Биостатистика

2. Статистическое оценивание и проверка гипотез.

Рубанович А.В.

Институт общей генетики им. Н.И. Вавилова РАН

Чем мы занимались на предыдущем занятии? Фактически теорией вероятностей!

- Мы вычисляли вероятность наблюдаемого расклада (комбинации событий) при условии случайности и независимости этих событий

- Эту вероятность мы вычисляли «в лоб», используя комбинаторику и биномиальное распределение Бернулли. Это была статистика «на пальцах», точнее говоря на монетах

- На этом пути мы освоили точный тест Фишера, предназначенный для сравнения частот событий

- К сожалению, для решения большинства других задач статистики такой «честный путь» невозможен.

- Вместо этого по результатам измерений вычисляется новая величина, т.н. статистика теста (t, χ^2, Z, …), и уже по ее значениям косвенно судят о неслучайности эффекта.

Несколько обязательных общих понятий

- Статистика - это экспериментальный анализ случайных величин. Мы пытаемся судить о неизвестных случайных величинах по конечной совокупности наблюдений за ними (выборке).

- Неизвестный нам закон распределения наблюдаемой случайной величины называется генеральным.

- Выборка - это последовательность чисел $x_1, \ldots, x_n$, полученных при n-кратном повторении эксперимента в неизменных условия, например это могут быть значения признака для n различных особей

- Характеристики выборки (среднее, дисперсия) являются приблизительными оценками истинных параметров неизвестного нам генерального распределения

Обычно по результатам биологического эксперимента появляется некий Excel-файл

	Признак 1	Признак 1	…
Особь 1			
Особь 1			
…			

Признаки могут быть:

- Количественные (непрерывные или счетные)
- Качественные (номинальные или порядковые)

Несколько советов по хранению данных:

- Вносите все данные в одну электронную таблицу. Не надо для каждой популяции создавать новый файл
- Тщательно продумывайте названия столбцов и обозначения для номинальных признаков
- При внесении текстовых данных следите за унификацией: Генотип «А С» - это не то же самое, что «АС» или « АС». Следите также за раскладкой клавиатуры

Познакомьтесь: наша учебная «база данных». Она будет использована для иллюстраций

	A	B	C	D	E	F	G	H	I	J
1	*ФИО*	*Регион*	*Пол*	*Возраст*	*Вес*	*Рост*	*Болезнь*	*АберХр*	*GSTM1*	*GSTP1_A313G*
2	1	Москва	М	47	111,8	181,8	0	0	del/del	A/A
3	2	Омск	Ж	53	81,0	154,9	0	0,610178	del/del	A/G
4	3	Омск								
5	4	Омск								
6	5	Киев								
7	6	Киев								
8	7	Москва	М	53	56,8	151,9				
9	8	Киев	Ж	49	58,5	142,9				
10	9	Москва	М	48	78,0	172,0				
11	10	Омск	М	43	85,2	189,8	0	0	del/del	A/A
12	11	Москва	М	42	76,7	179,1	0	0	del/del	A/A
13	12	Москва	Ж	58	78,7	145,6	1	0,013907	I/*	A/A
14	13	Москва	Ж	42	71,8	166,4	1	0,815971	del/del	G/G
15	14	Киев	Ж	53	60,7	145,1	1	0,75661	I/*	A/G
16	15	Москва	Ж	41	91,5	154,6	0	0,177917	I/*	A/G
17	16	Киев	М	41	104,3	180,2	0	0,208379	I/*	A/A
18	17	Омск	Ж	56	77,2	158,7	1	0,376719	I/*	A/G
19	18	Москва	Ж	44	77,9	145,4	1	0,504461	del/del	A/A
20	19	Москва	М	48	89,8	181,6	0	0,114204	del/del	A/A
21	20	Омск	Ж	43	58,0	142,4	0	0,191457	I/*	G/G
22	21	Киев	М	46	82,2	152,6	1	0,032043	I/*	A/A
23	22	Москва	Ж	56	80,7	164,1	0	0	I/*	A/G
24	23	Киев	М	47	86,3	163,1	0	0	del/del	A/A
25	24	Омск	Ж	43	46,9	140,4	0	0	del/del	A/A
26	25	Киев	Ж	53	61,4	152,4	0	0	I/*	A/G
27	26	Москва	Ж	48	85,6	193,5	0	0	del/del	A/A
28	27	Киев	М	41	84,7	141,6	0	0	I/*	A/A
29	28	Киев	Ж	50	94,1	166,2	1	0	del/del	A/G
30	29	Москва	Ж	42	60,3	142,1	0	0,701922	I/*	A/A
31	30	Москва	Ж	56	80,4	181,4	1	0,204385	I/*	A/A
32	31	Омск	Ж	46	87,0	143,0	0	0	del/del	A/A
33	32	Киев	М	38	92,5	178,5	1	0	I/*	A/G
34	33	Москва	М	43	79,2	130,9	0	0	I/*	A/G
35	34	Москва	Ж	46	61,7	136,2	0	0	I/*	A/A
36	35	Москва	Ж	43	85,2	147,6	0	0,295014	del/del	A/A
37	36	Омск	М	43	82,6	166,1	0	0,658415	del/del	A/A

Качестве номинальные

Количесвы призна

Качестве ый порядковый признак:
0 – контроль
1 - больной

Обзор данных: описательные статистики

- Среднее – основная характеристика «положения» случайной величины

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

Близкие характеристики «положения»

- Медиана – значения больше и меньше равновероятны
- Мода – наиболее вероятное значение случайной величины
- Среднее геометрическое $\overline{x_G} = \sqrt[n]{x_1 x_2 \ldots x_n}$

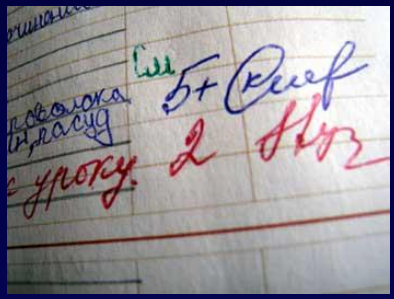
- Дисперсия – основная характеристика разброса случайной величины около среднего

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

Дисперсия имеет размерность $[x]^2$. Корень из дисперсии называется стандартным отклонением (*SD*) и имеет размерность $[x]$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Упражняемся…



Оценка	Число учеников (из 100)	
	Физика	Физкультура
2	10	0
3	50	10
4	30	20
5	10	70

- Чему равны средние оценки по физике и физкультуре?

Средняя оценка по физике = $0.1 \cdot 2 + 0.5 \cdot 3 + 0.3 \cdot 4 + 0.1 \cdot 5 = 0.2 + 1.5 + 1.2 + 0.5 = 3.4$

…. по физкультуре = $0 \cdot 2 + 0.1 \cdot 3 + 0.2 \cdot 4 + 0.7 \cdot 5 = 0 + 0.3 + 0.8 + 3.5 = 4.6$

- Для какого предмета дисперсия оценок выше?

Дисперсия оценок по физике =
$= 0.1 \cdot (2-3.4)^2 + 0.5 \cdot (3-3.4)^2 + 0.3 \cdot (4-3.4)^2 + 0.1 \cdot (5-3.4)^2 = 0.64$

Дисперсия оценок по физкультуре =
$= 0 \cdot (2-4.6)^2 + 0.1 \cdot (3-4.6)^2 + 0.2 \cdot (4-4.6)^2 + 0.7 \cdot (5-4.6)^2 = 0.44$

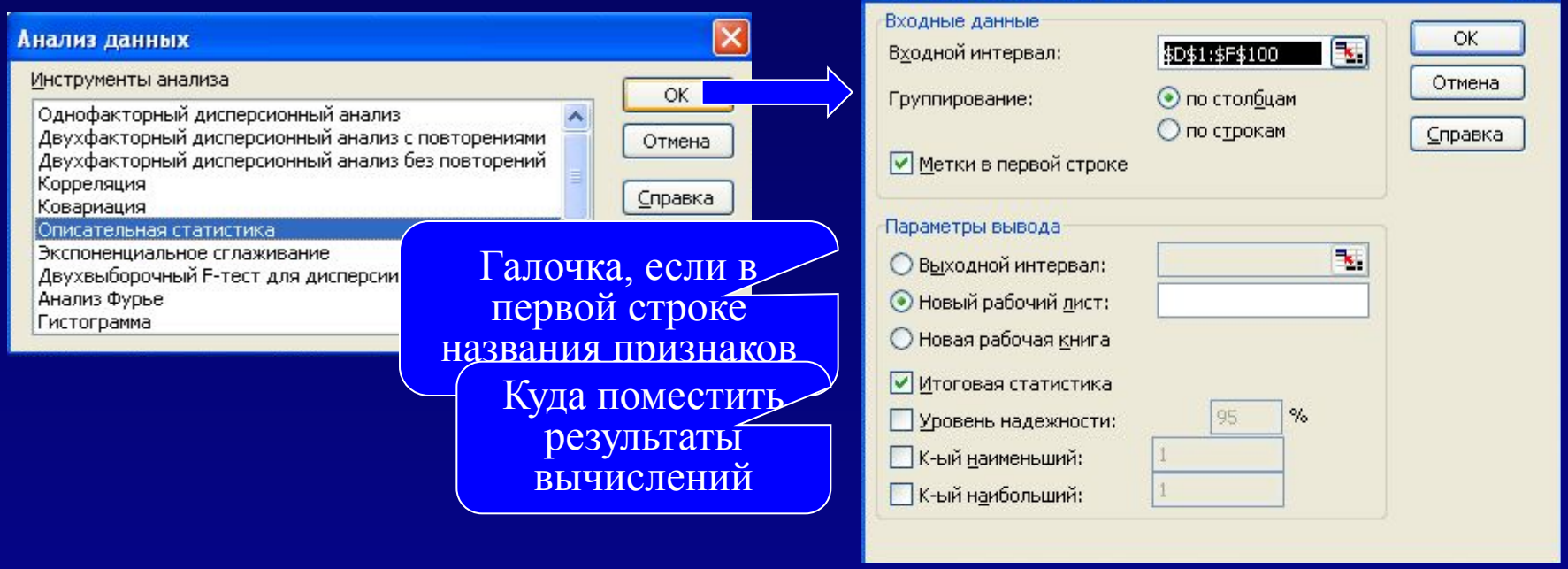
Обзор данных: описательные статистики с помощью Excel

- В Excel есть встроенные функции описательных статистик:

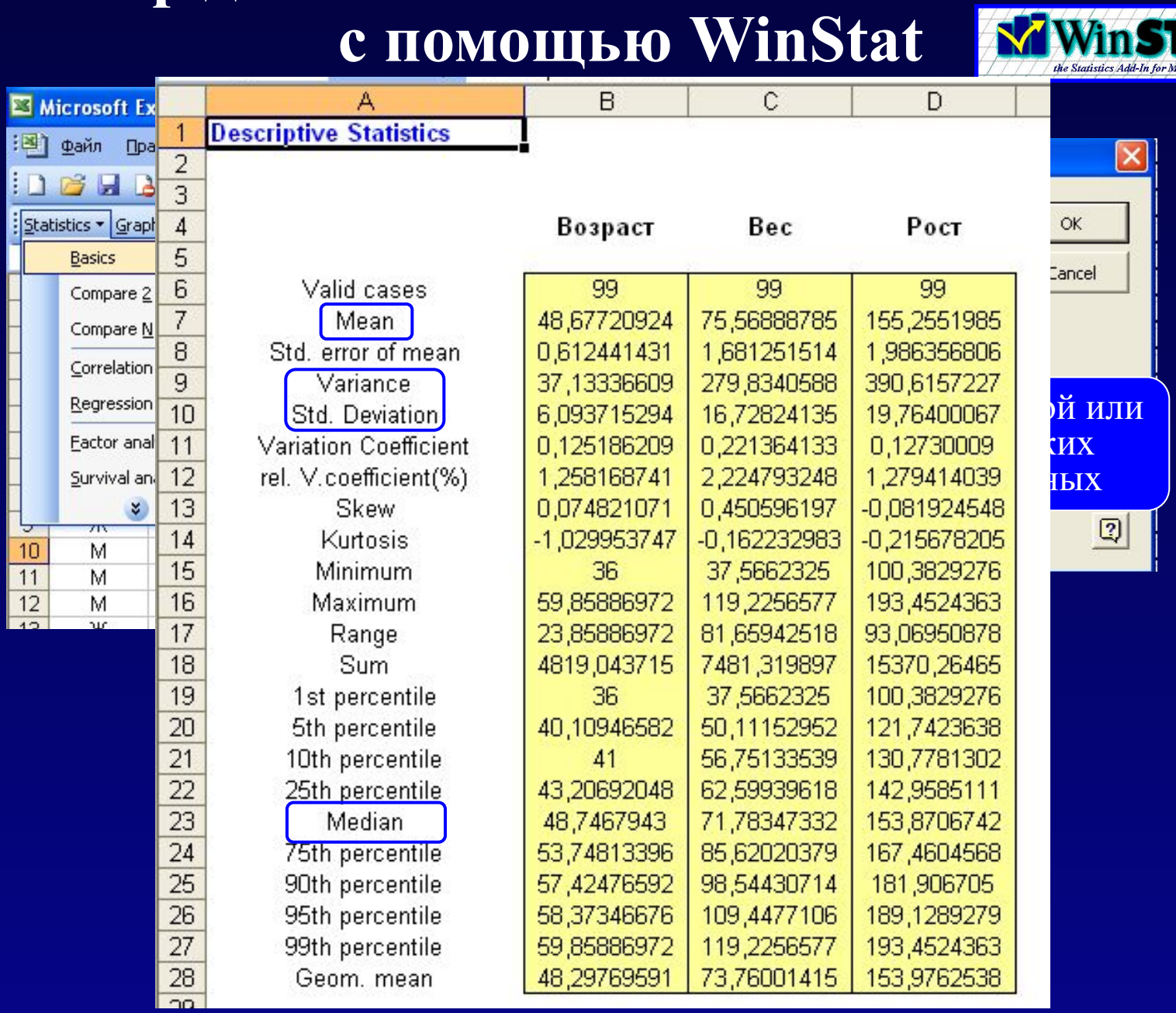
= СРЗНАЧ(число1; число2; …) или = СРЗНАЧ(диапазон)

= ДИСП(число1; число2; …) или = ДИСП(диапазон)

- Кроме того в пункте «Сервис» имеется пакет «А… держащий различные статистические процедуры



Обзор данных: описательные статистики с помощью WinStat

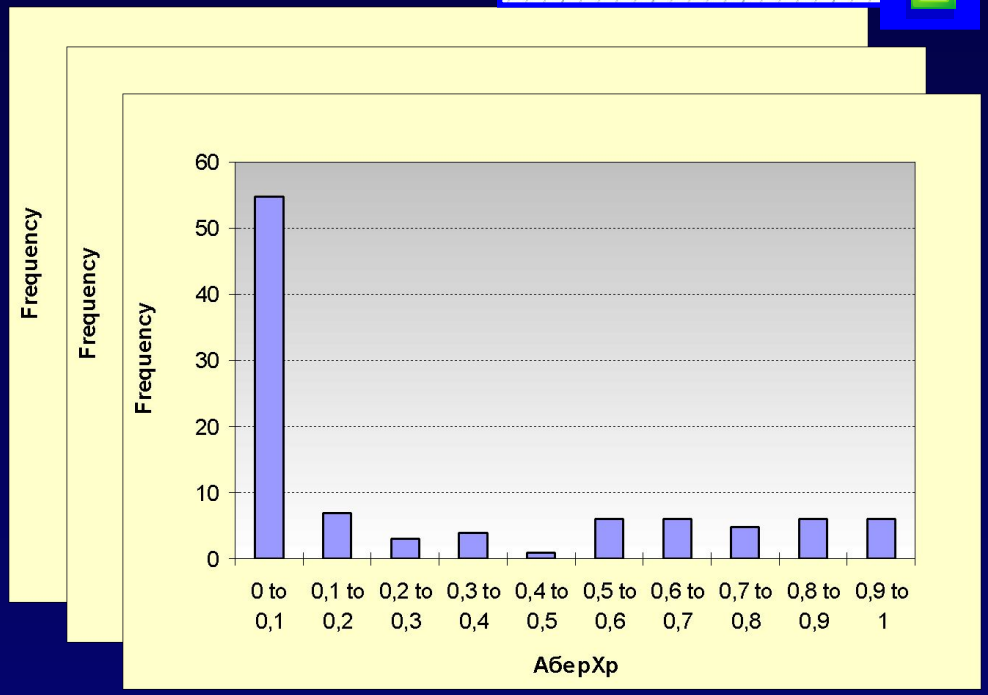


	A	B	C	D
1	Descriptive Statistics			
2				
3				
4		Возраст	Вес	Рост
5				
6	Valid cases	99	99	99
7	Mean	48,67720924	75,56888785	155,2551985
8	Std. error of mean	0,612441431	1,681251514	1,986356806
9	Variance	37,13336609	279,8340588	390,6157227
10	Std. Deviation	6,093715294	16,72824135	19,76400067
11	Variation Coefficient	0,125186209	0,221364133	0,12730009
12	rel. V.coefficient(%)	1,258168741	2,224793248	1,279414039
13	Skew	0,074821071	0,450596197	-0,081924548
14	Kurtosis	-1,029953747	-0,162232983	-0,215678205
15	Minimum	36	37,5662325	100,3829276
16	Maximum	59,85886972	119,2256577	193,4524363
17	Range	23,85886972	81,65942518	93,06950878
18	Sum	4819,043715	7481,319897	15370,26465
19	1st percentile	36	37,5662325	100,3829276
20	5th percentile	40,10946582	50,11152952	121,7423638
21	10th percentile	41	56,75133539	130,7781302
22	25th percentile	43,20692048	62,59939618	142,9585111
23	Median	48,7467943	71,78347332	153,8706742
24	75th percentile	53,74813396	85,62020379	167,4604568
25	90th percentile	57,42476592	98,54430714	181,906705
26	95th percentile	58,37346676	109,4477106	189,1289279
27	99th percentile	59,85886972	119,2256577	193,4524363
28	Geom. mean	48,29769591	73,76001415	153,9762538

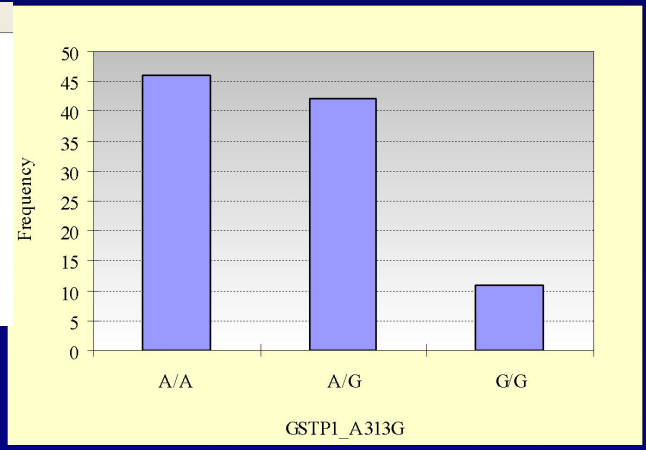
Обзор данных: смотрим характер распределений

Всегда необходимо просматривать:

- гистограммы распределений количественных признаков
- … и частоты встречаемости для качественных признаков, например, частоты генотипов



	A	B	C	D
1	**Frequencies**			
2				
3		**Frequency**	**Percent**	**Cumulative Percent**
4	**GSTP1_A313G**			
5	A/A	46	46,46	46,46
6	A/G	42	42,42	88,89
7	G/G	11	11,11	100,00

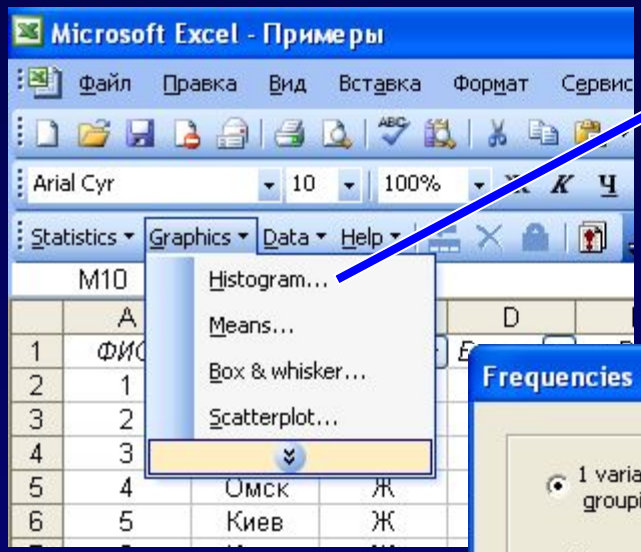


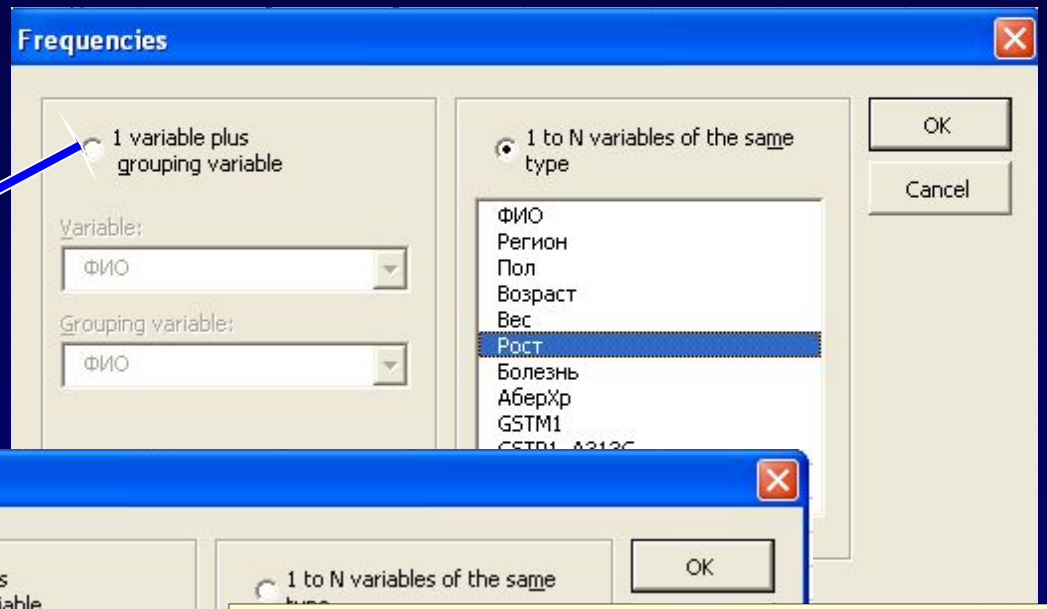
Можно использовать встроенный в Excel пакет «Анализ данных:

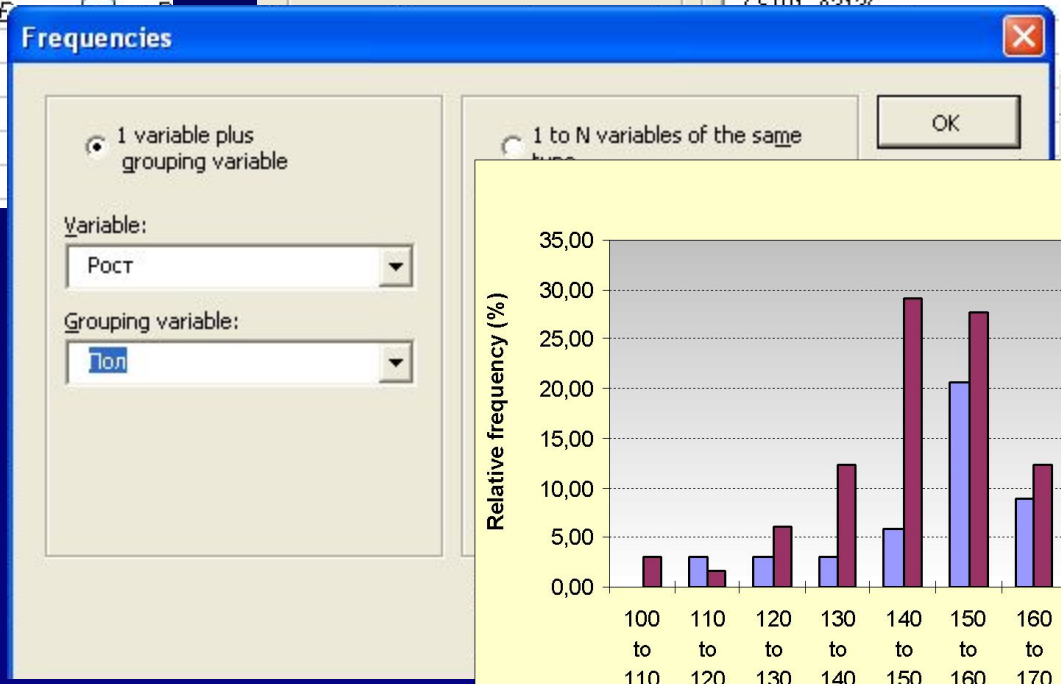
Обзор данных: смотрим характер распределений

С группировкой по номинальному признаку

Всегда необходимо просматривать:







Упражняемся…

Ошибки средних и доверительные интервалы

Выборочное среднее $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ является величиной случайной!

Стандартное отклонение этой случайной величины называется ошибкой среднего (*SE*). Можно показать, что

$$SE = \frac{SD}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

	A	B	C	D
1	Descriptive Statistics			
2				
3				
4		Возраст	Вес	Рост
5				
6	Valid cases	99	99	99
7	Mean	48,67720924	75,56888785	155,2551985
8	Std. error of mean	0,612441431	1,681251514	1,986356806
9	Variance	37,13336609	279,8340588	390,6157227
10	Std. Deviation	6,093715294	16,728	
11	Variation Coefficient	0,125186209	0,2213	

Слабо зависи… размеров выб…

…амих …да уменьшается …и увеличении …меров выборки

Почему 1.96 ? Мы еще об этом поговорим!

В отчетах можно писать: $\bar{x} \pm SE$

А можно указывать 95%-ый доверительный интер…

$$\left(\bar{x} - 1.96SE;\ \bar{x} + 1.96SE\right)$$

Это интервал, накрывающий истинное значение среднего с вероятностью 95%

Упражняемся…

Оценка	Число учеников (из 100)	
	Физика	Физкультура
2	10	0
3	50	10
4	30	20
5	10	70

Средняя оценка по физике = 3.4. Дисперсия = 0.64

Средняя оценка по физкультуре = 4.6. Дисперсия = 0.44

- Чему равны стандартные отклонения и ошибки самих оценок (*SD* и *SE*)?

По физике: 3.4±0.1 Можно записать так 3.40±0.08, но не так 3.4±0.08

$$SD = \sqrt{0.64} = 0.8 \qquad SE = \frac{0.8}{\sqrt{100}} = 0.08$$

По физкультуре: 4.6±0.1

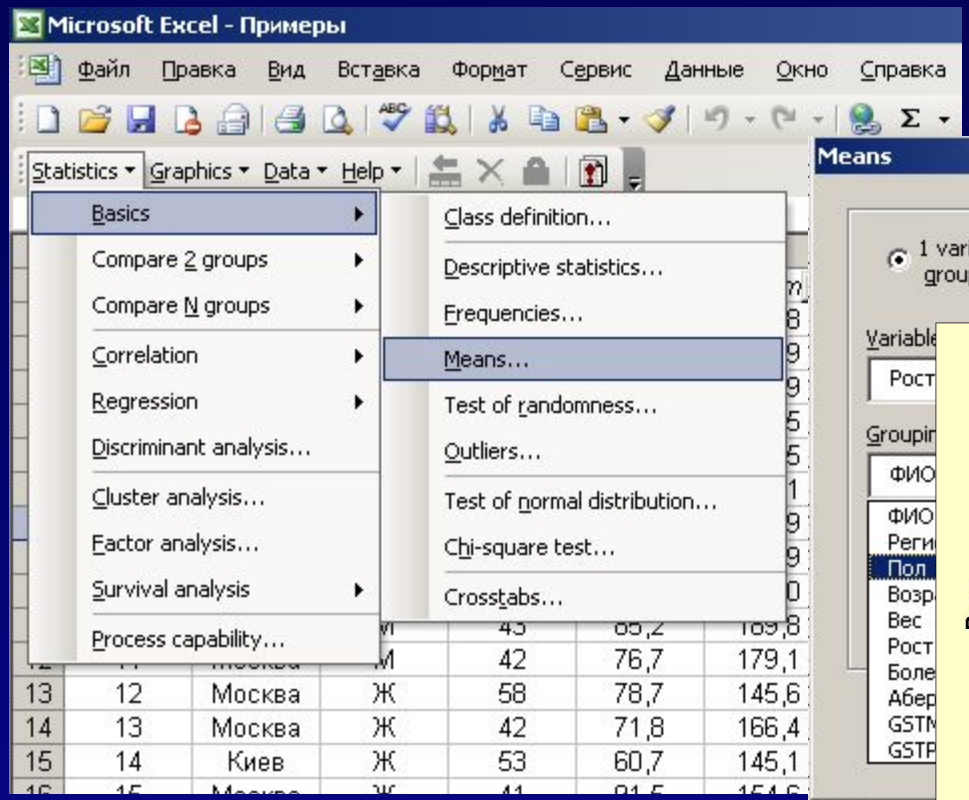
$$SD = \sqrt{0.44} = 0.66 \qquad SE = \frac{0.66}{\sqrt{100}} = 0.07$$

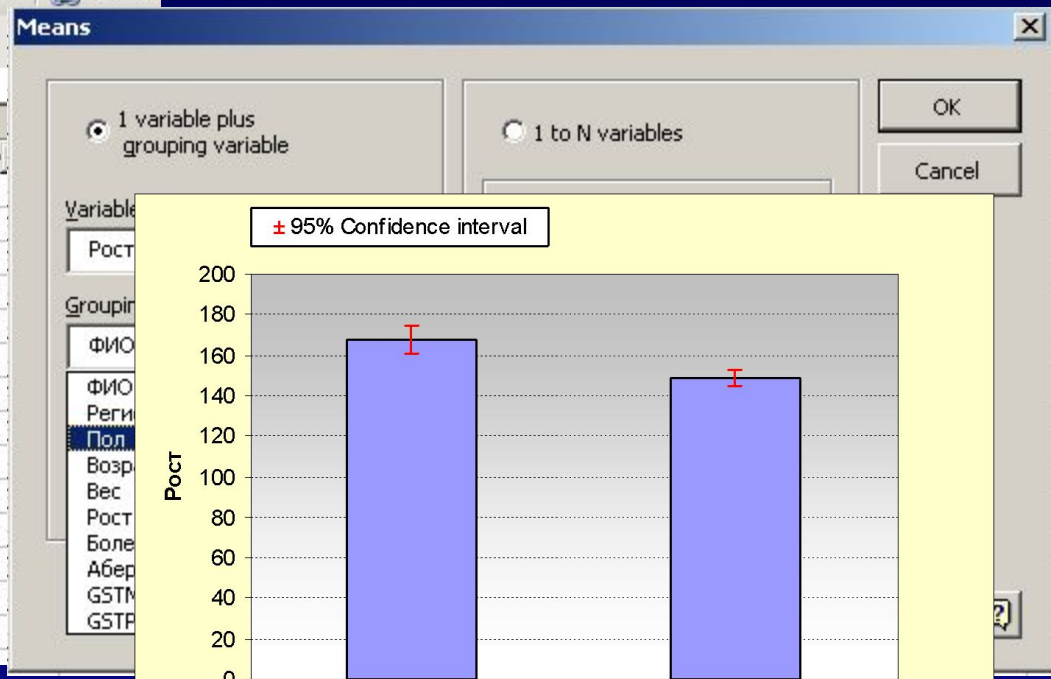
Упражняемся…

Конечно вручную это никто не считает!

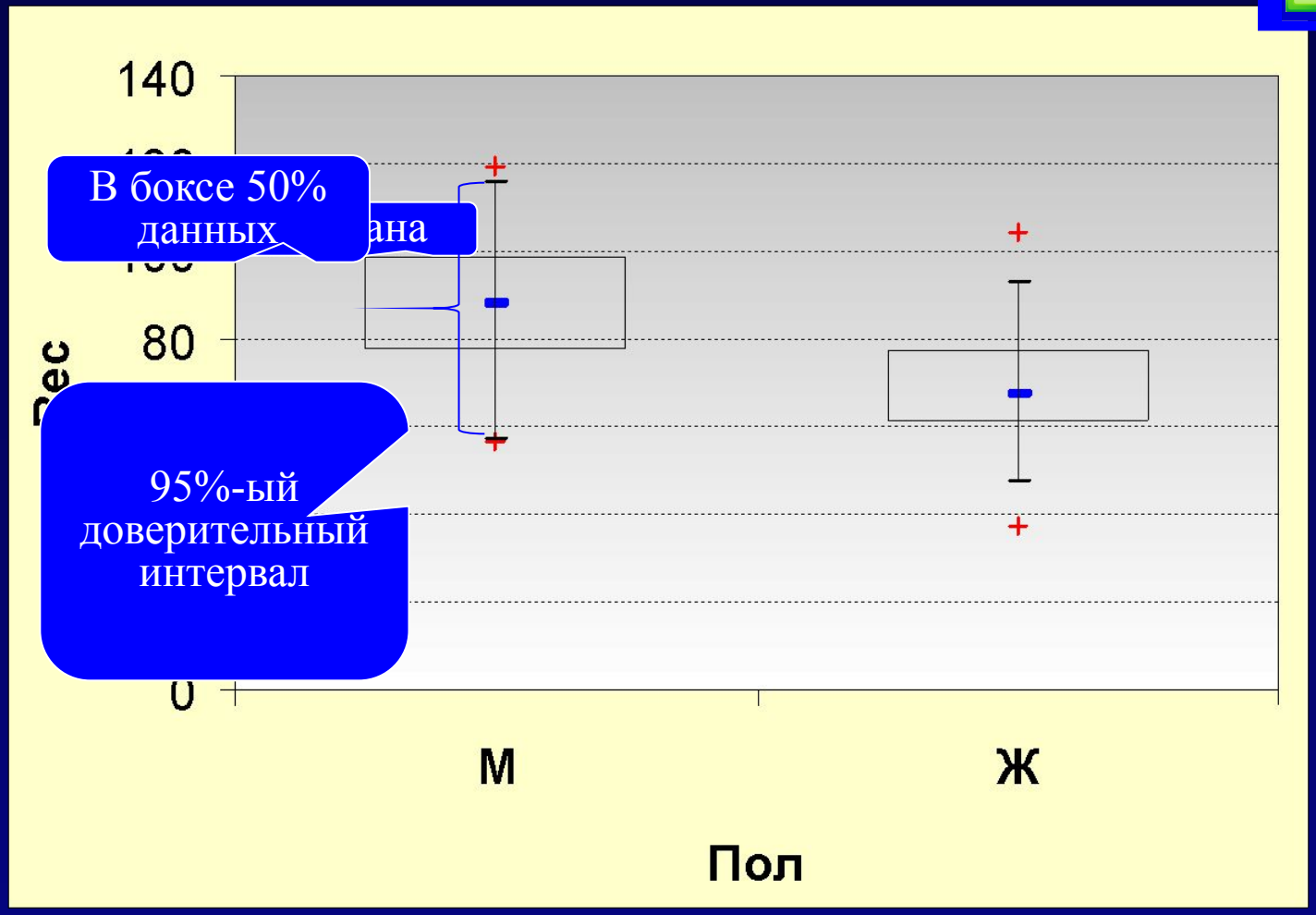
Можно использовать встроенный в Excel пакет «Анализ данных:

Еще удобней: WinSTAT® the Statistics Add-In for Microsoft® Excel





Боксы с усами (Box & Whisker) - еще один способ представления данных



Оценки частот тоже имеют ошибки и доверительные интервалы

	Количественный признак	Номинальный признак
Выборка	$\{x_1, x_2, \ldots x_n\}$	$\{m, n\}$
Среднее	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$p = \frac{m}{n}$
SD	$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	$\sqrt{p(1-p)}$
SE	$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{p(1-p)}{n}}$

95%-ый доверительный интервал для частоты:

$$\left(p - 1.96\sqrt{\frac{p(1-p)}{n}}; p + 1.96\sqrt{\frac{p(1-p)}{n}} \right)$$

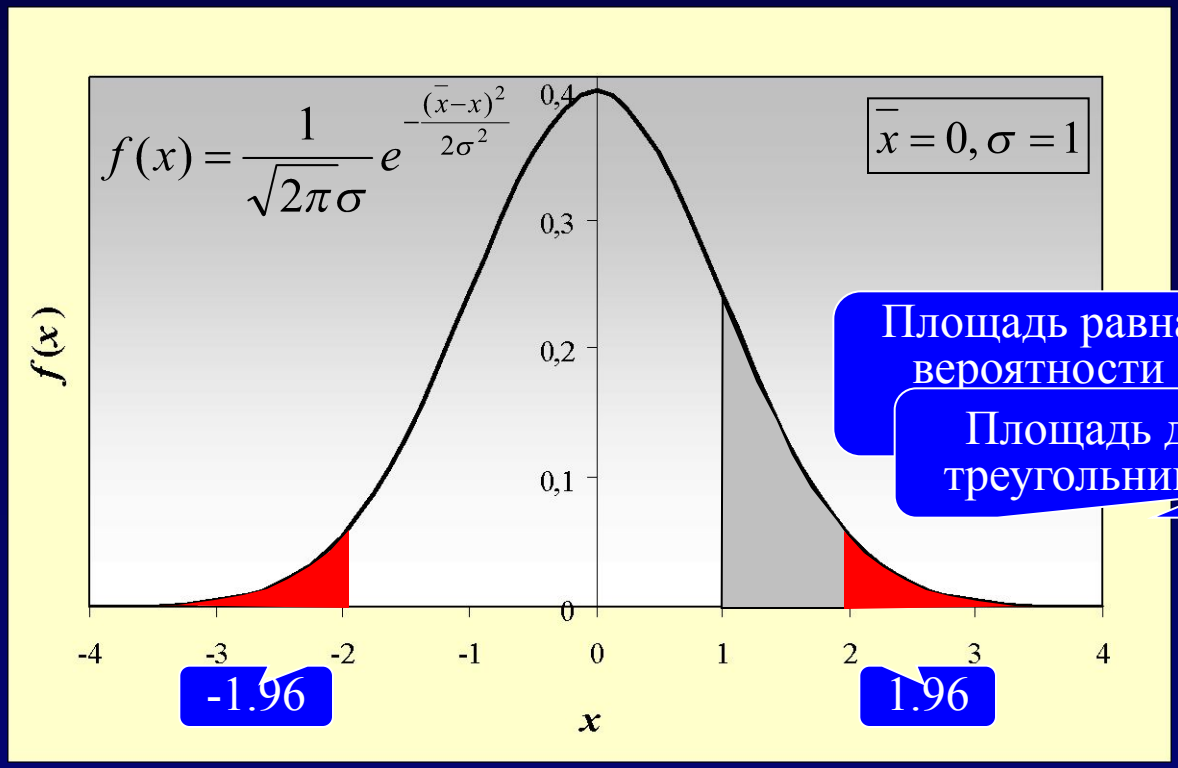
Еще лучше

WinPepi PORTAL Copyright J.H. Abramson, Jan. 9, 2010. Version 10.0

WhatIs/CI/Proportion

Поговорим о нормальном распределении

- Вы его много раз видели:



- Это плотность распределения (кривая, огибающая гистограмму). Площадь под кривой равна вероятности попадания *x* в соответствующий интервал.

- Площадь хвостов: $P(-1.96 < x < 1.96) = 0.95$

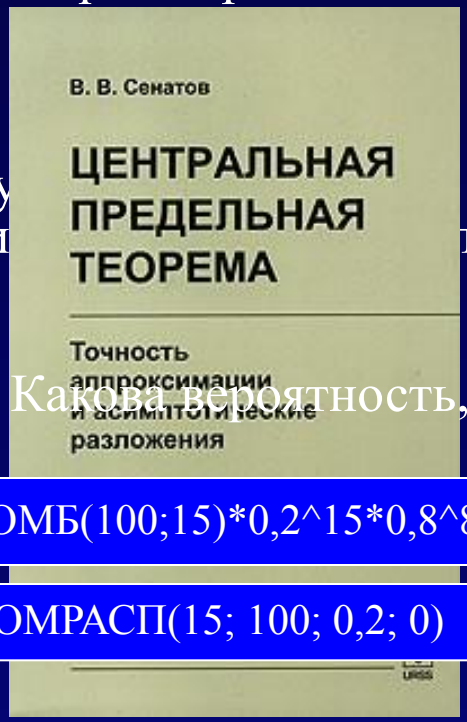
Отсюда 95%-ый доверительный интервал: $\left(\bar{x} - 1.96SE; \bar{x} + 1.96SE\right)$

Почему нормальное распределение встречается на каждом шагу?

- Нормальное распределение имеет любая величина, которая определяется суммой большого числа случайных слагаемых (**ЦПТ**).
  Чем больше слагаемых – тем «нормальней»!

- Например, биномиальный закон – это вероятность су
  независимых событий в *N* испытаниях. Поэтому, если
  распределение становится нормальным.

- Проверяем ... К 20 годам 80% молодых людей курит. Какова вероятность, что среди 100 окажется 15 некурящих?



С помощью биномиального распределения:

P(15) = 0.048

= ЧИСЛКОМБ(100;15)*0,2^15*0,8^85

или

= БИНОМРАСП(15; 100; 0,2; 0)

С помощью нормального распределения:

= НОРМРАСПР(15;20;4;0)

Среднее число некурящих $Np = 100 \cdot 0.2 = 20$,
дисперсия равна $Np(1-p) = 100 \cdot 0.2(1-0.2) = 16$, $\sigma = 4$.

P(15) = 0.046

Гипотезы и статистики

Ключевые понятия

- Гипотеза – это предположение о виде распределения или значении параметра генерального распределения (например о среднем)

- Нулевая гипотеза (H_0) - обычно предположение о случайном характере наблюдаемых различий или об отсутствии эффектов

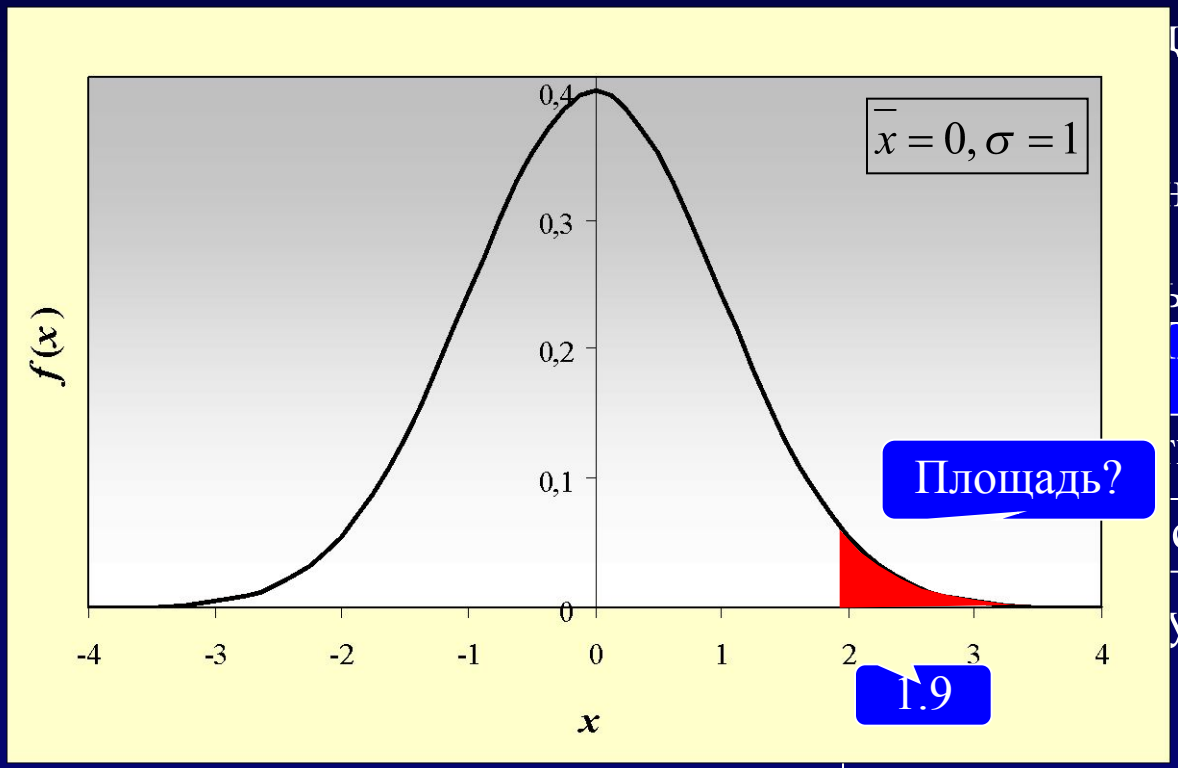
- Альтернативная гипотеза (H_1) формулируется в зависимости от характера теста – односторонний или двусторонний

- Статистический критерий – это правило, согласно которому принимается или отвергается гипотеза.

- Статистика – это функция от выборочных наблюдений на основе которой принимается или отвергается нулевая гипотеза

Гипотезы и статистики

Знакомый пример



ценка $p = 0.47$ при $n = 1000$

ний тест

ь «менее 470 из 1000» при

НОМРАСП(470; 1000; 0,5; 1)

ть правильную H_0 мала:

Считать сумму от 0 до 470

сть оши

у. Вычисляется некая теста), характер

Z - статистика $= \dfrac{p - 0.5}{\sqrt{\dfrac{p(1-p)}{n}}} = 1.9$

$$Z = \frac{x - \bar{x}}{\sigma_{x-\bar{x}}}$$

= (0,5-0,47)/КОРЕНЬ (0,47*0,53/1000)

$\alpha = 0.029$

=1- НОРМРАСП(1,9; 0; 1;1)

Однако по двустороннему тесту (p≠1/2) нам следует отвергнуть H_0: $2 \cdot 0.031 > 0.05$

О том же говорит размер доверительного интервала:

Вероятность упустить и вероятность обознаться

В жизни, а также при проведение статистических тестов возможны два типа ошибок:

- отвергнуть правильную нулевую гипотезу
- принять неправильную нулевую гипотезу

Нулевая гипотеза – обычно предположение об отсутствии различий, например, 2 выборки взяты из одной генеральной совокупности

Ошибка I рода (α)

Вероятность отвергнуть правильную нулевую гипотезу =
Вероятность обнаружить различия там, где их нет = Вероя
совершить фальшивое открытие

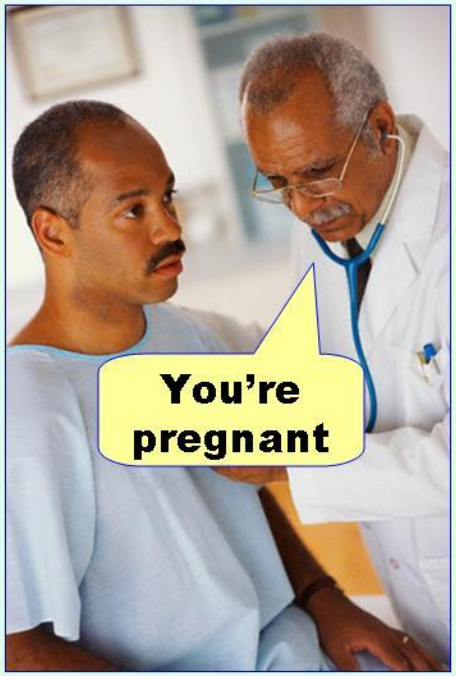
Ошибка II рода (β)

Вероятность принять неправильную нулевую гипотезу =
Вероятность не обнаружить существующие различия =
Вероятность упустить открытие

Вероятность упустить и вероятность обознаться

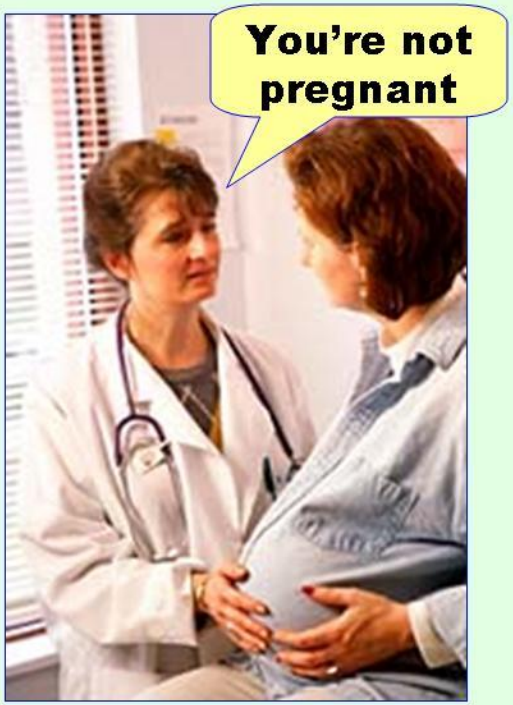
H_0 – беременности нет

Type I error
(false positive)



Отвергнута правильная нулевая гипотеза. Сделано фальш-положительное открытие

Type II error
(false negative)



Принята неправильная нулевая гипотеза. Фальш-негативный вывод. Открытие упущено

Вероятность упустить и вероятность обознаться

От чего зависят ошибки статистических тестов?

- От размаха реально существующих отличий и разброса данных
- От объемов выборок

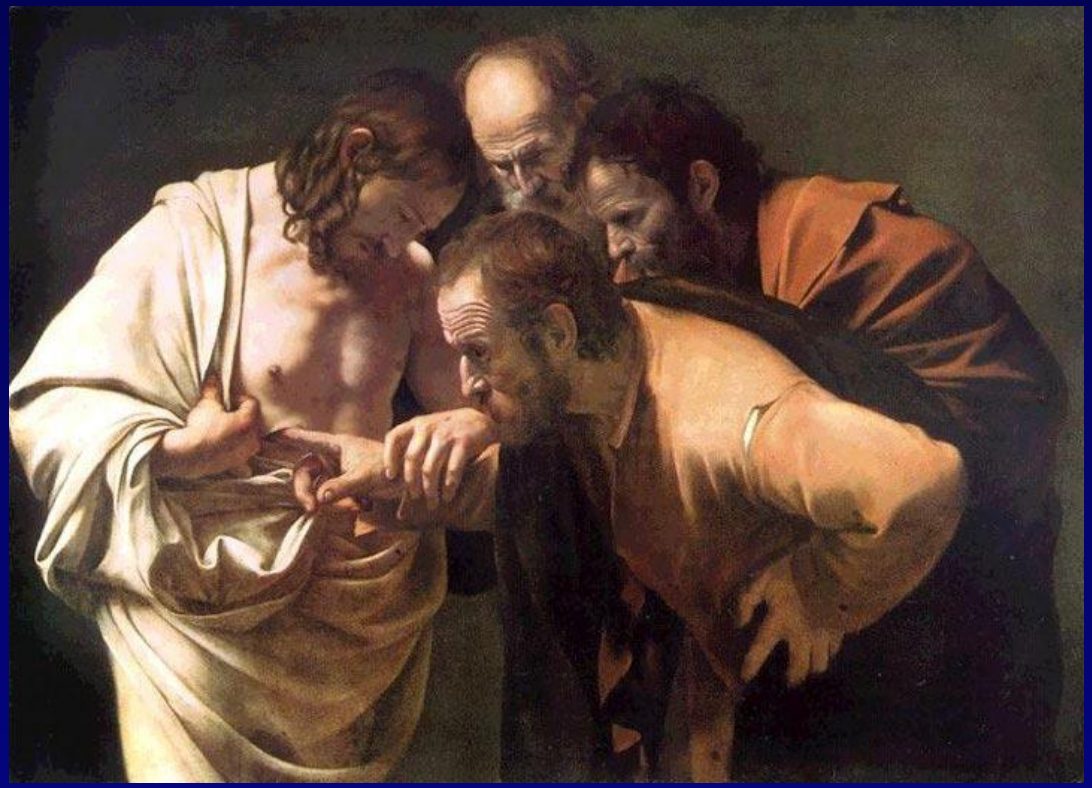
  • Ошибка I рода (вероятность фальшивого открытия) слабо зависит от объемов выборок, если они сравнимы по величине

  • С увеличением объема выборки вероятность ошибки II рода (вероятность упустить открытие) всегда уменьшается

- Ошибки I и II рода однозначно не связаны. В целом ошибка II рода растет при уменьшении ошибки I рода

Вероятность упустить и вероятность обознаться

«Критерий» св. Фомы Неверующего (0033):
всегда принимаем H_0
(т.е. различий нет, и все всегда случайно)



Караваджо (1573-1610). Фома Неверующий

Ошибка I рода = 0 $\Leftrightarrow$ Ошибка II рода = 1

Вероятность упустить и вероятность обознаться

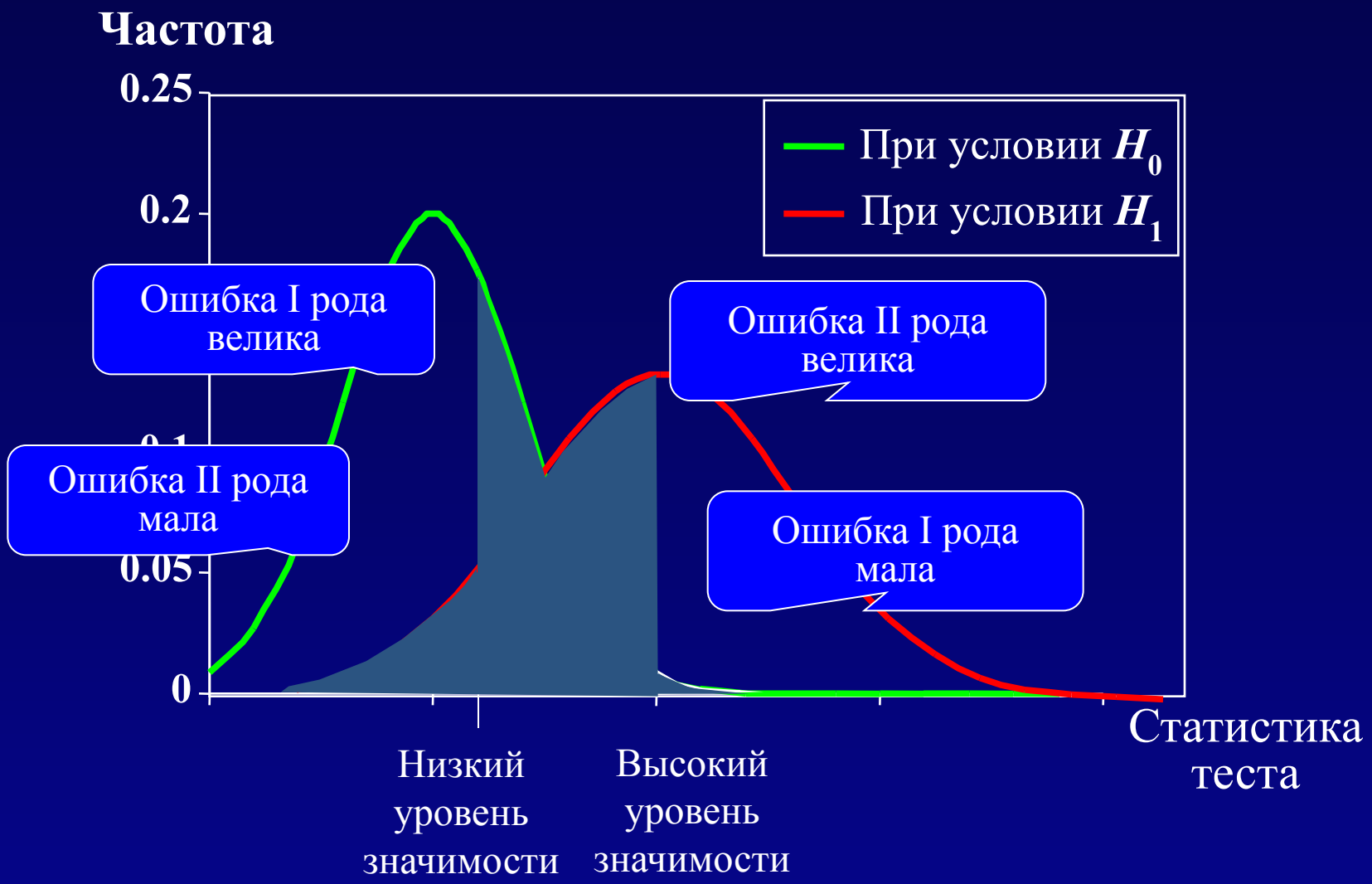
α vs. β : противоборство показателей теста

Всегда принимаем H_0 α=0, β=1

Всегда отвергаем H_0 ~~α=1~~, β=0

Уменьшая ошибку I рода, увеличиваем ошибку II рода, т.е. теряем мощность теста (*et converso)*

Вероятность упустить и вероятность обознаться



Вероятность упустить и вероятность обознаться

Мощность теста = 1- β

т.е. вероятность правильно отвергнуть нулевую гипотезу
или вероятность не упустить открытие

- Мощность 80% считается приемлемой
- Консервативный тест - это тест с низкой мощностью
- Мощностью теста резко возрастает при увеличении объемов выборок
- При планировании экспериментов имеет смысл прикинуть возможную мощность тестов

Например, Compare2/ Power/ Comparison of proportions
Size A - 100 Size B – 100
a/A – 0.2 b/B – 0.1

Мощность = 44%

… и необходимый объем выборок

Например, Compare2/ Sample size/ Proportions
Size A/ Size B =1
a/A – 0.2 b/B – 0.1

Общий объем выборок = 398

На сегодня это все

Напоследок хочу посоветовать:

- Если Вы этого никогда не делали, составьте базу данных в Excel и посчитайте самостоятельно описательные статистики

- Поставьте на свой компьютер WinPepi и оцените возможности этой программы

- Подумайте над тем, ошибки какого рода Вы чаще совершаете – I или II ? Это полезно для усвоения настоящего материала.