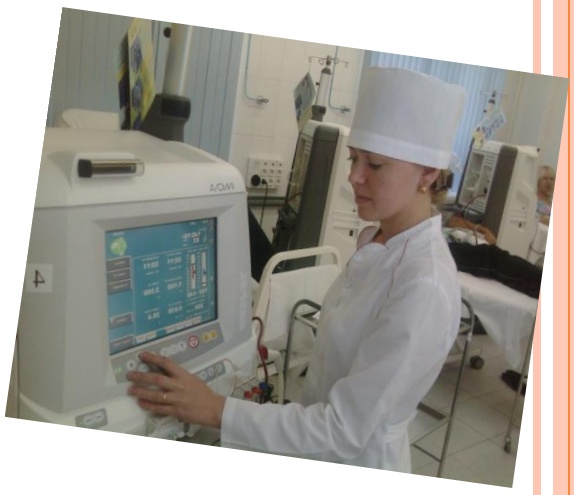


Лекция 8
Примеры задач анализа данных.
Методы подготовки данных к анализу.

Составитель: доц. Космачева И.М.

СТАТИСТИЧЕСКИЕ ПАКЕТЫ

Статистический пакет - программный продукт, предназначенный для статистической обработки данных. Существуют специализированные статистические пакеты и другие пригодные для проведения статистических расчетов приложения.

- ***Зарубежные:*** ***STATGRAPHICS, SPSS, SYSTAT, BMDP,SAS, CSS, STATISTICA, S-plus и др.,***
- ***Отечественные:*** ***STADIA, ЭВРИСТА, МЕЗОЗАВР, ОЛИМП: Стат-Эксперт, Статистик-Консультант, САНИ, КЛАСС-МАСТЕР, Deductor Academic (basegroup.ru) и др.***
- ***Mathcad, EXCEL***

Примеры анализа данных

Ошибка выборки - расхождение между характеристиками выборочной и генеральной совокупностей.

Наибольшая из возможных ошибок выборки Δ называется **предельной ошибкой выборки,** которая рассчитывается по формуле:

$$\Delta = t\mu,$$

где t — коэффициент доверия (коэффициент Стьюдента); $\mu = \sqrt{\frac{\sigma^2}{n}}$;

σ^2 — дисперсия генеральной совокупности; n — объем выборки.

В случае неизвестного σ используют приближенное значение $\mu = \sqrt{\frac{S^2}{n}}$.

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

где S^2- оценка дисперсии σ^2, вычисляемая по выборке x_1 x_2, x_n.

Примеры анализа данных

Доверительный коэффициент ***t*** находится из таблицы квантилей **нормального распределения при заданной надежности γ.** При стандартных значениях надежности **γ = 0,95 и γ = 0,99** соответствующие доверительные коэффициенты *t* равны $\mathbf{t_{0,95} = 1,96; t_{0,99} = 2,58.}$

На формулах расчета предельной ошибки выборки основан способ определения численности выборки, обеспечивающей заданную точность оценки. Тогда:

$$\Delta = t\frac{\sigma}{\sqrt{n}} \text{ или } \Delta = t\frac{S}{\sqrt{n}},$$

следует:

$$n = \frac{t^2\sigma^2}{\Delta^2} \text{ или } n = \frac{t^2S^2}{\Delta^2}.$$

Задача 1

Исходные данные

При изучении средней длительности пребывания больных в стационаре получены следующие данные: M = 20 дней, σ = 1,63 дня.

Задание

Определить необходимый объем выборки для получения достоверных результатов при изучении средней длительности пребывания больных в стационаре при заданном доверительном коэффициенте t_Y= **3** (**надежность γ = 0,9973**) и предельной ошибке **Δ = 0,5 дня**.

Расчет необходимого объема выборки для изучения средней длительности пребывания больных в стационаре:

$$n = \frac{t^2\sigma^2}{\Delta^2} = \frac{3^2 \cdot 1{,}63^2}{0{,}5^2} = \frac{9 \cdot 2{,}66}{0{,}25} = \frac{23{,}94}{0{,}25} = 95{,}8.$$

Для получения показателя средней длительности пребывания больных в стационаре с заданной точностью 0,5 дня необходимый объем выборки должен составить 96 больных.

Задача 2

Интервальные оценки математического ожидания нормального распределения при известном σ

$$P\left(\overline{X}_в - \frac{x_\gamma \sigma}{\sqrt{n}} < a < \overline{X}_в + \frac{x_\gamma \sigma}{\sqrt{n}}\right) = \gamma. \qquad \Phi(x_\gamma) = \frac{\gamma}{2},$$

Пример: Найти доверительный интервал для оценки математического ожидания, если σ=3, n = 36 и γ=0,95 . x_γ = 1,96 (определяем по таблице значений функции Лапласа). Тогда ε = 1,96*3/6 = 0,98. Таким образом, с надежностью 95 % оцениваемый параметр принадлежит доверительному интервалу

$$(x_в - 0.98; x_в + 0.98)$$

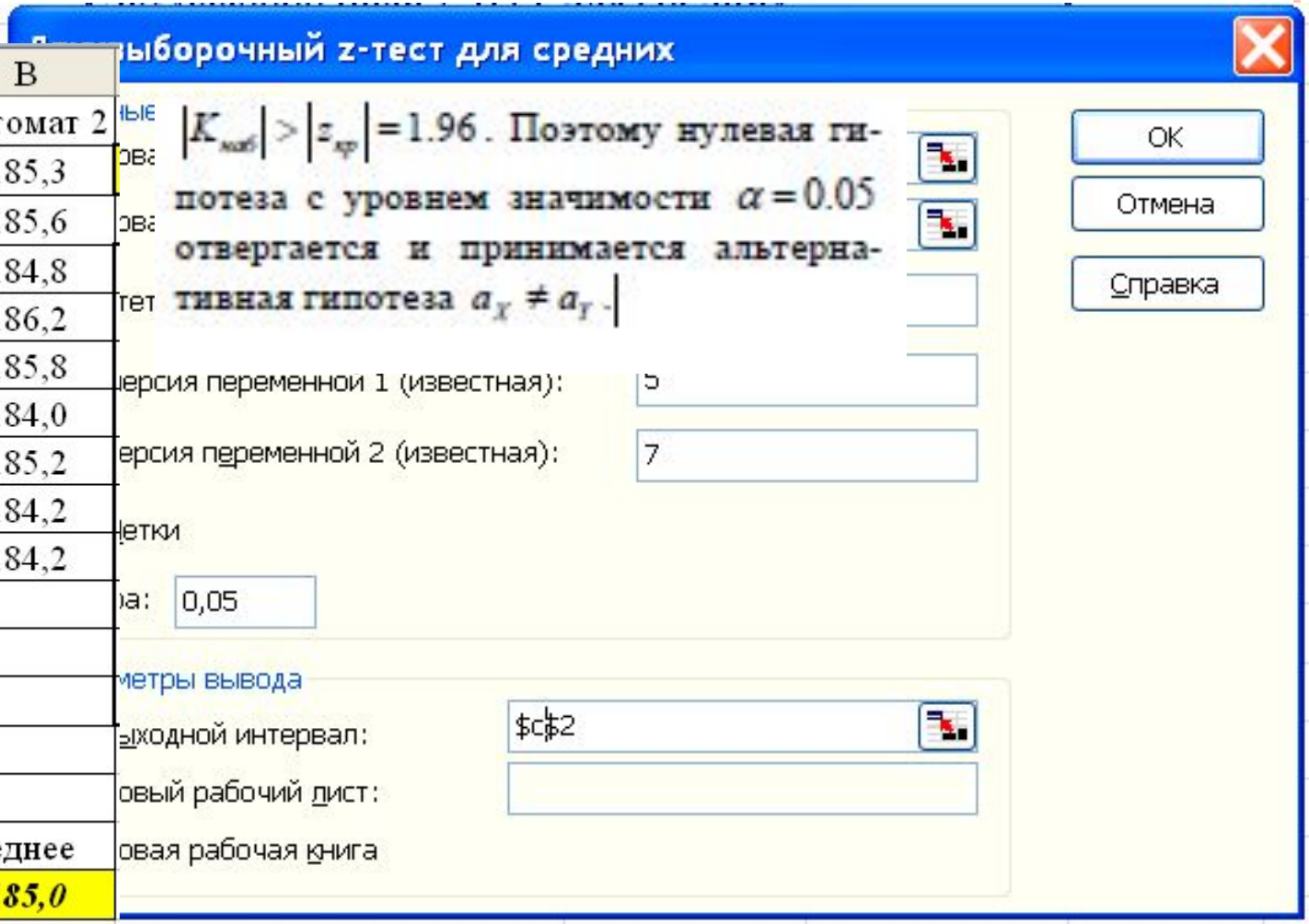
$$\left[\overline{X}_в - \frac{x_\gamma \sigma}{\sqrt{n}}, \overline{X}_в + \frac{x_\gamma \sigma}{\sqrt{n}}\right].$$

$$x_\gamma = \text{НОРМСТОБР}((\gamma + 1)/2)$$

Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.

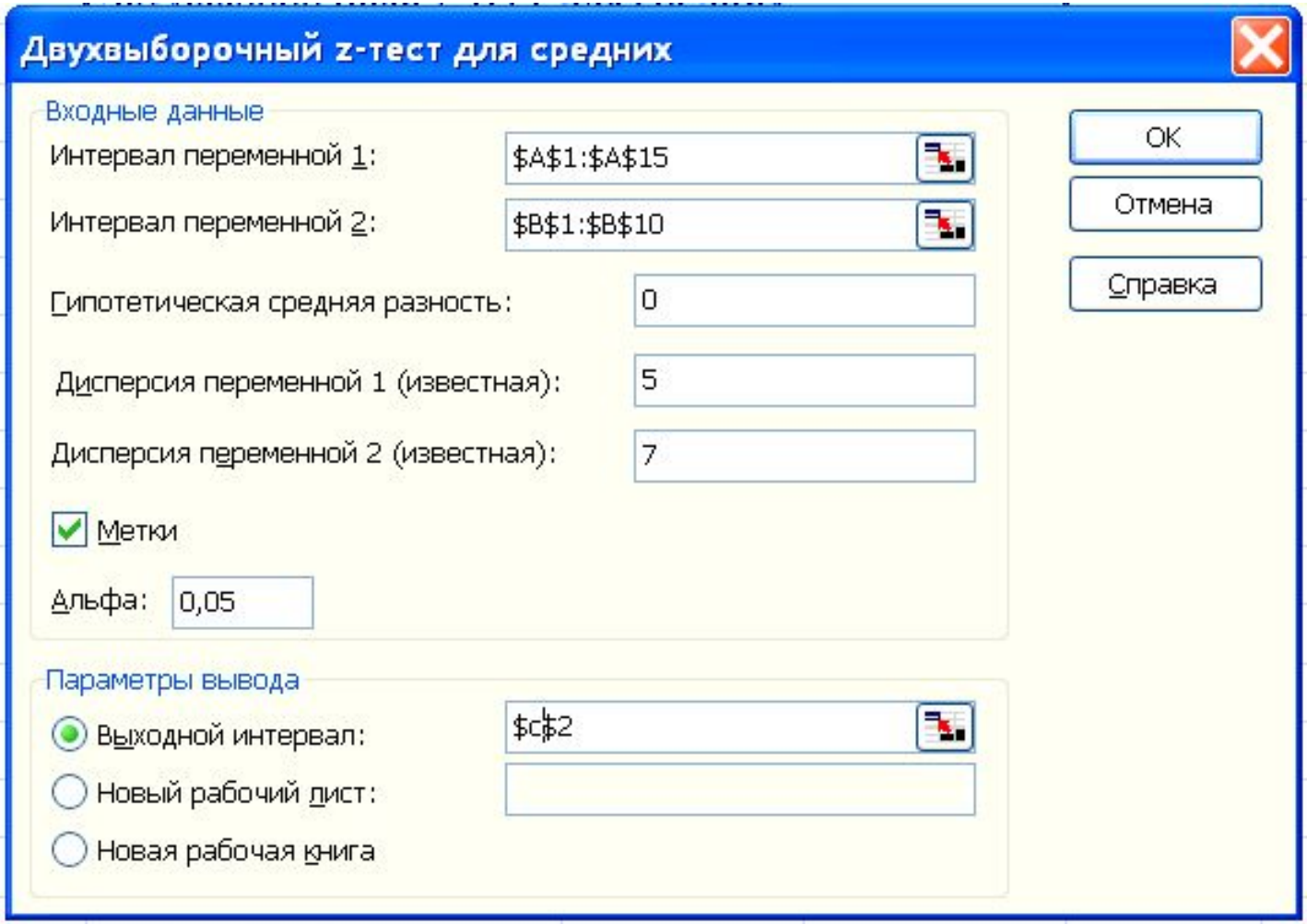
Необходимо проверить нулевую гипотезу $H_0: a_X = a_Y$ при альтернативной гипотезе $H_1: a_X \neq a_Y$.

	A	B
1	Автомат1	Автомат 2
2	182,3	185,3
3	183,0	185,6
4	181,8	184,8
5	181,4	186,2
6	181,8	185,8
7	181,6	184,0
8	183,2	185,2
9	182,4	184,2
10	182,5	184,2
11	179,7	
12	179,9	
13	181,9	
14	182,8	
15	183,4	
16	Среднее	Среднее
17	***182,0***	***185,0***



Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.

Необходимо проверить нулевую гипотезу $H_0: a_X = a_Y$ при альтернативной гипотезе $H_1: a_X \neq a_Y$.



Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.

Необходимо проверить нулевую гипотезу $H_0: a_X = a_Y$ при альтернативной гипотезе $H_1: a_X \neq a_Y$.

Двухвыборочный z-тест для средних		
	Автомат1	*Автомат 2*
$K_{наб} = z = -2.867$. Это значение попадает в критическую область $\lvert K_{наб} \rvert > \lvert z_{кр} \rvert = 1.96$. Поэтому нулевая гипотеза с уровнем значимости $\alpha = 0.05$ отвергается и принимается альтернативная гипотеза $a_X \neq a_Y$.	181,979	185,03
	5,000	7,00
	14,000	9,00
	0,000	
	-2,867	
	0,002	
	1,645	
	0,004	
	1,960	

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Вычисление величины $x_\gamma \sigma / \sqrt{n}$ **осуществляется с помощью функции ДОВЕРИТ:**

$$\Delta_{\bar{X}_s} = x_\gamma \sigma / \sqrt{n} = \text{ДОВЕРИТ}(\alpha; \sigma; n),$$

	F	G	H	I
1		Уровень значимости		0,05
2		Интервал	Левая граница	Правая граница
3		Матожидание		
4		Дисперсия		

=СРЗНАЧ(A1:A25)-ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)
=СРЗНАЧ(A1:A25)+ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)

Задача 3

Интервальные оценки математического ожидания нормального распределения при неизвестном σ

$$\left(\overline{X}_в - \frac{t(\gamma,n)S}{\sqrt{n}}, \overline{X}_в + \frac{t(\gamma,n)S}{\sqrt{n}}\right),$$

а точность интервальной оценки определить соотношением

$$\delta = \frac{t(\gamma,n)}{\sqrt{n}} S .$$

Пример . По выборке объема $n = 9$ из нормально распределенной генеральной совокупности найдены значения $x_в = 1.5$ и среднеквадратическое отклонение $s_в = 2$. Построить интервальную оценку для математического ожидания с надежностью $\gamma = 0.95$.

$$\delta = \frac{t(0.95,9)S}{\sqrt{n}} = \frac{2.31}{3} S = 0.77$$

$$\left(\overline{X}_в - 1.54, \overline{X}_в + 1.54\right)$$

Задача 3

Интервальные оценки математического ожидания нормального распределения при неизвестном σ

Вычисление величины $t(\gamma,n)$, входящей в доверительный интервал

$$\left[\overline{X}_в - \frac{t(\gamma,n)\cdot\sqrt{D_в}}{\sqrt{n-1}}, \overline{X}_в + \frac{t(\gamma,n)\cdot\sqrt{D_в}}{\sqrt{n-1}}\right],$$

осуществляют с использованием функции СТЬЮДРАСПОБР, обращение к которой имеет вид:

$$t(\gamma,n) = \text{СТЬЮДРАСПОБР}(\alpha;n),$$

Задача 4
Проверка независимости признаков.

- Значение выборочного коэффициента корреляции является оценкой «истинного» теоретического значения $\mathbf{r_{xy}}$ и отличается от него в силу различных случайных причин.
- Даже при очевидной независимости признаков, скорее всего, окажется $r_в$ *не равен 0.*
- Следует установить, отличие $r_в$ от нуля вызвано случайными причинами, связанными с выборкой (незначимо), или же оно принципиально, т.е. объясняется именно зависимостью признаков (значимо).
- Таким критерием является статистика, имеющая распределение Стьюдента

$$T = \frac{R_B\sqrt{n-2}}{\sqrt{1-R_B^2}},$$

Задача 4
Проверка независимости признаков.

Пример . Получена корреляционная таблица, составленная по выборке студентов возраста 20 - 22 лет.

СВ X – стаж курильщика (количество лет), СВ Y – жизненная емкость легких (ЖЕЛ) в мл.

При 77 наблюдениях, требуется определить, зависит ли в генеральных совокупностях значение показателя ЖЕЛ (Y) от стажа курильщика (X). Распределение случайных величин X и Y предполагается нормальным.

Решение

Используем критерий Стьюдента для проверки гипотезы H_0: $\mathbf{r_{xy}=0}$

В примере число наблюдений $n = 77$, выборочный коэффициент корреляции $r_B = -0{,}7535$.

$$T_{набл} = \frac{-0{,}7535\sqrt{77-2}}{\sqrt{1-(-0{,}7535)^2}} = -9{,}9255.$$

Выберем уровень значимости $\alpha = 0{,}01$ и по таблице критических значений распределения Стьюдента находим $t_{кр} = t_{кр}(0{,}01;\ 75) = 2{,}643$.

Так как $|Tнабл| > t_{кр}$, то при выбранном уровне значимости нулевую гипотезу отвергаем; следовательно, случайные величины X и Y зависимы.

Проверка независимости признаков.

- Квадрат коэффициента корреляции зависимой и независимой переменных представляет долю дисперсии зависимой переменной, обусловленной влиянием независимой переменной, и называется ***коэффициентом детерминации***. Коэффициент детерминации показывает, в какой степени изменчивость одной переменной обусловлена (детерминирована) влиянием другой переменной.
- Если обе переменные, между которыми изучается связь, представлены в порядковой шкале, или одна из них - в порядковой, а другая - в метрической, то применяются ранговые коэффициенты корреляции: **Спирмена или τ-Кенделла**. И тот, и другой коэффициент требует для своего применения **предварительного ранжирования обеих переменных.**
- Коэффициент ранговой корреляции целесообразно применять при наличии небольшого количества наблюдений.

КОЭФФИЦИЕНТ СПИРМЕНА

1. Нужно упорядочить данные по возрастанию и заменить реальные значения их рангами. *Рангом* значения называется его номер в упорядоченном ряду. Например, в ряду 1,4, 8, 8, 12 ранг числа 4 равен 2.
2. Если в ряду встретятся одинаковые значения, им следует присвоить один и тот же ранг, равный среднему занимаемых ими.
3. Затем, беря вместо самих значений их ранги, рассчитывают обычный коэффициент корреляции Пирсона. Это и будет коэффициент ранговой корреляции Спирмена.

КОЭФФИЦИЕНТ СПИРМЕНА

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

где d — разность рангов для каждого члена выборки.

$$r_s = 1 - \frac{6\left[(-1)^2 + (-1)^2 + 2^2 + 0^2 + 0{,}5^2 + (-0{,}5)^2 + 0^2 + 0^2 + 0^2\right]}{10^3 - 10} =$$

$$= 0{,}96.$$

Вычисление коэффициента ранговой корреляции Спирмена

Рост		Вес		
Значение, см	Ранг	Значение, г	Ранг	Разность рангов
31	1	7,7	2	–1
32	2	8,3	3	–1
33	3	7,6	1	2
34	4	9,1	4	0
35	5,5	9,6	5	0,5
35	5,5	9,9	6	–0,5
40	7	11,8	7	0
41	8	12,2	8	0
42	9	14,8	9	0
46	10	15,0	10	0

КОЭФФИЦИЕНТ КОНКОРДАЦИИ

	1	2	…	n
1	r_{11}	r_{12}	…	r_{1n}
2	r_{21}	r_{22}	…	r_{2n}
…	…	…	…	…
N	r_{N1}	r_{N2}	…	r_{Nn}
Сумма рангов	r_1	r_2	…	r_n

N- количество экспертов, n – количество объектов,

в N+1 строке таблицы стоят суммы рангов, полученных от экспертов объектами,

r_{ij} - ранг j-го объекта, присвоенного i-м экспертом.

Все n объектов упорядочиваются в соответствии с величиной $r_s = \sum_{i=1}^{N} r_{is}$,

на первое место в ранжировании ставится объект, которого r_s минимально, и т.д.

КОЭФФИЦИЕНТ КОНКОРДАЦИИ

Степень согласованности мнений экспертов при получении итогового ранжирования определяется через расчет **коэффициента конкордации**:

$$W = \frac{12 \cdot \sum_{i=1}^{n} \left[r_i - \frac{1}{2} \cdot N \cdot (n+1) \right]^2}{N^2 \cdot (n^3 - n)}$$

КОЭФФИЦИЕНТ КОНКОРДАЦИИ

В случае, когда ранжирование нестрогое (то есть допускается наличие равноценных объектов), коэффициент конкордации вычисляется по другой формуле:

$$W = \frac{12 \cdot \sum_{i=1}^{n} [r_i - \frac{1}{2} \cdot N \cdot (n+1)]^2}{N^2 \cdot (n^3 - n) - N \sum_{i=1}^{N} \sum_{j=1}^{k_i} (t_{ij}^3 - t_{ij})}$$

где k_i – число групп равных рангов, введенных i-ым экспертом, t_{ij} – количество равных рангов в j-ой группе, введенной i-ым экспертом. Коэффициент конкордации равен 1, если все ранжировки экспертов одинаковы, и равен нулю, если все ранжировки различны. Согласованность экспертов считается высокой, если $W \geq 0{,}8$.

Непараметрические статистические гипотезы

- При обработке статистических данных большого объема часто возникает ситуация, когда закон распределения генеральной совокупности не известен заранее.
- Сравнение гистограммы с известными кривыми функций плотностей позволяет выдвинуть гипотезу о виде распределения генеральной совокупности.
- Часто возникает необходимость проверить гипотезу о предполагаемом законе неизвестного распределения.
- Такая проверка осуществляется с помощью критериев согласия, например **критерия Пирсона χ^2, Колмогорова, Смирнова.**
- Обычно эмпирические и теоретические частоты различаются, но возможно расхождение случайно, незначимо и объясняется малым числом наблюдений, способом группировки, другими причинами.
- Возможно, что расхождение вызвано неверным предположением, например, о нормальном распределении генеральной совокупности.

Распределения

- **Нормальное** (гауссово, симметричное, колоколообразное) распределение **(normal, Gaussian distribution)–** описывает совместное воздействие на изучаемое явление небольшого числа случайно сочетающихся факторов (по сравнению с общей суммой факторов), число которых неограничено велико.
- Встречается в природе наиболее часто, за что и получило название «нормального».
- Характеризует распределение непрерывных случайных величин.

Распределения

- **Биномиальное** распределение (распределение Бернулли) **(binomial distribution, Bernoulli distribution)** – описывает распределение частоты события, обладающего постоянной вероятностью появления при многократных испытаниях. При большом числе испытаний стремиться к нормальному.
- Крайним вариантом биномиального распределения является **альтернативное** распределение, при котором вся совокупность распределяется на две части (две альтернативы).
- Это вероятностное распределение, связанное с двумя взаимоисключающими исходами, например, наличием или отсутствием симптома или лабораторного показателя, смерть или выживание.
- Биномиальное распределение характеризует распределение дискретных случайных величин.

Распределения

- **Распределение Пуассона** – описывает события, при которых с возрастанием значения случайной величины, вероятность появления ее в совокупности резко уменьшается.
- Распределение Пуассона характерно для редких событий (редких заболеваний) и может рассматриваться также как крайний вариант биномиального.
- Характеризует распределение дискретных случайных величин.

Критерий Пирсона

- Статистикой критерия является величина

$$F_{набл} = \chi^2 = \sum^{k} \frac{(n - n \cdot p_m)^2}{n \cdot p_m}$$, где k – число интервалов, n – объем выборки.

$$p_m = \Phi\left(\frac{x_m - \bar{x}}{\sigma}\right) - \Phi\left(\frac{x_{m-1} - \bar{x}}{\sigma}\right)$$, где $\bar{x}$ - среднее арифметическое, Ф – интеграл вероятностей,

$\sigma = \sqrt{s^2}$ – среднее квадратичекое отклонение.

- Эта величина является мерой расхождения эмпирических частот и теоретических частот .
- Критическое значение критерия равно обратному распределению хи-квадрат со степенями свободы(k – r – 1):
- $\chi^2_{kr} = \chi^2_{1-\alpha}(k - r - 1)$, где k–количество интервалов эмпирического распределения, r– число оцениваемых параметров закона распределения, α–заданный уровень значимости.
- Если $\chi^2 > \chi^2_{kr}$, то гипотеза H_0 отвергается; если выполняется условие $\chi^2 < \chi^2_{kr}$, то распределение можно считать соответствующим теоретическому, другими словами гипотеза H_0 не противоречит опытным данным.

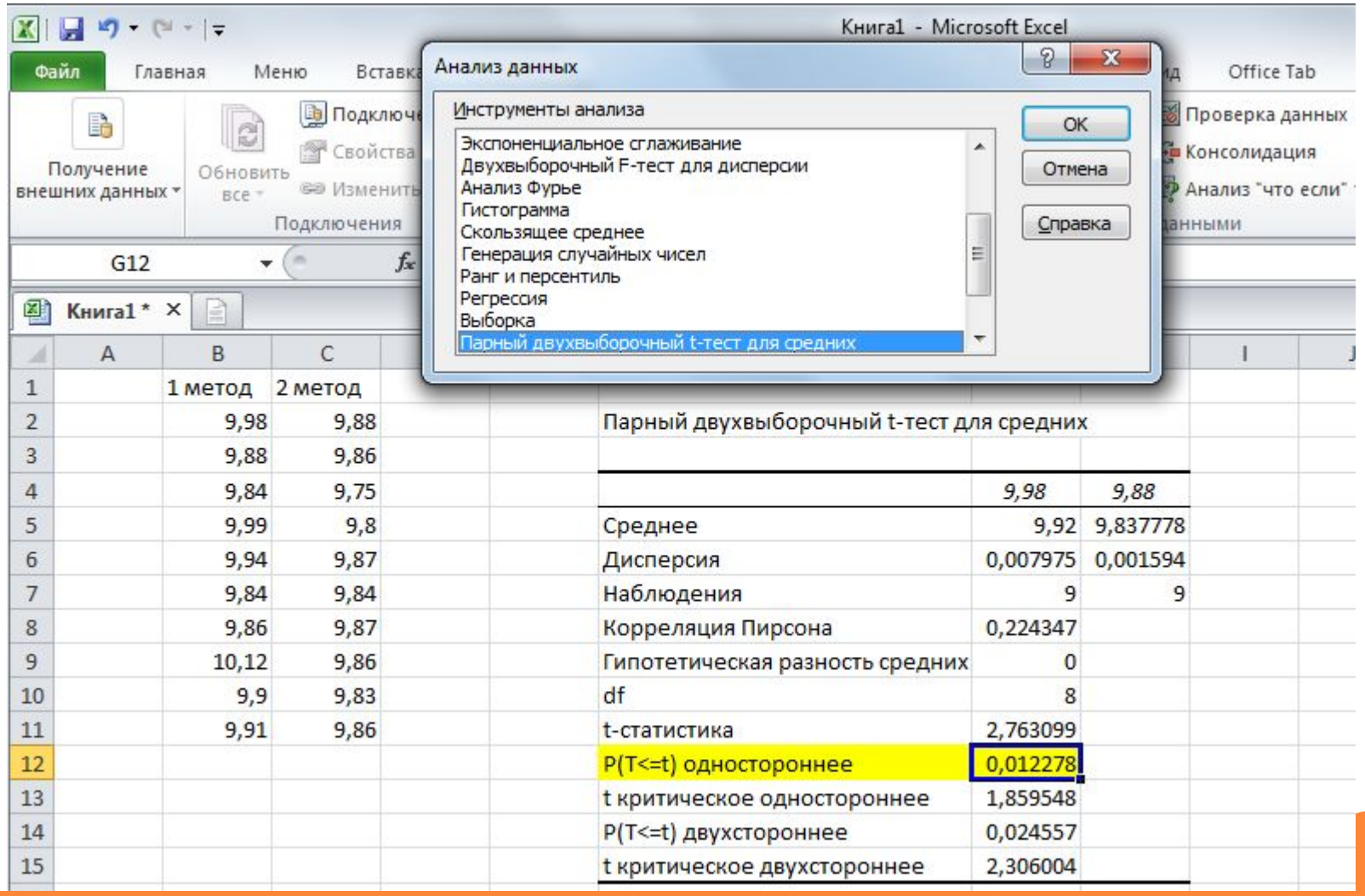
Статистические гипотезы

С помощью статистических расчетов вычисляется значение p, которое затем сравнивается с заранее выбранным *уровнем значимости*, часто обозначаемому греческой буквой α (альфа) (не путать с ошибкой 1-го типа). Обычно в биомедицинских исследованиях уровень значимости устанавливается на уровне $\alpha \leq 0{,}05$ ($\leq 5\%$) . Если выбран уровень значимости $\alpha = 0{,}05$, то все выборки, которые для выдвинутой гипотезы возвращают величину $p \leq 0{,}05$, отвергают эту гипотезу, а выборки с величиной $p > 0{,}05$ не дают оснований для того, чтобы её отвергнуть. Величину уровня значимости следует понимать в том смысле, что мы задаём, что не более чем в 5% попыток сравнения (какого-либо параметра в разных группах) обнаруженная разница может быть обусловлена чистой случайностью, а не тем, что разница действительно существует.

Задача 4
Использование *t*-критерия

- Критерий t Стьюдента направлен на оценку различий величин средних двух выборок X и Y, которые распределены по нормальному закону. Одним из главных достоинств критерия является широта его применения. Он может быть использован для сопоставления средних у связных и несвязных выборок, причем выборки могут быть не равны по величине.
- Для сравнения двух **независимых выборок** используется **непарный t-критерий**, для двух **зависимых** выборок используется **парный t-критерий**.
- Рассмотрим пример, когда из первой генеральной совокупности извлекается случайная выборка, имеющая объем n_1 а из второй — случайная выборка, объем которой равен n_2. Необходимо проверить гипотезу о равенстве средних.

Задача 4
Использование *t*-критерия



	A	B	C
1		1 метод	2 метод
2		9,98	9,88
3		9,88	9,86
4		9,84	9,75
5		9,99	9,8
6		9,94	9,87
7		9,84	9,84
8		9,86	9,87
9		10,12	9,86
10		9,9	9,83
11		9,91	9,86
12			
13			
14			
15			

Парный двухвыборочный t-тест для средних

	9,98	*9,88*
Среднее	9,92	9,837778
Дисперсия	0,007975	0,001594
Наблюдения	9	9
Корреляция Пирсона	0,224347	
Гипотетическая разность средних	0	
df	8	
t-статистика	2,763099	
P(T<=t) одностороннее	0,012278	
t критическое одностороннее	1,859548	
P(T<=t) двухстороннее	0,024557	
t критическое двухстороннее	2,306004	

Поскольку *p*-значение равно 0,01 и меньше $\alpha < 0{,}05$, нулевую гипотезу H_0 следует отклонить.

Задача 5

Сравнение 2-х средних нормальных генеральных совокупностей, дисперсии которых известны (независимые выборки).

По двум независимым выборкам, объемы которых соответственно n=60 и m=50, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние =1250 и =1275. Генеральные дисперсии известны: D(X)=120, D(Y)=100. При уровне значимости **0,01** проверить нулевую гипотезу H_0 : M(X) =M(Y) при конкурирующей H_1: M(X) ≠M(Y).

$$F_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/m}}$$

$$\Phi_{F_{кр}} = (1-\alpha)/2, \text{ где}$$

$$\Phi(t) = \frac{1}{\sqrt{2 \cdot \pi}} \int_0^t e^{-\frac{x^2}{2}} dx.$$

Если $|F_{наб}| < F_{кр}$, то нет оснований отвергать нулевую гипотезу.

кр нао кр средние различаются значимо.

Задача 6
Сравнение 2-х дисперсий нормальных генеральных совокупностей (F-критерий).

Правило 1: Для того, чтобы при заданном уровне значимости α (достоверности) проверить нулевую гипотезу $H0 : D(x) = D(y)$ равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H1: D(x) > D(y)$, надо вычислить наблюдаемое значение критерия (отношение большей исправленной дисперсий к меньшей) .

$F_{набл} = S_Б^2/S_М^2$ и по таблице критических точек распределения Фишера по заданному уровню значимости α и числам степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ найти критическую точку $F_{кр}(\alpha, k_1, k_2)$. Если $F_{наб} < F_{кр}$, что нет оснований отвергнуть нулевую гипотезу.

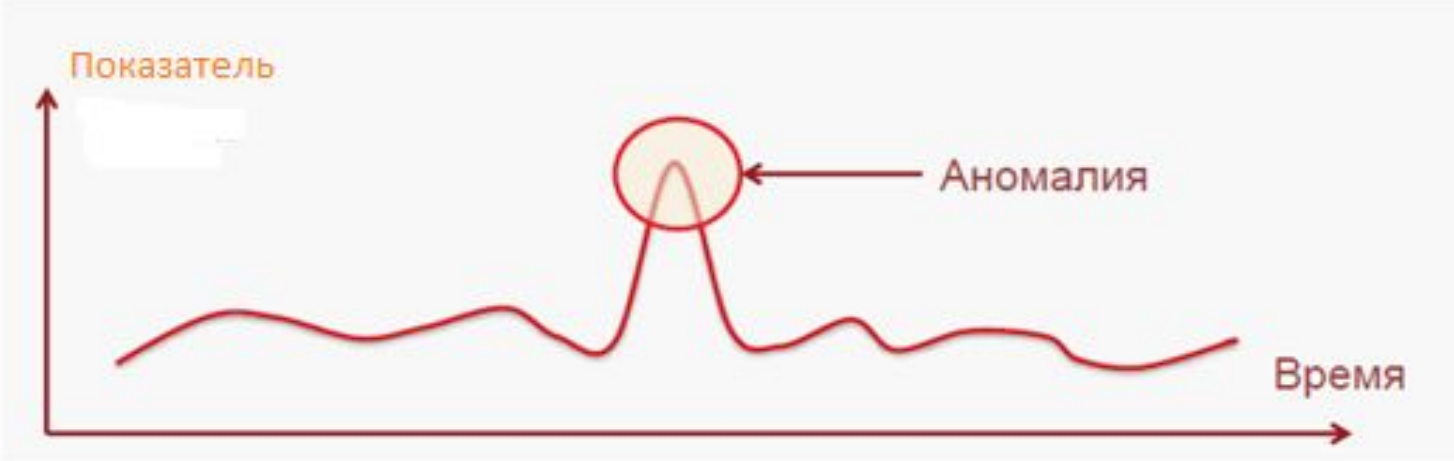
Если $F_{наб} > F_{кр}$, нулевую гипотезу отвергают.

МЕТОДЫ ПОДГОТОВКИ ДАННЫХ К АНАЛИЗУ

- Реальные данные для анализа редко бывают хорошего качества
- С целью повышения качества данных используется комплекс методов и алгоритмов, получивших название «очистка данных» (cleaning, refinement)
- Использование «*грязных*» данных может привести к **выявлению ложных закономерностей, ошибочных прогнозов и к неверным управленческим решениям**.

МЕТОДЫ ПОДГОТОВКИ ДАННЫХ К АНАЛИЗУ

- Очистка от шумов и сглаживание рядов данных
- Редактирование аномальных значений
- Восстановление пропущенных значений
- Обработка дубликатов и противоречий
- Снижение размерности входных данных
- Устранение незначащих факторов



Причины аномальных данных

- **искусственные** — связаны с ошибками ввода данных, некорректной работой программ или технических систем регистрации и ввода данных;
- **естественные** — отражают факты и события, имевшие место в действительности, но вызванные исключительными обстоятельствами, которые встречаются очень редко или в единичных случаях.

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

Атрибут Возраст представлен следующими двадцатью значениями:

*{3, 56,23, 39, **156**, 52, 41, 22, 9, 28, **139**, 31, 55, 20, **–67**, 37, 11, 55, 45, 37}*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 39.6 \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 45.89, \qquad T = \bar{x} \pm 2\sigma$$

Потенциальные аномалии: 156, 139 и –67 (ошибки ввода).

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

- В основе метода лежит оценка мер расстояния между всеми наблюдениями в n-мерном пространстве данных
- Значение S_i множества данных S является аномальным, если хотя бы **часть значений *p*** из множества S **расположена на большем расстоянии, чем *d***, от остальных значений.
- **Пример**

S - множество двумерных наблюдений, где требованием для аномальности является значение порогов ***p* ≥4 и *d*≥ 3**.

$$S = \{S_1, S_2, S_3, S_4, S_5, S_6, S_7\} = \{(2,4),(3,2),(1,1),(4,3),(1,6),(5,3),(4,2)\}.$$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

$p \geq 4$ и $d \geq 3$.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1		2,236	3,162	2,236	2,236	3,162	2,828
S_2			2,236	1,414	4,472	2,236	1,000
S_3				3,605	5,000	4,472	3,162
S_4					4,242	1,000	1,000
S_5						5,000	5,000
S_6							1,414

Значение	S_1	S_2	S_3	S_4	S_5	S_6
Параметр p	2	1	5	2	5	3

$S3$ и $S5$ - кандидаты в аномальные, для них значение $p = 5$ превышает заданный порог $p \geq 4$.

МЕТОДЫ КОРРЕКТИРОВКИ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

- Удаление записи с аномальным значением
- Ручная замена аномальных значений
- Сглаживание и фильтрация данных
- Интерполяция данных
- Замена на наиболее вероятное значение

ПРОИСХОЖДЕНИЕ ПРОПУСКОВ В ДАННЫХ

- В процессе ввода данных, ошибки.
- При сбое в работе автоматических систем регистрации.
- В процессе загрузки данных случаи пропуска могут возникать на месте значений, имеющих некорректный тип или формат.

МЕТОДЫ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

- Ручная обработка пропусков (применим только для небольших выборок данных)
- Подстановка констант
- Предсказание пропущенных значений (нейронная сеть, дерево решений)
- Подстановка среднего значения

DATA MINING – КЛАССЫ РЕШАЕМЫХ ЗАДАЧ

- Классификация
- Регрессия
- Кластеризация
- Ассоциация
- Последовательность

Классификация

Нахождение функциональной зависимости между входными атрибутами и **дискретным выходным** атрибутом.

Классификация позволяет отнести объект к одному из известных классов.

Регрессия

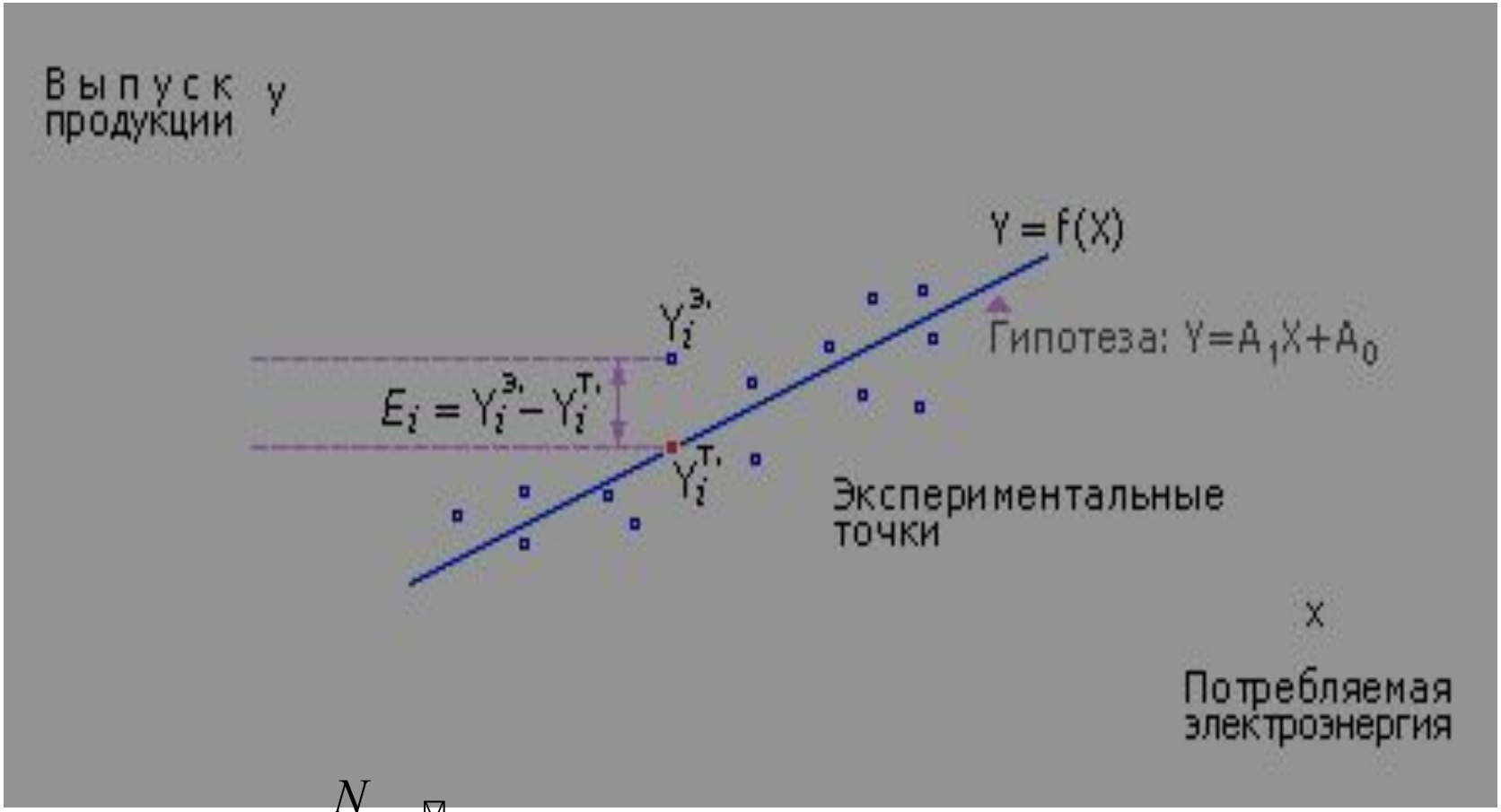
Регрессией называется зависимость среднего значения одной случайной величины от некоторой другой (или от нескольких случайных величин).

- Цель регрессионного анализа – по результатам наблюдений за входными и выходными величинами найти зависимость между входами и выходом, т.е. получить математическую модель.
- Нахождение функциональной зависимости между входными атрибутами и **непрерывным выходным** атрибутом.

Задачи регрессионного анализа :

- Прогнозирование ухудшения состояния пациента.
- Оценка вероятности повторных рецидивов заболевания.
- Расчет загруженности докторов при обслуживании населения.
- Анализ влияния различных факторов на исследуемый.

РЕГРЕССИОННЫЙ АНАЛИЗ



$$Q = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \to \min.$$

РЕГРЕССИОННЫЙ АНАЛИЗ

$$\frac{\partial F}{\partial A_0} = -2\sum_{i=1}^{n}(Y_i - A_0 - A_1 X_i) = 0$$

$$\frac{\partial F}{\partial A_1} = -2\sum_{i=1}^{n}(Y_i - A_0 - A_1 X_i)X_i = 0$$

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \cdot \begin{pmatrix} A_0 \\ A_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

РЕГРЕССИОННЫЙ АНАЛИЗ

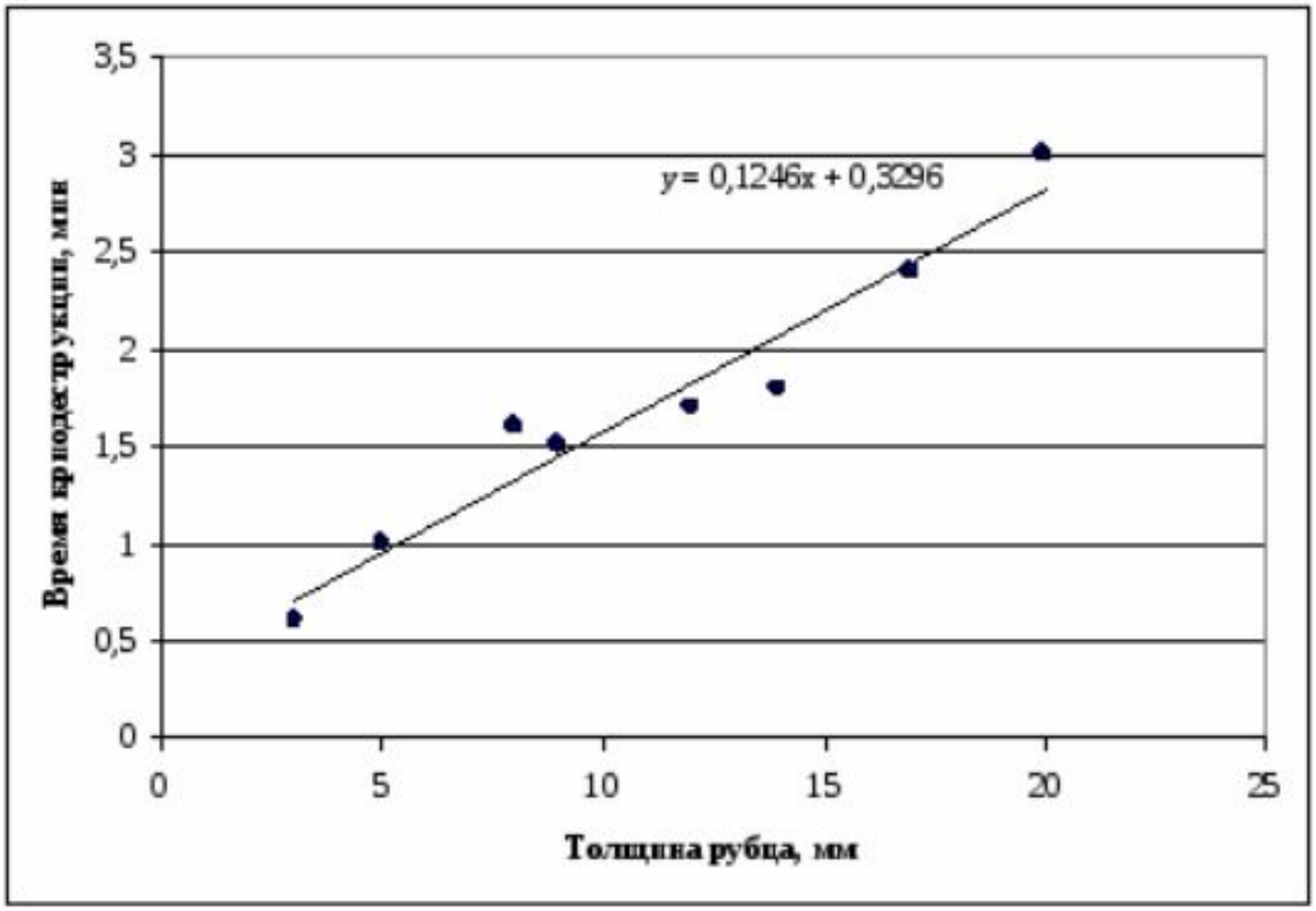
$$A_0 = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

Пересечение с осью ОУ

$$A_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

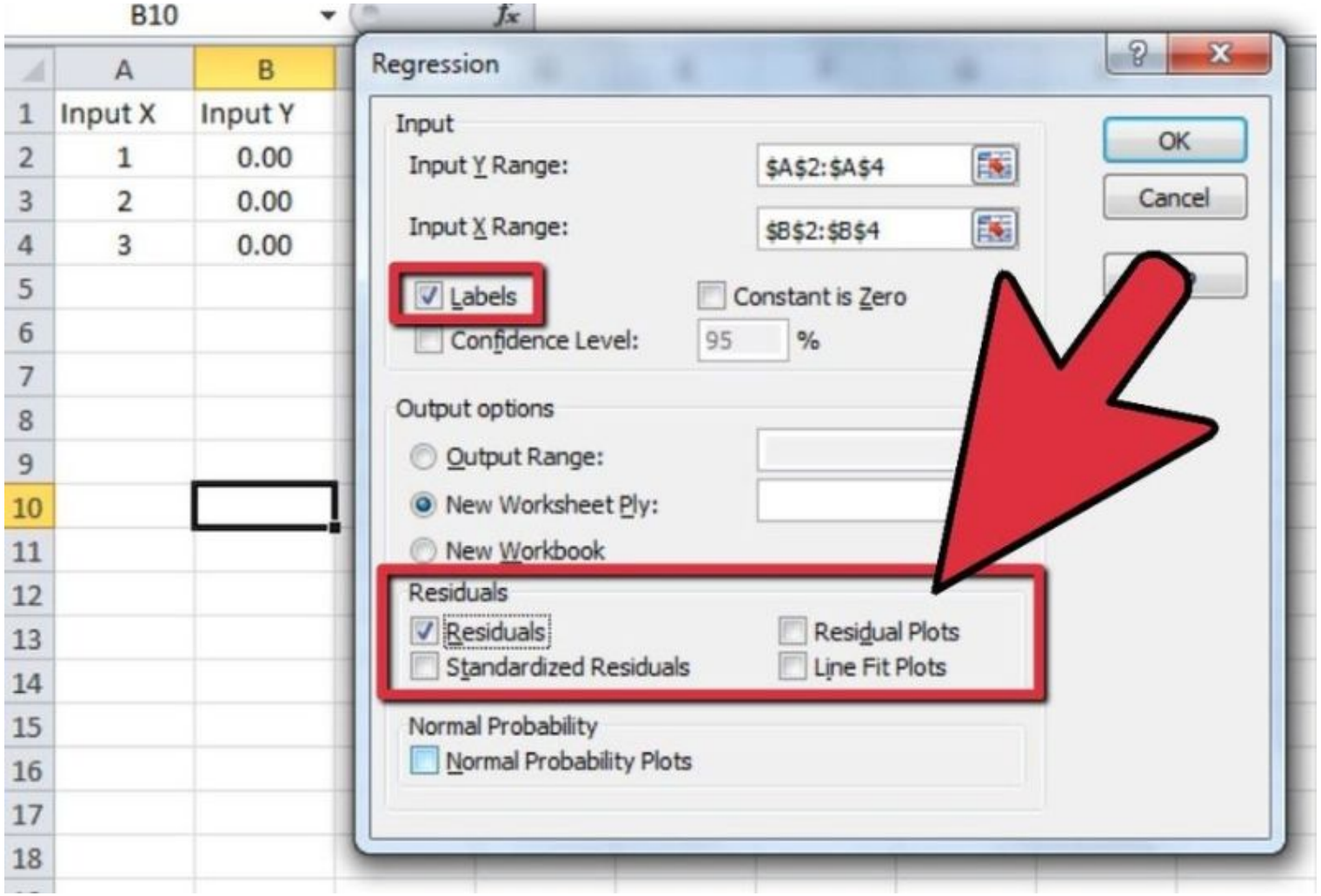
Наклон

Построение диаграммы



Ось X ввести текст «**Толщина рубца, мм**»; ***Ось Y*** – «**Время криодеструкции, мин**».

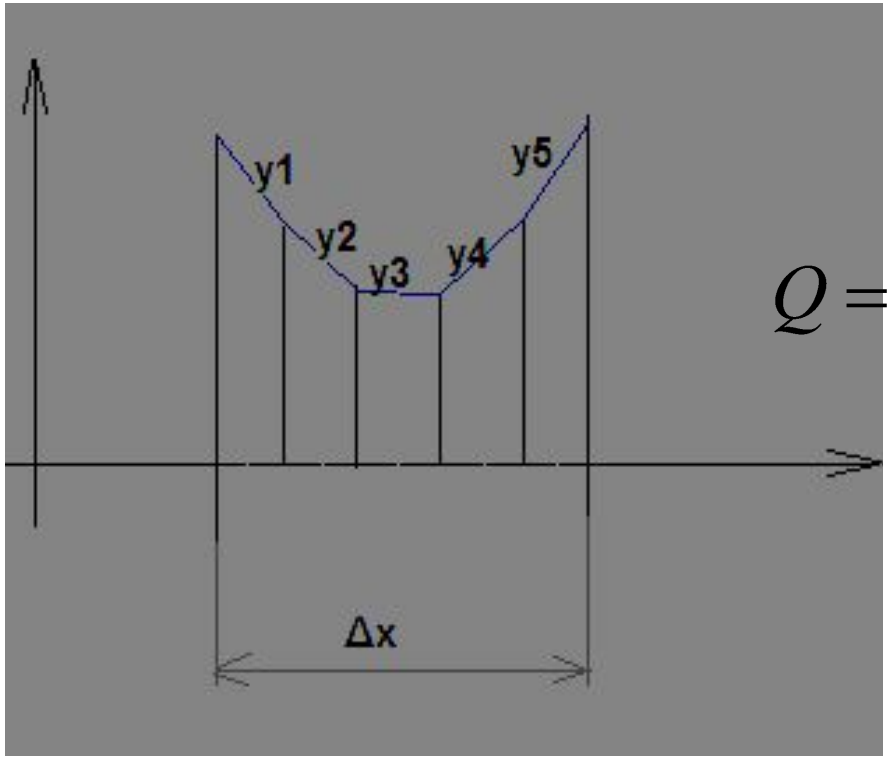
Регрессионный анализ данных



РЕГРЕССИОННЫЙ АНАЛИЗ

$y=a_0+a_1x_1+\ldots+a_nx_n.$

$a_0+a_1x_1+\ldots+a_nx_n+a_{12}x_1x_2+\ldots+a_{11}x_1^2+a_2x_2^2+\ldots+a_{nn}x_n^2$



$$Q=\sum_{i=1}^{N}(\hat{y}_i-y_i)^2\rightarrow \min.$$

РЕГРЕССИОННЫЙ АНАЛИЗ

$$y=a_0+a_1x_1+\ldots+a_nx_n,$$

$$A = (X^{T}\cdot X)^{-1}\cdot X^{T}\cdot Y$$

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{2i} & \cdots & \sum_{i=1}^{n} X_{mi} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}X_{1i} & \sum_{i=1}^{n} X_{2i}X_{1i} & \cdots & \sum_{i=1}^{n} X_{mi}X_{1i} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{1i}X_{2i} & \sum_{i=1}^{n} X_{2i}X_{2i} & \cdots & \sum_{i=1}^{n} X_{mi}X_{2i} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} X_{mi} & \sum_{i=1}^{n} X_{1i}X_{mi} & \sum_{i=1}^{n} X_{2i}X_{mi} & \cdots & \sum_{i=1}^{n} X_{mi}X_{mi} \end{pmatrix} \cdot \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ \cdots \\ A_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_iX_{1i} \\ \sum_{i=1}^{n} Y_iX_{2i} \\ \cdots \\ \sum_{i=1}^{n} Y_iX_{mi} \end{pmatrix}$$

РЕГРЕССИОННЫЙ АНАЛИЗ

- После проведения эксперимента необходимо убедиться в существовании линейной зависимости, адекватности линейной модели в пределах выбранного диапазона значений входной величины. Оценка отклонения от

$$R_{x/y}{}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$, где $\hat{y}_i$ - расчетное, y_i - измеренное $\bar{y}$ -среднее

- выбрать такую, которая дает наибольшее значение R^2
- Чем больше R^2, т. е. чем больше числитель, тем больше изменение факторного признака объясняет изменение результативного признака и тем, следовательно, лучше уравнение регрессии, лучше выбор функции.

Кластеризация

- Разбиение объектов на кластеры, т.е. **группы схожих элементов**:
- Кластеризация пациентов со схожей историей болезни, особенностями восстановления после болезни.
- Анализ спроса на медицинские услуги в зависимости от комбинации входных показателей.
- Обнаружение аномальных отклонений.

Ассоциация

- Анализ транзакций, т.е. событий, **происходящих вместе**.
- Обнаружение зависимости вида **«Из события А с определенной вероятностью следует событие В»:**
- Прогноз реакции организма пациента при появлении определенного симптома.

Последовательность

- Анализ событий, связанных между собой по времени.
- **«После события А спустя определенное время произойдет событие В»:**
- Анализ потребности пациентов в лекарствах.

СПАСИБО ЗА ВНИМАНИЕ.