Биостатистика

3. Анализ количественных признаков

Рубанович А.В.

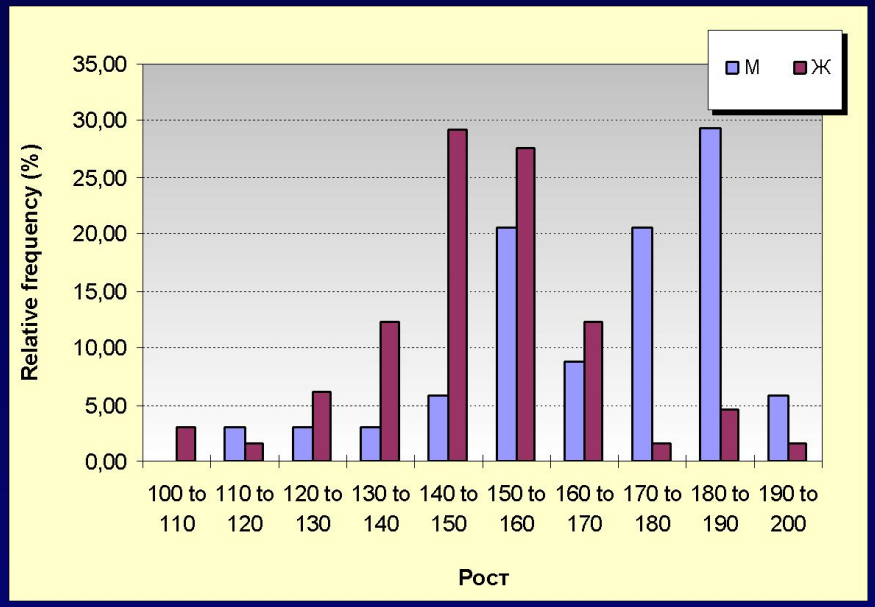
Институт общей генетики им. Н.И. Вавилова РАН

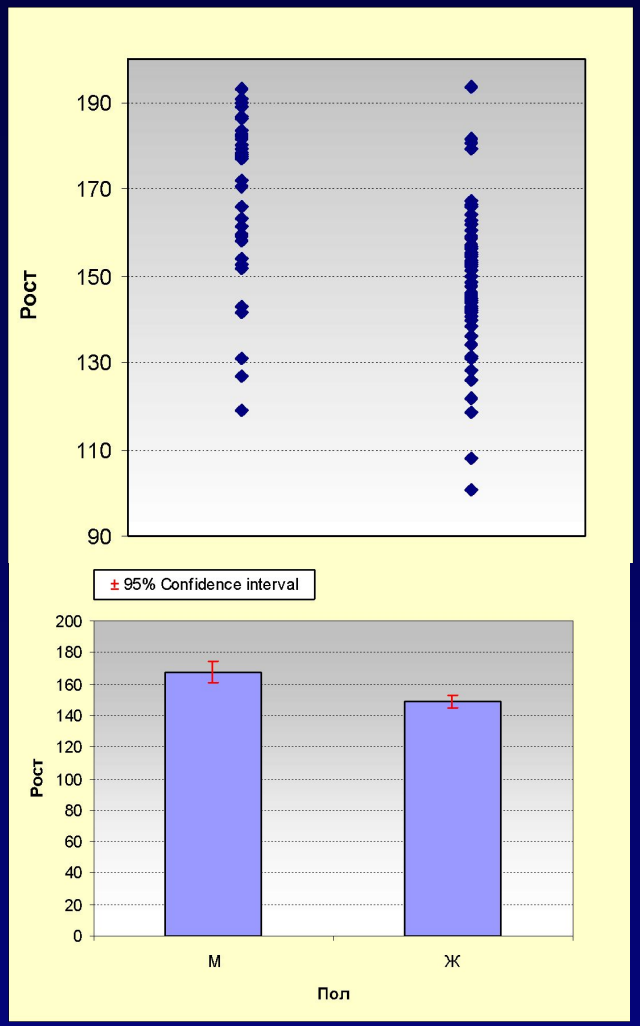
Чем мы занимались на предыдущем занятии?

- Мы вспомнили общепринятые методы описания и представления данных
- На примере качественных признаков (данных о частотах) познакомились с принципами построения и проверки статистических гипотез
- Поговорили о вероятностях возможных ошибок, возникающих при использовании всякого статистического теста
- При этом мы сознательно не затрагивали ряд традиционных для статистики тем: сравнение средних, критерий Стьюдента и т.д.
- Отчасти потому, что вы об этом наверняка наслышаны, но в основном из методических соображений

Сравнение средних

Перейдем, наконец, к задаче
о сравнении средних для двух выборок.
Например, рост в выборках «М» и «Ж»





Нулевая гипотеза состоит в предположении, что обе выборки изъяты из одной генеральной совокупности (т.е. различий нет):

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

H_1: $\bar{x}_1 \neq \bar{x}_2$ (двусторонний тест)

Дальше надо предложить способ оценить вероятность ошибки I рода

Сравнение средних

На прошлом занятии мы рассмотрели достаточно универсальный способ построения статистических критериев: Z – статистика, т.е. $Z = \xi / \sigma_\xi$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

, т.е. разность средних, деленная на стандартное отклонение этой разности.

Есть надежда, что эта величина имеет нормальное распределение со средним 0 и дисперсией 1. Так оно и есть, но только при больших объемах выборок!

Для не очень больших выборок распределение величины $t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$ следует распределению Стьюдента.

Это распределение случайной величины, равной

$$t = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \xi_2^2 + \ldots + \xi_k^2)}}$$

, где все ξ_i - нормальны

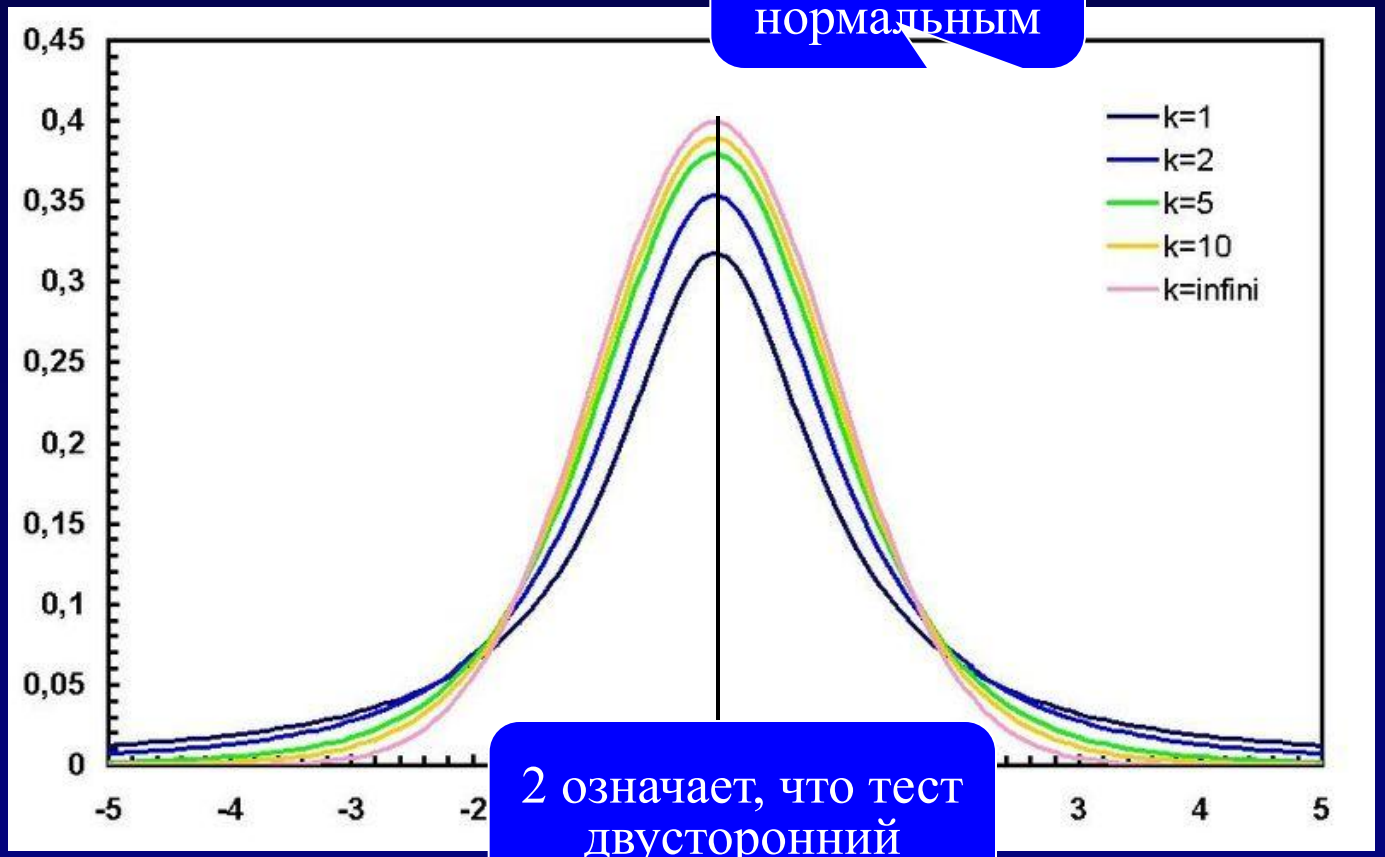
k – число степеней свободы

Вильям Стьюдент (Госсет) (1876-1936)
Работал на пивоваренном заводе Гиннесса
Опубликовал «распределение Стьюдента» в 1908 г.

Сравнение средних

Распределение Стьюдента очень похоже на нормальное, но имеет большую дисперсию: $D(t) = k/(k\text{-}2) > 1$

При $k \to \infty$ становится нормальным



2 означает, что тест двусторонний

Excel умеет вычислять «хвосты» распределения Стьюдента:

0.024 = СТЬЮДРАСП(2; 100; 1)

Сравнение средних

3 варианта использования теста Стьюдента:

- Сравнение выборочного среднего с известным числом

- Сравнение двух зависимых выборок

  Для каждой особи проводят 2 однотипных замера:
  - до и после приема лекарства,
  - в этом году и в прошлом году и т.д.

- Сравнение двух выборочных средних для независимых выборок

  Возможно раного объема

Упражняемся …

15 октября 2011 г. президент Д. Медведев сообщил, что средняя продолжительность жизни в РФ

В этом месяце в районном мор… получена другая оценка: 62±3 года. Отличается … средне… стране?

Эта запись означает, что наша величина имеет распределение Стьюдента с n-1 степенями свободы

2 означает, что тест двусторонний

Вычисляем величину $\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-}{SE}$ … 1)

$P = 0.022$ = СТЬЮДРАСП((69-62)/3; 100-1; 2)

Вывод: нулевая гипотеза отвергается. Вероятность того, что при этом отвергли правильную нулевую гипотезу равна 0.022 (ошибка I рода). Выборка по данным районного морга не соответствует среднему по стране.
Различия статистически значимы.

Никогда не пишите, что различия достоверны!
Достоверно это то, что происходит с вероятностью 1

В данном примере среднее для одной выборки сравнивалось с заранее известной величиной. Это так называемый одновыборочный тест
(мы это уже делали: помните 470 из 1000?)

Сравнение средних

в случае зависимых выборок

Это простой случай. Вычисляется *t*-статистика

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}}$$

и вес хвостов распределения Стьюдента с n_1+n_2-2 степенями свободы.

Можно ни о чем этом не думать и использовать

=TTECT(массив1; массив2; 2; 1)

2 означает, что тест двусторо

1 означает, что выборки зависимы

Для независимых выборок все несколько сложнее…

Сравнение средних
в случае независимых выборок

При сравнении средних двух независимых выборок возможны 2 ситуации:

- $\sigma_1 = \sigma_2$, т.е. изменчивость данных в обеих выборках одинакова

  Тогда все просто: вычисляется статистика $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}} \sim t(n_1 + n_2 - 2)$

- $\sigma_1 \neq \sigma_2$, т.е. изменчивость данных в выборках неодинакова, и эти различия статистически значимы. Тогда вычисляется объединенная дисперсия для двух выборок. Число степеней свободы тоже модифицируется.

Не будем расписывать, как это делается, а запустим Excel

=ТТЕСТ(массив1; массив2; 2; 2)

2 означает, что ~~тест~~ двусторонний

2 - $\sigma_1 = \sigma_2$
3 - $\sigma_1 \neq \sigma_2$

Надо сказать, что Excel не проверяет статистическую значимость $\sigma_1 \neq \sigma_2$, Более адекватно поступает WinStat

Сравнение средних
в случае независимых выборок

При сравнении средних двух независимых выборок возможны 2 ситуации:

- $\sigma_1 = \sigma_2$, т.е. изменчивость данных в обеих выборках одинакова

  Тогда все просто: вычисляется статистика $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}} \sim t(n_1 + n_2 - 2)$

- $\sigma_1 \neq \sigma_2$, т.е. изменчивость данных в выборках неодинакова, и эти различия статистически значимы. Тогда вычисляется объединенная дисперсия для двух выборок. Число степеней свободы тоже модифицируется.

Не будем расписывать, как это делается, а запустим Excel

=ТТЕСТ(массив1; массив2; 2; 2)

2 означает, что ~~тест~~ двусторонний

2 - $\sigma_1 = \sigma_2$
3 - $\sigma_1 \neq \sigma_2$

Надо сказать, что Excel не проверяет статистическую значимость $\sigma_1 \neq \sigma_2$, Более адекватно поступает WinStat

Упражняемся…

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

Считаем t-статистику:

$$t = \frac{4.6 - 3.4}{\sqrt{0.08^2 + 0.07^2}} = 11.3$$

= СТЬЮДРАСП(11,3; 100-2; 2)

Значимо! $P = 10^{-19}$

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

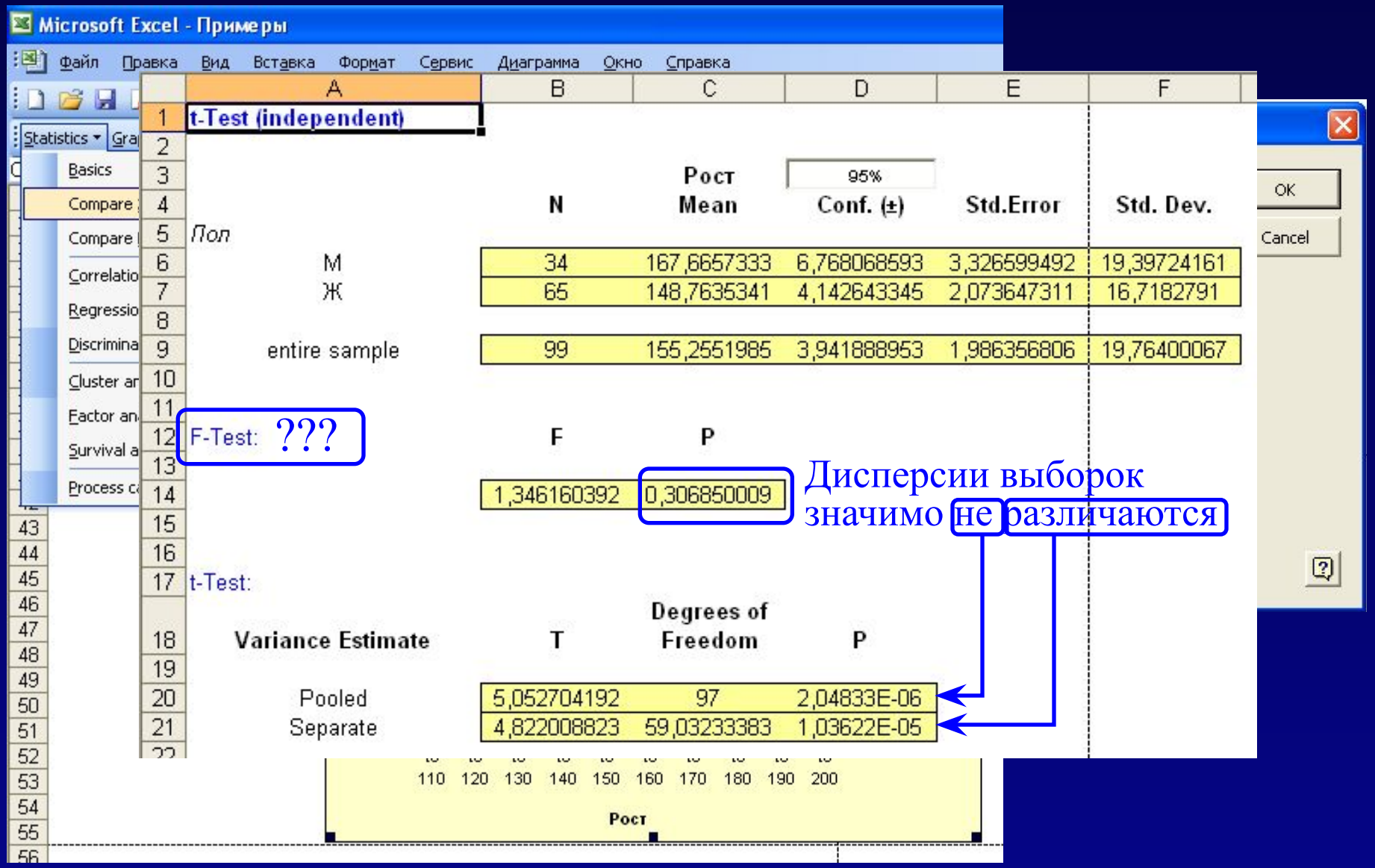
$$SD = \sqrt{0.64} = 0.8 \qquad SE = \frac{0.8}{\sqrt{100}} = 0.08$$

По физкультуре: 4.6±0.1

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

Сравнение средних

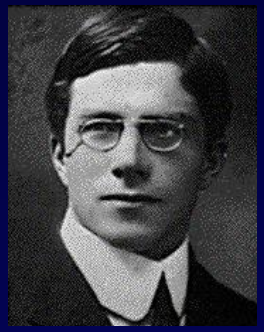
С помощью WinSTAT® the Statistics Add-In for Microsoft® Excel



	A	B	C	D	E	F
1	t-Test (independent)					
2						
3			Рост	95%		
4		N	Mean	Conf. (±)	Std.Error	Std. Dev.
5	Пол					
6	М	34	167,6657333	6,768068593	3,326599492	19,39724161
7	Ж	65	148,7635341	4,142643345	2,073647311	16,7182791
8						
9	entire sample	99	155,2551985	3,941888953	1,986356806	19,76400067
10						
11						
12	F-Test: ???	F	P			
13						
14		1,346160392	0,306850009			
15						
16						
17	t-Test:					
18	Variance Estimate	T	Degrees of Freedom	P		
19						
20	Pooled	5,052704192	97	2,04833E-06		
21	Separate	4,822008823	59,03233383	1,03622E-05		

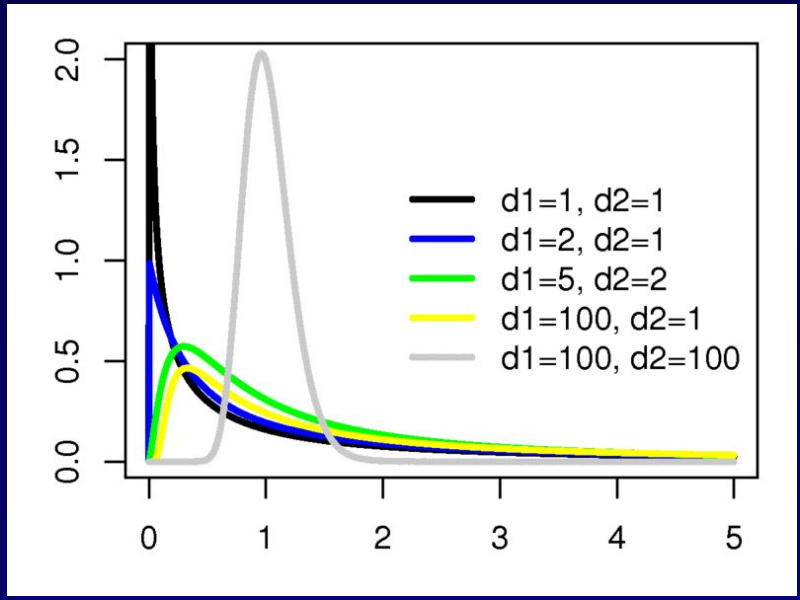
Дисперсии выборок значимо не различаются

Сравнение дисперсий



Р. Фишер построил критерий (односторонний) для сравнения дисперсий (*F*-тест) и вычислил функцию распределения соответствующей статистики.

$$F = \frac{\sigma_2^2}{\sigma_1^2}$$ (большая на меньшую),



В Excel имеется функция, вычисляющая это распределение

=FРАСП(1,5;100;100)

Можно также сравнить дисперсии двух выборок

=ФТЕСТ(массив1; массив2)

$H_0: \sigma_1 = \sigma_2$ против $H_1: \sigma_1 < \sigma_2$

Не путайте статистику (критерий) Фишера с точным тестом Фишера!

Сравнение дисперсий

Дисперсионный анализ (ANOVA) – сравнение нескольких выборок

Рассмотрим набор k выборок:
(при $k = 2$ все сведется к критерию Стьюдента)

	Среднее	Дисперсия
Выборка 1	$\bar{x}_1$	σ_1^2
Выборка 2	$\bar{x}_2$	σ_2^2
…….	…	…
Выборка k	$\bar{x}_k$	σ_k^2
Все выборки	$\bar{x}$	σ^2

Р. Фишер показал, что

$$\sigma^2 = \sigma_W^2 + \sigma_B^2$$

т.е. дисперсию объединенной выборки можно разложить на сумму средней дисперсии внутри выборок (σ_W^2) и межвыборочную дисперсию (σ_B^2):

$$\sigma^2 = \frac{\sum_{i=1}^{k} \sigma_k^2}{k} + \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2}{k}$$

Ничего, кроме школьной алгебры!

Средняя дисперсия

Дисперсия средних

Статистика $F = \dfrac{\sigma_B^2}{\sigma_W^2}$

Внутривыборочная изменчивость

Межвыборочная изменчивость

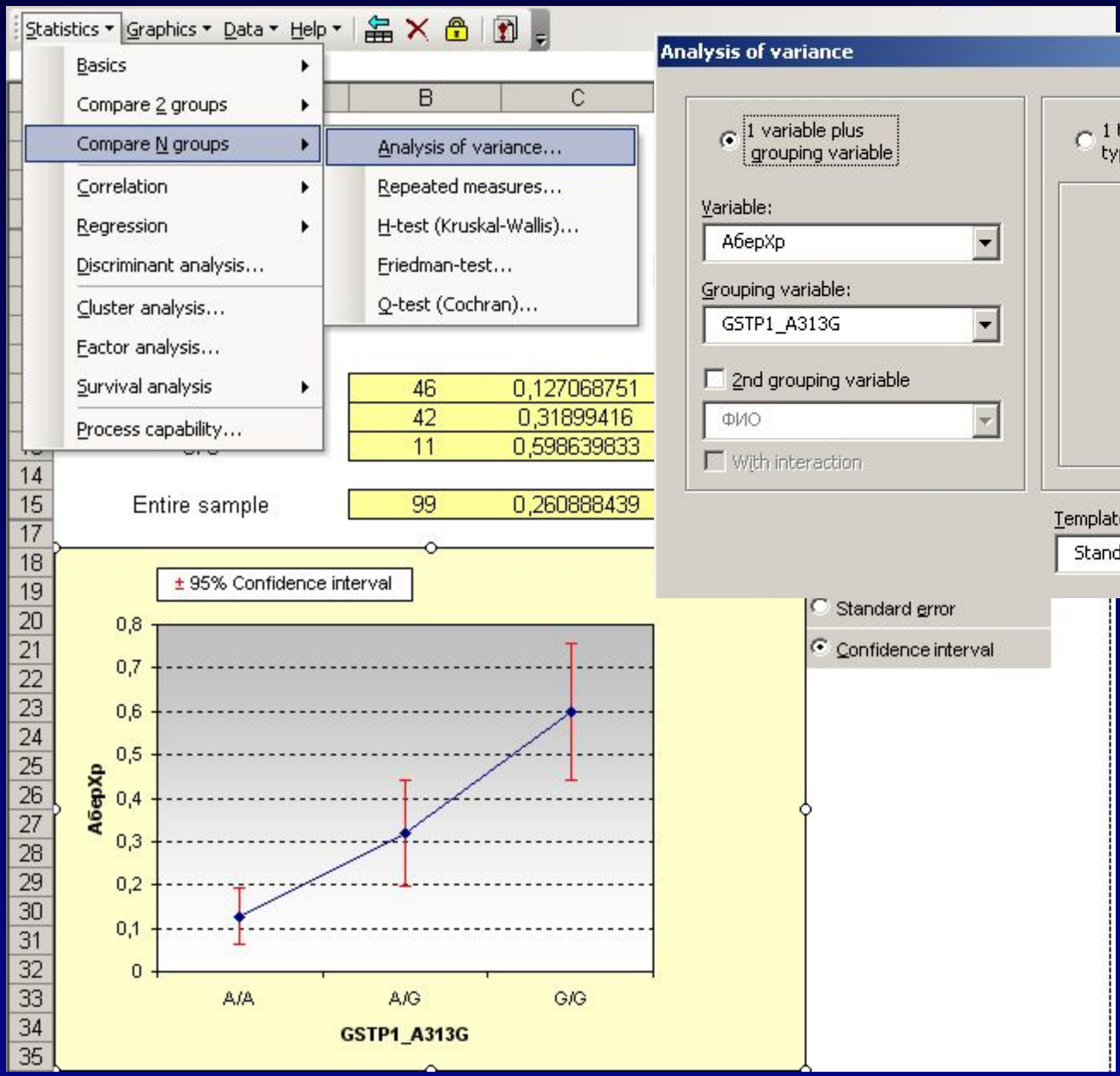
Остаточная изменчивость

Факториальная изменчивость

H_0: $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$

H_1: хотя бы одно среднее отличается

Сравнение нескольких выборок



Сравнение нескольких выборок
Упражняемся…

Для нашей учебной базы данных сравним частоты аберраций хромосом для носителей различных генотипов по локусу *GSTP1*

WinSTAT® the Statistics Add-In for Microsoft® Excel

Analysis of Variance

Variable: АберХр
grouped by: GSTP1_A313G

	Sum of Squares	Degrees of Freedom	Mean Square	F	P
Between Groups	2,220394197	2	1,110197098	11,78092131	2,65812E-05
Within Groups	9,046739097	96	0,094236866		
Total	11,26713329	98	0,114970748		

Межгруп
раз выше,

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
AA	46	5,845163	0,127069	0,048833
AG	42	13,39775	0,318994	0,153666
GG	11	6,585038	0,59864	0,054892

Дисперсионный анализ

очник вари	*SS*	*df*	*MS*	*F*	*P-Значени*	*критическое*
Между гру	2,220394	2	1,110197	11,78092	2,66E-05	3,091191
Внутри гру	9,046739	96	0,094237			
Итого	11,26713	98				

Можно обойтись пакетом

«Анализ данных» в Excel

Важное предупреждение

t-тест (Стьюдента)
F-тест (Фишера)
Дисперсионный анализ
} только для нормально распределенных данных!

В противном случае можно получить совершенно абсурдный результат:

	Фирма 1	Фирма 2
	100	120
	100	120
	100	120
	100	120
	110	120
	110	500
Средние	103.3	183.3

В какой фирме зарплата выше?

=ТТЕСТ(массив1; массив2; 2; 3)

$P = 0.235$ 

Эти средние значимо не различаются по тесту Стьюдента!

На этом примере видно, что в ряде случаев надо сравнивать не сами данные, а их порядковые ранги (номера в последовательности)

Ранговые статистики

Данные

Фирма 1	Фирма 2
100	120
100	120
100	120
100	120
110	120
110	500
Средние 103.3	183.3

Ранги

Фирма 1	Фирма 2
1	7
2	8
3	9
4	10
5	11
6	12
3.5	9.5

0.0002 =ТТЕСТ(массив1; массив2; 2; 2)

Другое дело! Хотя и это некорректно…

Ранговые критерии

Ранговые критерии являются непараметрическими, т.е. такими, которые не зависят от характера распределения данных. В частности они нечувствительны к выбросам отдельных точек

Самый простой тест – критерий знаков для пары зависимых выборок

Плацебо	Лекарство	Разность
105	120	+
110	115	+
120	110	-
103	125	+
115	120	+
121	134	+
107	110	+
114	117	+

1 минус из 8

Приводит ли лекарство к увеличению систолического давления?

0.035 =БИНОМРАСП (1;8;0,5;1)

Различия значимы по одностороннему тесту (но не по двустороннему!)

Ранговые критерии

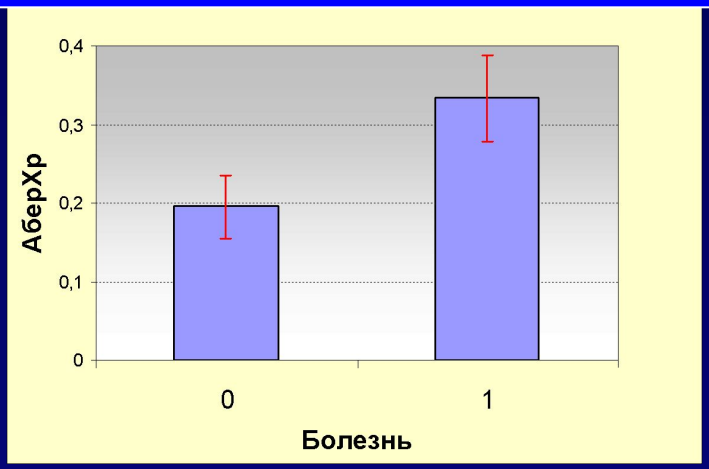
Для сравнения 2 независимых выборок используется тест Манна – Уитни, который основан на вычислении суммы рангов для каждой из выборок

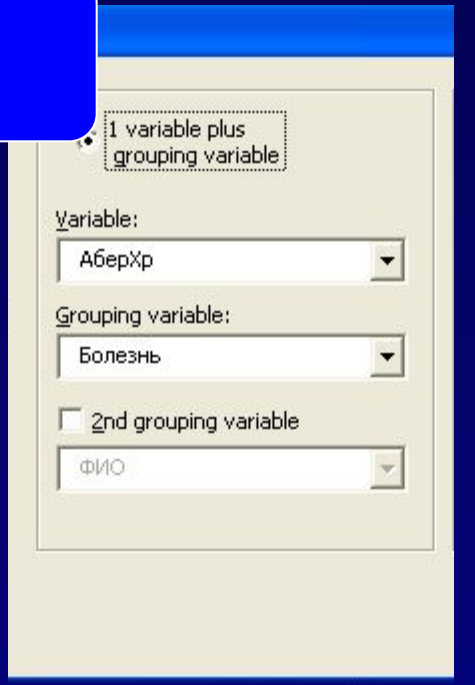
Как всегда H_0: выборки взяты из одной генеральной совокупности.

Упражняемся …

Но что там с нормальностью?

Видим различия средних:





Проверяем значимость различий по Стьюденту:

t-Test:

Variance Estimate	T	Degrees of Freedom	P
Pooled	-2,057004362	97	0,042370006
Separate	-2,030799745	86,41661279	0,045349121

Различия значимы по Стьюденту (независимо от условия равенства дисперсий)

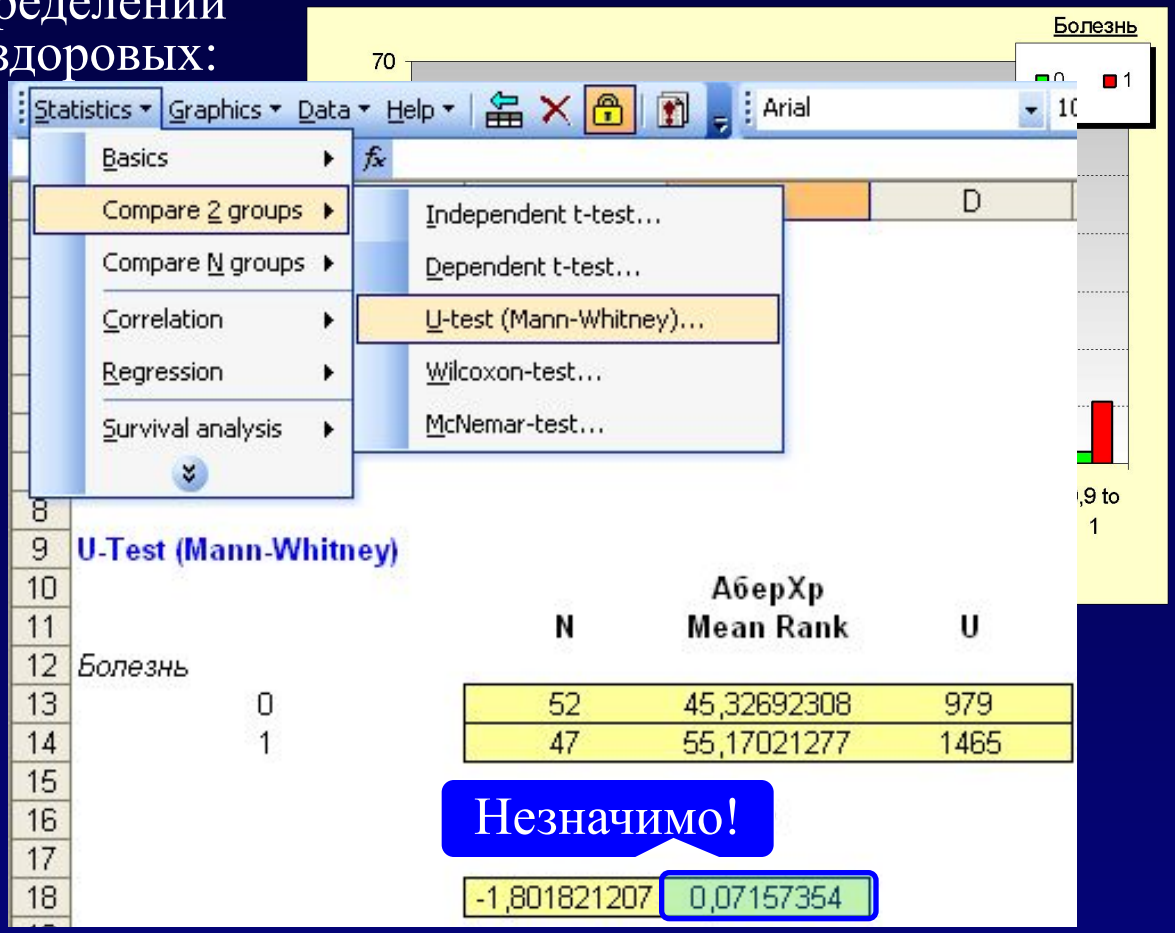
Проверяем нормальность …

Строим гистограммы распределений аберраций для больных и здоровых:

Необходимо использовать непараметрический тест Манна-Уитни

Попробуем все это воспроизвести:

WinSTAT® the Statistics Add-In for Microsoft® Excel



Проверяем значимость различий по Стьюденту:

t-Test:

Variance Estimate	T	Degrees of Freedom	P
Pooled	-2,057004362	97	0,042370006
Separate	-2,030799745	86,41661279	0,045349121

Различия значимы по Стьюденту (независимо от условия равенства дисперсий)

Что значит «незначимо»?

Допустим мы не обнаружили статистическую значимость различий, о чем с грустью сообщаем в публикации. Достаточно ли этого?

НЕТ! Мы должны продемонстрировать, что объемы наших выборок достаточны, чтобы обнаружить эффект, если он существует.

Мощность (чувствительность) используемых тестов должна быть не ниже 80% (тогда упускаем не более 20% открытий)

Только в этом случае незначимые различия можно рассматривать как отрицательный результат

Что значит «незначимо»?

Допустим, что для 2 выборок имеем:

	n	$\bar{x}$	SE	SD
Выборка 1	100	10	1	10
Выборка 1	100	12	1	10

Тогда по тесту Стьюдента различия незначимы и P = 0.159

Compare2/ Numerical observations/ Normal distributin/mean value

Проверим мощность данного теста

Compare2/ Power/ Comparison of means
Size A - 100 Size B – 100
DETECT a difference 2

$\Rightarrow$ Мощность всего 29% !

т.е. доля упущенных открытий более 70% !

О чем мы обязаны сообщить в публикации (правда биологи этого почти никогда не делают)

Чтобы выйти на мощность 80% объемы выборок должны быть 400 и 400

Compare2/ Sample size/ Means

На сегодня это все

Напоследок хочу посоветовать:

- Проверяйте характер распределения сравниваемых величин. Или хотя бы стройте гистограммы распределений – для себя.

- Поставьте на свой компьютер WinStat и постройте пример использования дисперсионного анализа
- На всякий случай проверяйте значимость различий параметрическими и непараметрическими методами.
- Оценивай мощность теста в случае получения незначимых результатов