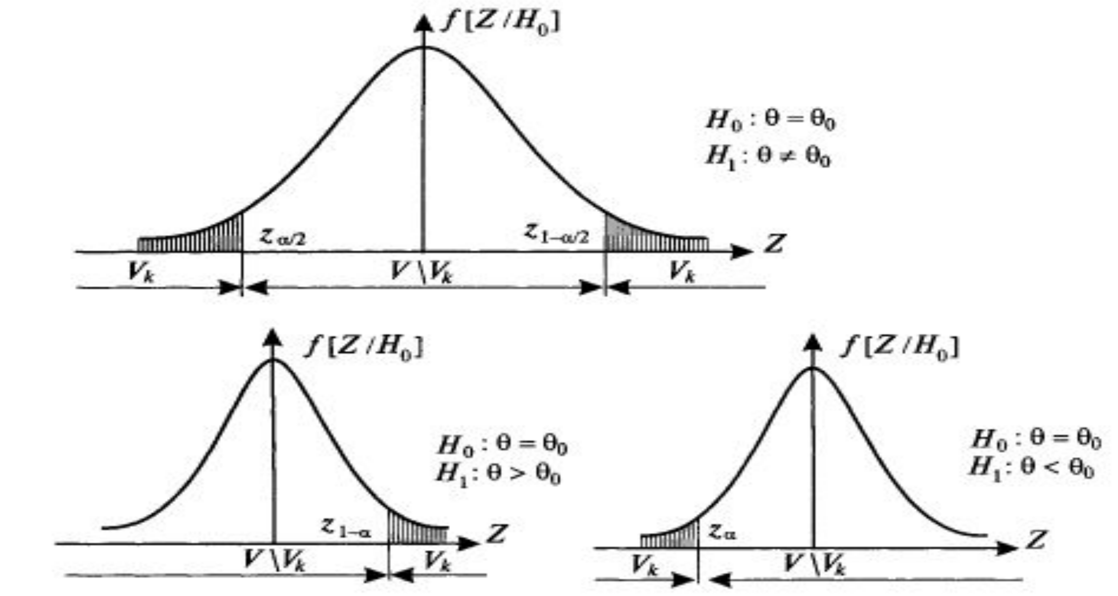
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Пусть ***V*** – множество значений статистики ***Z***, а V_k – такое подмножество, что при условии истинности гипотезы H_0, вероятность попадания статистики критерия в V_k равна ***α***: $P[Z \in V_k | H_0] = \alpha$. ***α*** – уровень значимости (малое число). Пусть $z_в$ – выборочное значение статистики ***Z***, вычисленное по выборке. Формулировка критерия: отклонить гипотезу H_0, если $z_в \in V_k$ (попадание в **критическую область**); принять гипотезу H_0, если $z_в \in V \setminus V_k$ (попадание в **область принятия гипотезы**).

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Уровень значимости ***α*** определяет размер критической области ***V_k***. Положение критической области на множестве значений статистики ***Z*** зависит от формулировки альтернативной гипотезы ***H_1***. Если проверяется гипотеза ***H_0: $\theta = \theta_0$***, а альтернативная гипотеза ***H_1***: ***$\theta > \theta_0$*** (***$\theta < \theta_0$***), то критическая область размещается на правом (левом) хвосте плотности распределения статистики ***Z***, то есть имеет вид неравенства: $Z > z_{1-\alpha}$ $(Z < z_\alpha)$, где $z_{1-\alpha}$ и z_α - квантили распределения статистики ***Z***, вычисленные при условии истинности ***H_0***.

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ



Размещение критической области при различных альтернативных гипотезах

Если альтернативная гипотеза $\boldsymbol{H_1: \theta \neq \theta_0}$, критическая область размещается на обоих хвостах плотности распределения ***Z***, то есть определяется неравенствами: $Z < z_{\frac{\alpha}{2}}$ и $Z > z_{1-\frac{\alpha}{2}}$. В этом случае критерий – двусторонний.

При проверке гипотез необходимо:

1. Сформулировать гипотезы $\boldsymbol{H_0}$ и $\boldsymbol{H_1}$, назначить уровень значимости ***α.***
2. Выбрать статистику ***Z*** критерия для проверки $\boldsymbol{H_0}$.
3. Определить критическую область $\boldsymbol{V_k}$, при условии, что верна $\boldsymbol{H_0}$.
4. Вычислить выборочное значение статистики $\boldsymbol{z_в}$ критерия.

$z_в \in V_k$ $z_в \in V \setminus V_k$

Сравнение двух дисперсий нормальных генеральных совокупностей

При заданном уровне значимости ***α*** по двум выборкам объемов n_1 и n_1 проверить нулевую гипотезу $H_0 : D(X) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе
$H_1 : D(X) > D(Y)$. В качестве критерия выбирается отношение большей
дисперсии к меньшей $F = \frac{s_Б^2}{s_М^2}$ ~$F(n_1 - 1, n_2 - 1)$, имеющей распределение
Фишера (n_1 – объем выборки с большей дисперсией). Критическая область – односторонняя и определяется по правилу: $F > F_{1-\alpha}(n_1 - 1, n_2 - 1)$
, где $F_{1-\alpha}(n_1 - 1, n_2 - 1)$ - квантиль распределения Фишера. Введем критическое число $F_{кр} = F_{1-\alpha}(n_1 - 1, n_2 - 1)$, отделяющее критическую

Сравнение выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности

При заданном уровне значимости ***α*** по выборке объемом ***n*** проверить нулевую гипотезу $H_0 : D(X) = \sigma_0^2$ о равенстве генеральной дисперсии $D(X)$ гипотетическому значению σ_0^2 при конкурирующей гипотезе $H_1 : D(X) > \sigma_0^2$. В качестве критерия выбирается статистика

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

имеющая распределение χ^2 с $n-1$ степенью свободы. Наблюдаемое значение критерия вычисляется по выборке

$$\chi^2_{набл} = \frac{(n-1)s^2}{\sigma_0^2}$$

Критическая область правосторонняя и определяется по правилу $\chi^2 > \chi^2_{1-\alpha}(n-1)$, где $\chi^2_{1-\alpha}(n-1)$ - квантиль распределения χ^2 с $n-1$ степенью свободы порядка $1-\alpha$. Введем критическое число $\chi^2_{кр} = \chi^2_{1-\alpha}(n-1)$.

Если $\chi^2_{набл} < \chi^2_{кр}$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{набл} > \chi^2_{кр}$ – нулевая гипотеза отвергается.

Сравнение двух средних генеральных совокупностей, дисперсии которых известны

При заданном уровне значимости α по двум выборкам объемами n и m проверить нулевую гипотезу $H_0: M(X)=M(Y)$ о равенстве математических ожиданий двух нормальных генеральных совокупностей с известными дисперсиями при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$. В качестве критерия выбирается статистика $z=\dfrac{\bar{x}-\bar{y}}{\sqrt{\dfrac{D(X)}{n}+\dfrac{D(Y)}{m}}}$ ~ $N(0,1)$.

Наблюдаемое значение $z_{набл}$ критерия вычисляется по выборкам. Критическая область - двусторонняя и определяется условиями $z<u_{\frac{\alpha}{2}}$ или $z>u_{1-\frac{\alpha}{2}}$, где u_α - квантиль нормального нормированного распределения порядка α. Учитывая, что $u_{\frac{\alpha}{2}}=-u_{1-\frac{\alpha}{2}}$ область принятия гипотезы определяется неравенством $|z|<u_{1-\frac{\alpha}{2}}$. Пусть $z_{кр}=u_{1-\frac{\alpha}{2}}$.

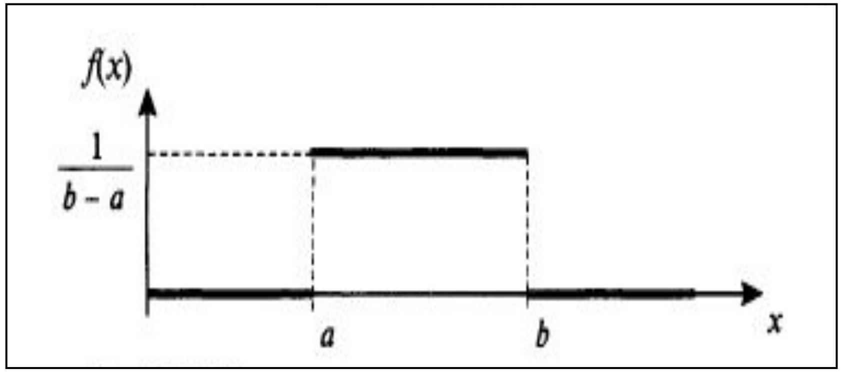
$$\left|z_{набл}\right|<z_{кр}$$

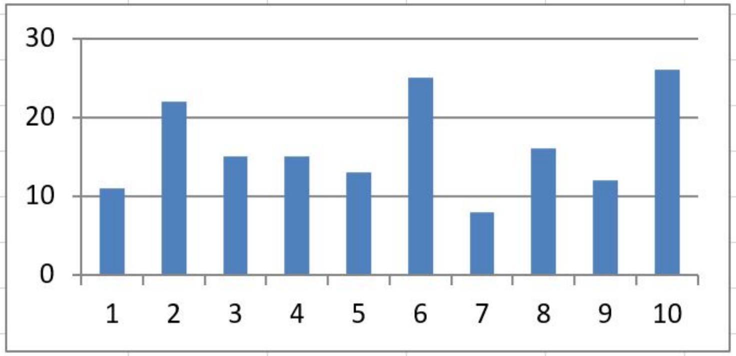
КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА

Пусть по выборке случайной величины X объема n необходимо проверить гипотезу H_0 о том, что X имеет предполагаемый закон распределения. Пусть построена гистограмма для заданной выборки случайной величины X объема n с частичными интервалами Δ_j одинакового размера h. По форме гистограммы делается предположение о типе закона распределения и по выборке вычисляются точечные оценки его параметров. Используя предполагаемый закон распределения случайной величины X, находим вероятности p_j того, что значение X принадлежит частичным интервалам Δ_j, то есть $p_j = P[X \in \Delta_j]$,
где $j = 1, \dots, N$. Величины $m_j = np_j$ равны теоретическим частотам попадания случайной величины X в каждый частичный интервал. Выборочное (наблюдаемое) значение статистики критерия вычисляется

$$\chi^2_{набл} = \sum_{j=1}^{N} \frac{(n_j - m_j)^2}{m_j}$$

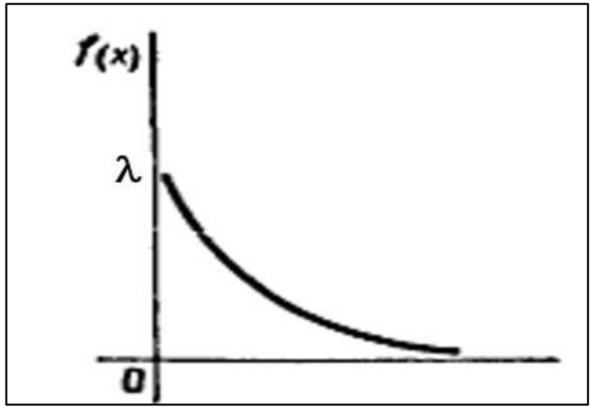
КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА

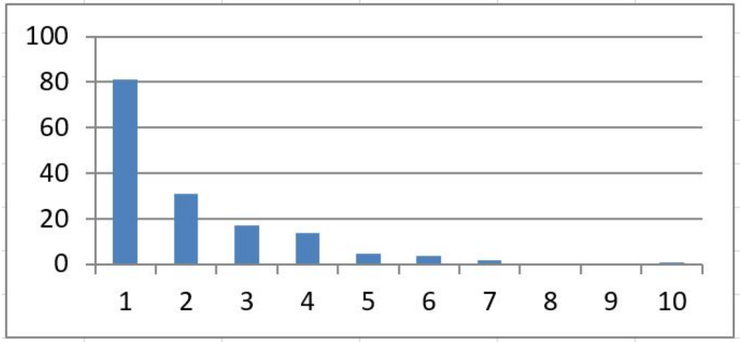




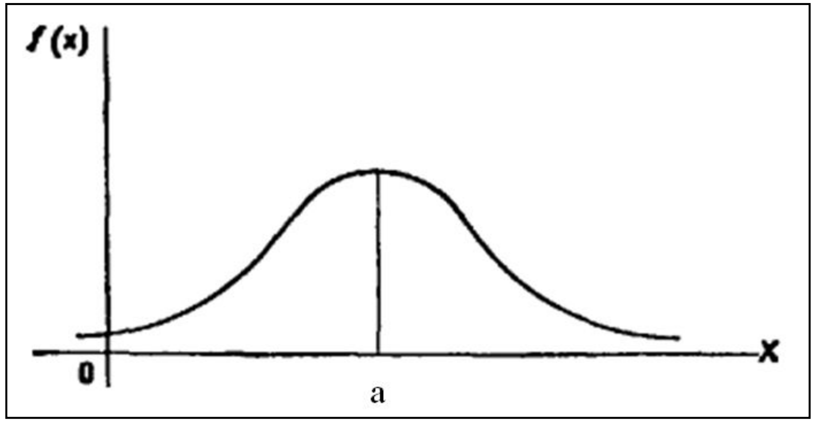
Законы распределения:

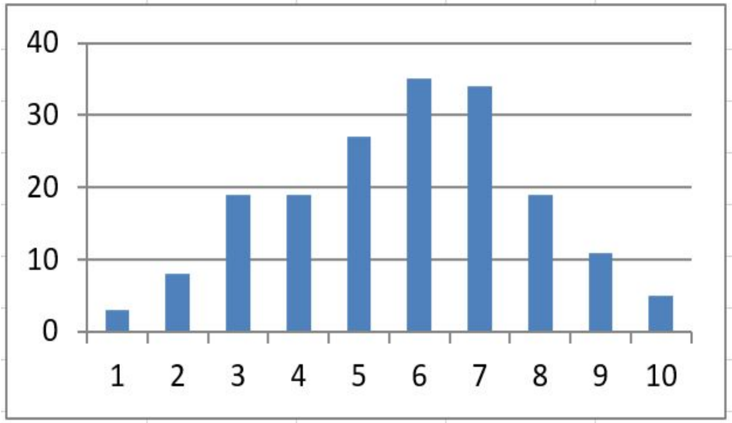
а) равномерный;





б) экспоненциальный;





в) нормальный

КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА

Гипотеза H_0 согласуется с результатами наблюдений на уровне значимости α, если $\chi^2_{набл} < \chi^2_{1-\alpha}(N - m_p - 1)$, где $\chi^2_{1-\alpha}(N - m_p - 1)$ - квантиль порядка

$1 - \alpha$ распределения χ^2 с $N - m_p - 1$ степенями свободы, где m_p – число неизвестных параметров распределения, оцениваемых по выборке. Если же $\chi^2_{набл} > \chi^2_{1-\alpha}(N - m_p - 1)$, то гипотеза H_0 отклоняется.

При заданном уровне значимости α используется следующий расчет.

1) Для каждого частичного интервала определяются средние точки y_j частичных интервалов.

$$m_j = \frac{hn}{x_{max} - x_{min}}, \; j = 1, \ldots, N$$

2) Вычисляются теоретические частоты предполагаемого

распределения: а) для равномерного закона

КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА

б) для нормального закона $m_j = \frac{hn}{s}\varphi(u_i),\ j=1,\dots,N$, где $u_j = \frac{y_j - \bar{x}_в}{s}$,

$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ - плотность нормального нормированного распределения;

в) для показательного закона $m_j = n\left(e^{-\lambda_в x_j} - e^{-\lambda_в x_{j+1}}\right),\ j=1,\dots,N$, где $\lambda_в = \frac{1}{\bar{x}_в}$ - выборочный параметр показательного закона распределения:

.

Находится наблюдаемое значение критерия ***χ²***: $\chi^2_{набл} = \sum_{j=1}^{N}\frac{(n_j - m_j)^2}{m_j}$.

3) Находится критическое значение критерия ***χ²*** по уровню значимости **α** и числу степеней свободы ***k = N - m_p - 1***, где ***m_p*** - число параметров распределения, равное ***m_p = 2*** для нормального и равномерного законов ***m_p = 1*** для показательного распределения: $\chi^2_{крит} = \chi^2_{1-\alpha}(N - m_p - 1)$

.

4) Если $\chi^2_{набл} < \chi^2_{крит}$ - то выдвинутая гипотеза о типе закона

МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

Пусть есть результаты некоторого эксперимента **(x_i, y_i), $i = 1, \ldots, n$**. Надо подобрать (“подогнать”) функцию **$y = f(x) = ax + b$**, наилучшим образом описывающую экспериментальные точки **(x_i, y_i)**. Критерием качества “подгонки” кривой выберем минимум функционала:

$$F = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \to min$$

Данный принцип называется методом наименьших квадратов и является одним из самых распространенных методов подбора функциональной зависимости, наилучшим образом описывающим эмпирические данные.

Используя необходимые условия экстремума:

$$\frac{\partial F}{\partial a} = 0, \frac{\partial F}{\partial b} = 0$$

Получим:

$$b^* = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad a^* = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b^*}{n}\sum_{i=1}^{n} x_i$$

МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

Введем обозначения: $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \vec{s} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

Тогда $\vec{y}^* = a\vec{s} + b\vec{x}$ - вектор “подогнанных” с помощью регрессионного уравнения значений **y**, лежащий в плоскости (гиперплоскости) векторов $\vec{x}$ и $\vec{s}$; $\vec{e} = \vec{y} - \vec{y}^*$ - вектор разности между выборочными значениями $\vec{y}$ и “подогнанными” $\vec{y}^*$ с помощью функции ***y = f(x)***. Наилучшее приближение $\vec{y}^*$ к $\vec{y}$ будет, если вектор $\vec{e}$ перпендикулярен обоим векторам $\vec{x}$ и $\vec{s}$:

Это условие эквивалентно равенству нулю двух скалярных произведений: $\vec{s}^T \vec{e} = 0$ и $\vec{x}^T \vec{e} = 0$. Обозначим:

$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \vec{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}$. $\vec{e} = \vec{y} - X\vec{\beta}$

МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

Условие ортогональности вектора $\vec{e}$ к плоскости векторов $\vec{x}$ $\vec{s}$ и : $X^T \vec{e} = 0$. После преобразований: $\vec{\beta} = \left(X^T X\right)^{-1} X^T \vec{y}$. Это уравнение для коэффициентов регрессионного уравнения ***a*** и ***b*** совпадает с формулами по методу наименьших квадратов.

Пусть реальное (теоретическое) уравнение зависимости ***y*** от ***x*** имеет вид: ***y = a + bx + ε,*** где ***x*** **-** неслучайная (детерминированная) величина*,* ***y*** и ***ε*** - случайные величины. Считаем,, что ***ε ~ N*(*0, σ*),** где ***σ²*** **-** дисперсия ошибок. По данным наблюдений следует найти оценки a^* и b^* для коэффициентов уравнения регрессии и s^2 - для дисперсии ошибок σ^2.

Имеет место **теорема Гаусса-Маркова**, утверждающая, что при использовании метода наименьших квадратов для определения оценок a^* и b^* коэффициентов уравнения регрессии по формуле

$$\vec{\beta} = \left(X^T X\right)^{-1} X^T \vec{y}$$

МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

В соответствии с теоремой Гаусса-Маркова оценки $a^* \sim N(a, \sigma_a)$, $b^* \sim N(b, \sigma_b)$. Имеют место следующие формулы для дисперсий оценок:

$$D(a^*) = \sigma_a^2 = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2\right)}, \quad D(b^*) = \sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2}.$$

Оценки дисперсий:

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - a^* - b^* x_i\right)^2}{n-2}, \quad s_a^2 = s^2 \frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2\right)}, \quad s_b^2 = \frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2}.$$

При проверке гипотезы $H_0: b = b_0$ с альтернативой $H_1: b \neq b_0$ используется статистика $t = \frac{b^* - b_0}{s_b}$, имеющая распределение Стьюдента с $(n - 2)$ степенями свободы. Доверительный интервал для

$$b \in \left(b^* - t_\gamma s_b ; b^* + t_\gamma s_b\right)$$

МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

Доверительный интервал для коэффициента ***a*** с надежностью **γ** имеет вид: $\left(a^* - t_\gamma s_a ; a^* + t_\gamma s_a\right)$.

Коэффициентом детерминации (или долей объясненной дисперсии) называется величина: $R^2 = 1 - \frac{ESS}{TSS} = \frac{TSS - ESS}{TSS} = \frac{RSS}{TSS}$, где $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $ESS = \sum_{i=1}^{n}(y_i - y_i^*)^2$, $RSS = \sum_{i=1}^{n}(y_i^* - \bar{y})^2$. Очевидно, что $0 \le R^2 \le 1$. Если, $R^2 = 0$, то регрессия не улучшает качество прогноза **y** по сравнению с тривиальным: $y = \bar{y}$. Если $R^2 = 1$, то это означает точную “подгонку”: все наблюдаемые точки лежат на регрессионной кривой.

Проверяется гипотеза $H_0: b = b_0$ с альтернативой $H_1: b \neq b_0$. Статистика

$$F = \frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{s^2} \sim F(1; n-2).$$

Если при заданном уровне значимости ***α*** критическое значение $F_{кр}(\alpha; 1; n - 2)$ меньше наблюдаемого

МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Рассматривается следующее уравнение модели:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_k x^k+\varepsilon,$$

где $\vec{\beta}=\left(\beta_1\ \beta_2\ \ldots\ \beta_k\right)^T$ - вектор коэффициентов модели, $x_1, x_2, \ldots, x_k$ - регрессоры (независимые переменные), ε - случайная компонента, $\varepsilon \sim N(0, \sigma)$, где σ^2 - дисперсия ошибок. По результатам экспериментов $(x_i; y_i)$, $i=1,\ldots,n$ надо определить значения коэффициентов β.

$$\vec{y}=\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{pmatrix},\quad \vec{y}^*=\begin{pmatrix} y_1^* \\ y_2^* \\ \ldots \\ y_k^* \end{pmatrix}$$

Пусть $y_i^*=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\ldots+\beta_k x_i^k$, где $\vec{y}^*=\left(y_1^+\ \ y_2^+\ \ \ldots\ \ y_k^+\right)^T$ - прогноз значений объясняемой переменной. Пусть $\vec{\beta}$ - вектор прогнозируемых значений. Тогда $\vec{e}=\vec{y}-\vec{y}^*$ - вектор разности между выборочными и прогнозируемыми значениями **y**. Будем

$$\left|\vec{e}\right|^2=\vec{e}^T\vec{e}=\left(\vec{y}-\vec{y}^*\right)^T\left(\vec{y}-\vec{y}^*\right)$$

МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Для оценки коэффициентов множественной регрессии необходимо использовать формулу, совпадающую со случаем парной регрессии:

$$\vec{\beta}^{\,*} = \left(X^T X\right)^{-1} X^T \vec{y}$$

. Формула для оценки дисперсии ошибок σ^2 :

$$s^2 = \frac{\vec{e}^{\,T}\,\vec{e}}{n-(k+1)}$$

. Коэффициент детерминации определяется так же, как и для парной регрессии:

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}\left(y_i^* - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \frac{\vec{\beta}^{\,T} X^T \vec{y} - n\left(\bar{y}\right)^2}{\vec{y}^{\,T}\,\vec{y} - n\left(\bar{y}\right)^2}$$

Для определения дисперсий оценок коэффициентов регрессии вычисляется матрица ковариаций:

$$V\left(\vec{\beta}^{\,*}\right) = s^2\left(X^T X\right)^{-1}$$

. В качестве дисперсий оценок:

$$D\left(\vec{\beta}^{\,*}\right) = \left(s_1^2 \quad s_2^2 \quad \dots \quad s_k^2\right)^T$$

берутся диагональные элементы q_{ii} матрицы

$$V\left(\vec{\beta}^{\,*}\right) : s_i^2 = s^2 q_{ii}$$

. Доверительный интервал

$$\beta_i \in \left(\beta_i^* - t_\gamma s_i ;\ \beta_i^* + t_\gamma s_i\right)$$

МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Для проверки значимости коэффициентов регрессии $\boldsymbol{\beta_i}$ можно использовать гипотезу $\boldsymbol{H_0: \beta_i = 0}$ при альтернативе $\boldsymbol{H_1: \beta_i \neq 0}$. Двусторонняя критическая область формируется из условия $t = \left|\frac{\beta_i^*}{s_i}\right| > t_{кр}$,

где $\boldsymbol{t_{кр}(\alpha; n - (k + 1))}$ - критическая точка распределения Стьюдента для уровня значимости $\boldsymbol{\alpha}$ и числа степеней свободы $\boldsymbol{n - (k + 1)}$.

Для оценки значимости уравнения регрессии, то есть проверки факта улучшения прогноза значения $\boldsymbol{y}$ по сравнению с тривиальным: $y^* = \bar{y}$ используется критерий Фишера. Гипотеза $\boldsymbol{H_0: \beta_1 = \beta_2 = ... = \beta_k = 0}$ (свободный член $\boldsymbol{\beta_0}$ - не учитывается) проверяется с использованием

статистики $F = \frac{\left(\sum_{i=1}^{n}(y_i^* - \bar{y})^2\right)/k}{s^2}$ $\boldsymbol{\sim F(k; n - (k + 1))}$. Если наблюдаемое значение $\boldsymbol{F_{набл}}$ больше критического $\boldsymbol{F_{кр}(\alpha; k; n - (k + 1))}$, то гипотеза

"ПОДГОНКА" КРИВОЙ

Пусть данные некоторого эксперимента представлены в виде таблицы значений переменных ***x*** и ***y***. Ставится задача об отыскании аналитической зависимости между ***x*** и ***y***, то есть некоторой формулы

y = f(x), явным образом выражающей ***y***, как функцию ***x***. Естественно требовать, чтобы график искомой функции ***y = f(x)*** не слишком уклонялся от экспериментальных точек (x_i, y_i). Эту задачу называют сглаживанием" экспериментальных данных или "подгонкой" кривой. Для решения можно использовать метод наименьших квадратов (МНК). Согласно МНК указывается вид эмпирической формулы $y = Q(x, a_0, a_1, \dots, a_n)$, где $a_0, a_1, \dots, a_n$ – числовые параметры. Наилучшими значениями параметров $a_0, a_1, \dots, a_n$ считаются те, для которых сумма квадратов уклонений функции

$$S(a_0, a_1, \dots, a_n) = \sum_{i=1}^{m} (Q(x_i, a_0, a_1, \dots, a_n) - y_i)^2 \to \min$$

"ПОДГОНКА" КРИВОЙ

Используя необходимые условия экстремума функции нескольких переменных, получаем систему уравнений для определения параметров

$\frac{\partial S}{\partial a_0}=0, \frac{\partial S}{\partial a_1}=0, \ldots, \frac{\partial S}{\partial a_n}=0$. Если система имеет единственное решение, то оно

является искомым и аналитическая зависимость между экспериментальными данными определяется формулой: ***y = f(x).*** В общем случае система уравнений для определения оптимальных значений параметров нелинейна.

Пусть для переменных ***x*** и ***y*** соответствующие значения экспериментальных данных (***x_i, y_i***) не располагаются вблизи прямой. Тогда можно выбрать новые переменные ***X = φ (x, y)*** и ***Y = ψ (x, y)*** так, чтобы преобразованные экспериментальные данные ***X_i = φ (x_i, y_i)*** и

"ПОДГОНКА" КРИВОЙ

Числа k и b определяются по встроенной функции **ЛИНЕЙН**, где вместо x_i и y_i подставляют соответствующие значения X_i и Y_i. Функциональная зависимость $y = f(x)$ определена неявно уравнением

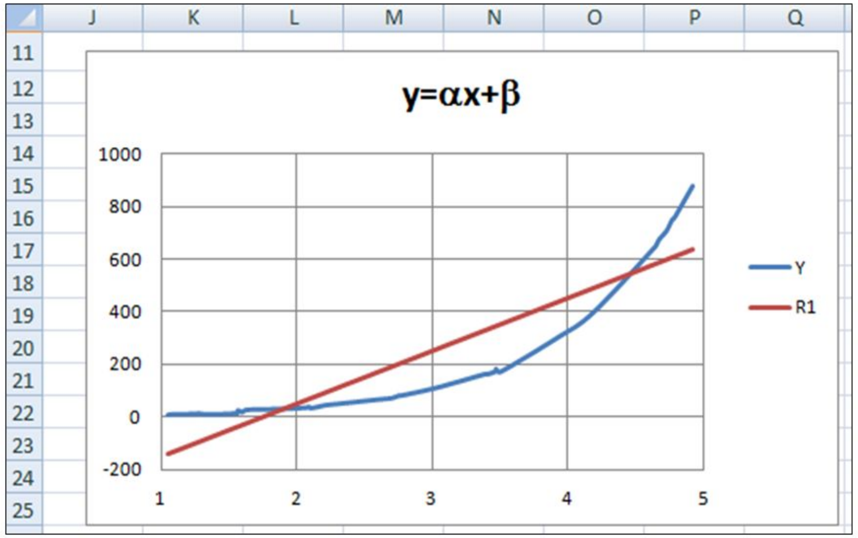
$\psi(x, y) = k\phi(x, y) + b$, которое разрешимо относительно y в частных случаях.

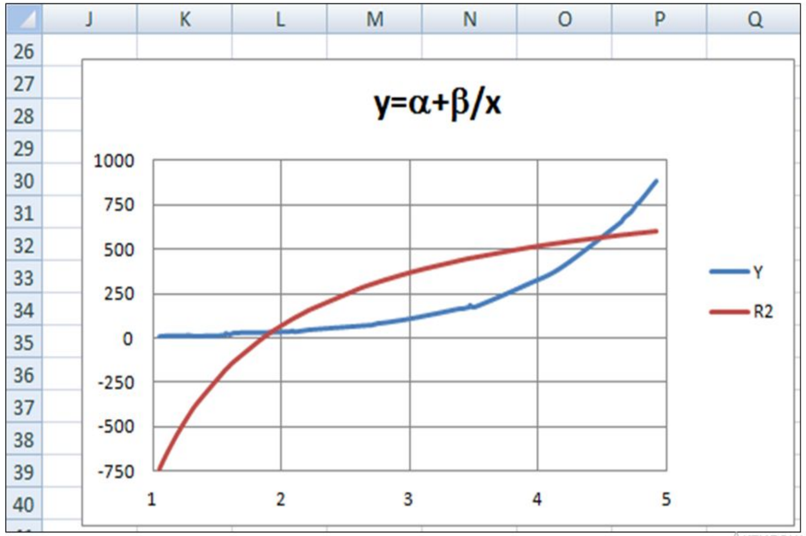
Рассмотрим семь вариантов преобразования переменных, в которых возможно явное выражение переменной y из уравнения $\psi(x, y) = k\phi(x, y) + b$. В таблице приведены формулы преобразования переменных, явная эмпирическая формула $y = f(x)$, зависящая от двух параметров α и β, выражение этих параметров через коэффициенты, полученные с помощью МНК.

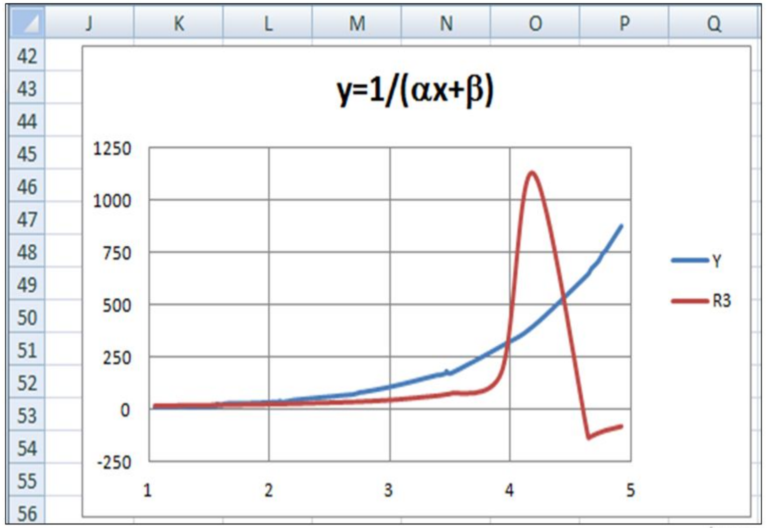
"ПОДГОНКА" КРИВОЙ

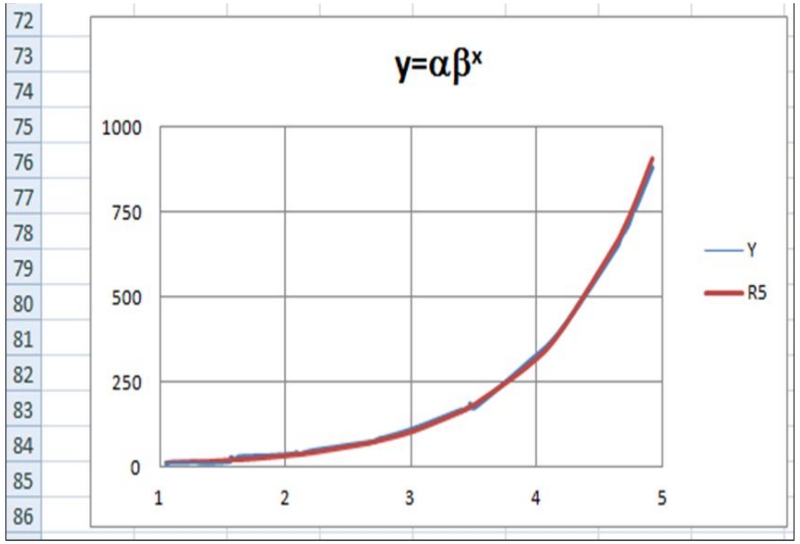
№	Эмпирическая формула	Выравнивание данных (преобразование переменных)
1.	$y = \alpha x + \beta,\ \alpha = k,\ \beta = b$	$X = x,\ Y = y$
2.	$y = \alpha + \dfrac{\beta}{x},\ \alpha = k,\ \beta = b$	$X = x,\ Y_1 = xy$
3.	$y = \dfrac{1}{\alpha x + \beta},\ \alpha = k,\ \beta = b$	$X = x,\ Y_2 = \dfrac{1}{y}$
4.	$y = \dfrac{x}{\alpha x + \beta},\ \alpha = k,\ \beta = b$	$X = x,\ Y_3 = \dfrac{x}{y}$
5.	$y = \alpha \beta^{x},\ \alpha = e^{b},\ \beta = e^{k}$	$X = x,\ Y_4 = \ln y$
6.	$y = \alpha \ln x + \beta,\ \alpha = k,\ \beta = b$	$X_1 = \ln x,\ Y = y$
7.	$y = \alpha x^{\beta},\ \alpha = e^{b},\ \beta = k$	$X_1 = \ln x,\ Y_4 = \ln y$

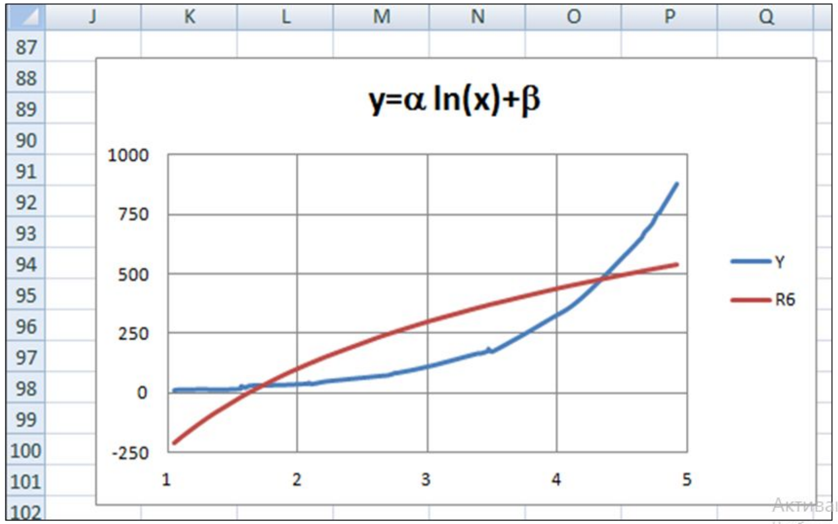
"ПОДГОНКА" КРИВОЙ

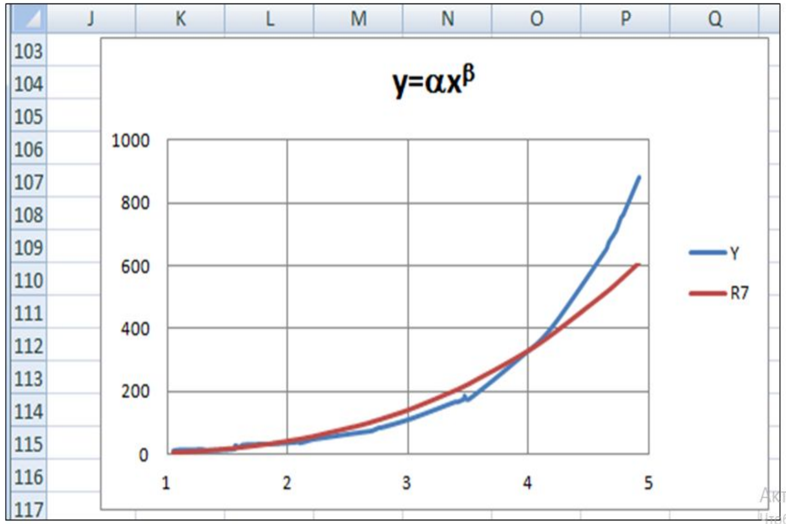












Экспериментальные данные наилучшим образом аппроксимируются регрессионной кривой $y = \alpha\beta^x = 3{,}584 \cdot 3{,}083^x$.