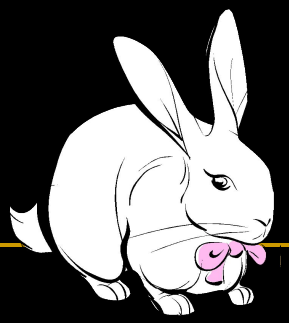
Занятие 3

Мощность статистического теста. Дисперсионный анализ ANOVA



Мощность - вероятность <u>отвергнуть</u> H_0 в эксперименте, когда H_0 действительно <u>неверна</u>.

	Истинное (но неизвестное нам) положение дел	
	Верна H_0	Верна H_1
Мы «приняли» H_0	ПРАВИЛЬНО!	ОШИБКА 2-го рода = β
Мы отвергли H_0	ОШИБКА 1-го рода = α	ПРАВИЛЬНО! **мощность** критерия = $1-\beta$

Мощность

Т.е., масса землероек в Заповеднике на самом деле **больше**, чем 90 г.

Мощность – вероятность того, что проведённое нами исследование установит этот факт.

H_0: $\mu \leq 90$ г;

H_1 : $\mu > 90$ г

Ошибка 2-го рода + мощность = 1

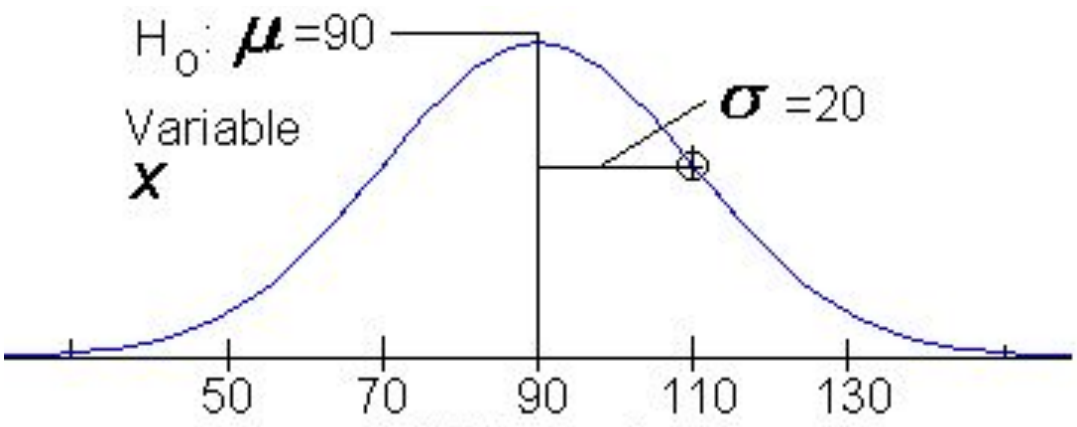
$$\beta + (1-\beta) = 1$$

(это 2 возможных результата теста, если H_0 не верна)

Мощность

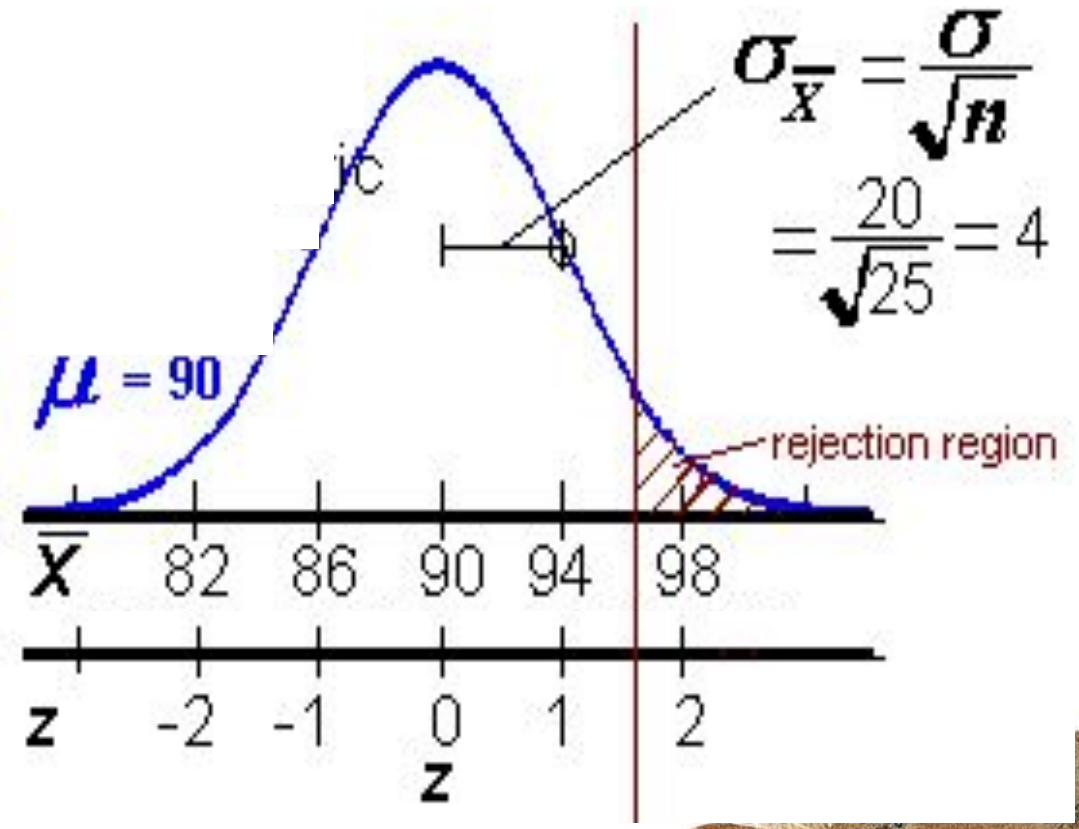
Мощность предполагаемого статистического теста - ключевой элемент планирования исследования

«Реальное значение» параметра:
Во всей мировой популяции землероек μ = 90 г.
Пусть «реальное значение» средней массы в заповеднике = 94 г.



Мощность

Нарисуем распределения выборочных средних для $\mu = 90$ и $\mu = 94$ (стандартное отклонение $\sigma = 20$).

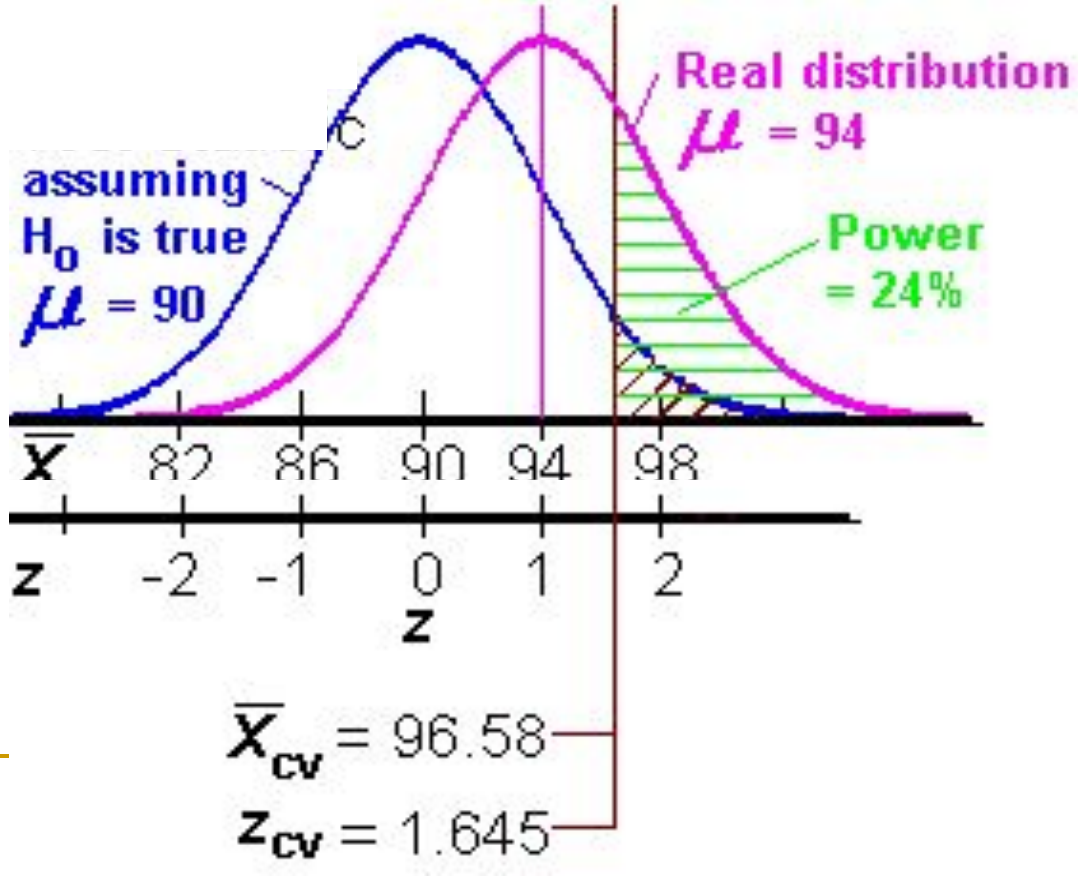


Размер выборки n = 25 зверей

Мощность

Если мы поймаем 25 землероек в заповеднике, у нас есть вероятность лишь 24%, что мы найдём различия!

Т.к. лишь в 24% случаев среднее из нашей будущей выборки попадёт в критическую область.



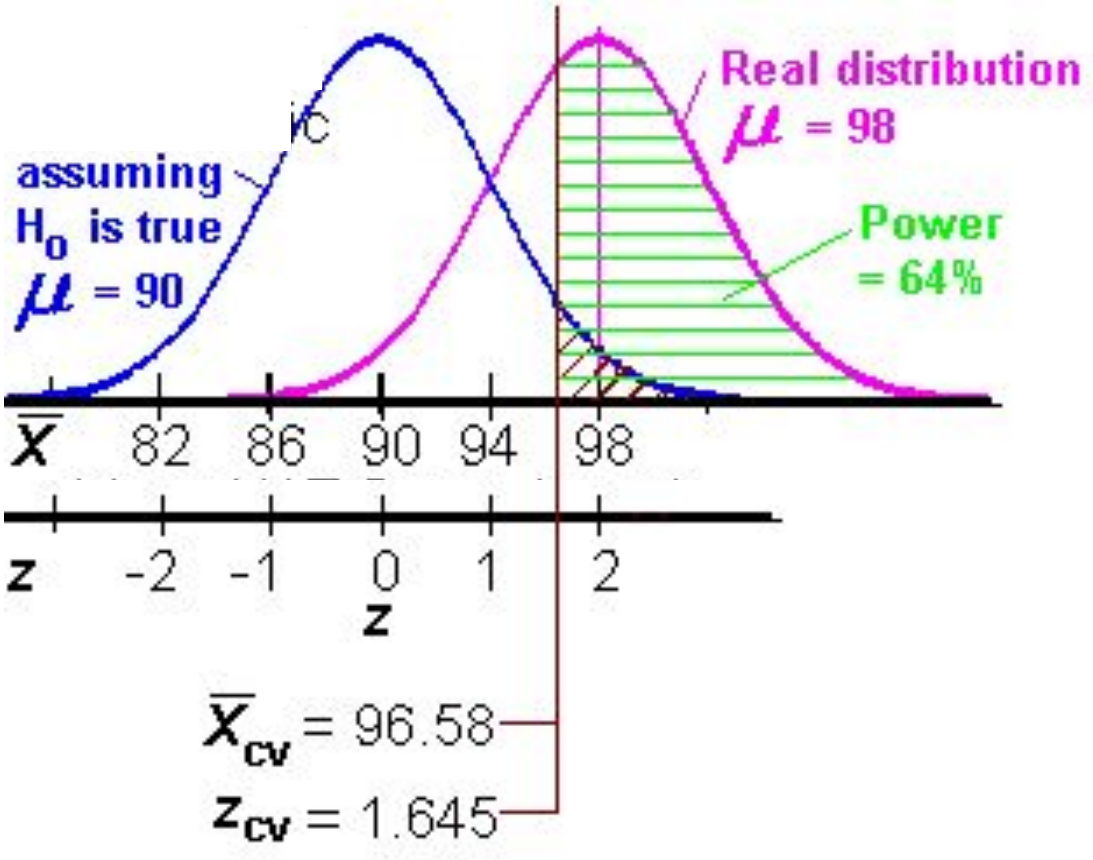
Как увеличить мощность?

Большей МОЩНОСТИ критерия способствуют:

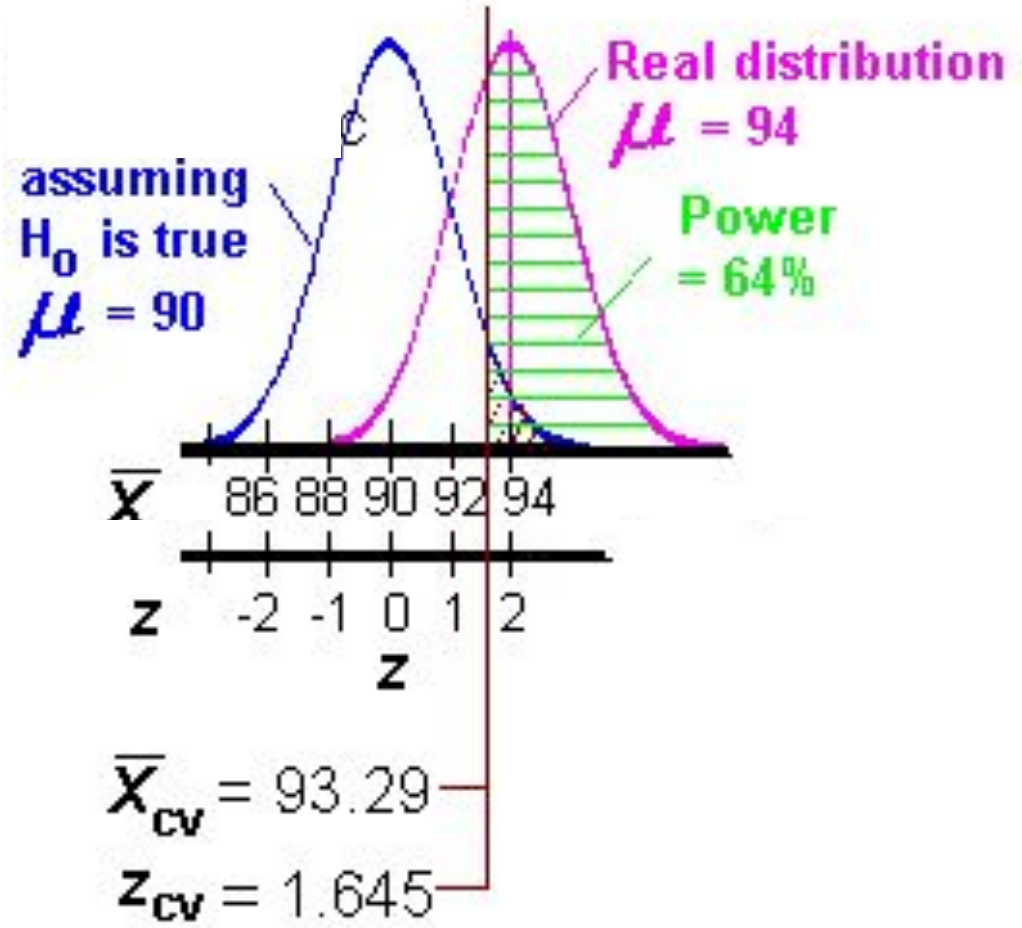
1. Большой размер выборки;
2. Большие различия между популяциями (effect size);
3. Маленькое стандартное отклонение;
4. Большой уровень значимости (α=0.05 а не α=0.01);
5. Выбор одностороннего теста вместо двустороннего

Мощность

Если в действительности средняя масса землероек в заповеднике равна 98 г, мощность теста будет уже 64%.



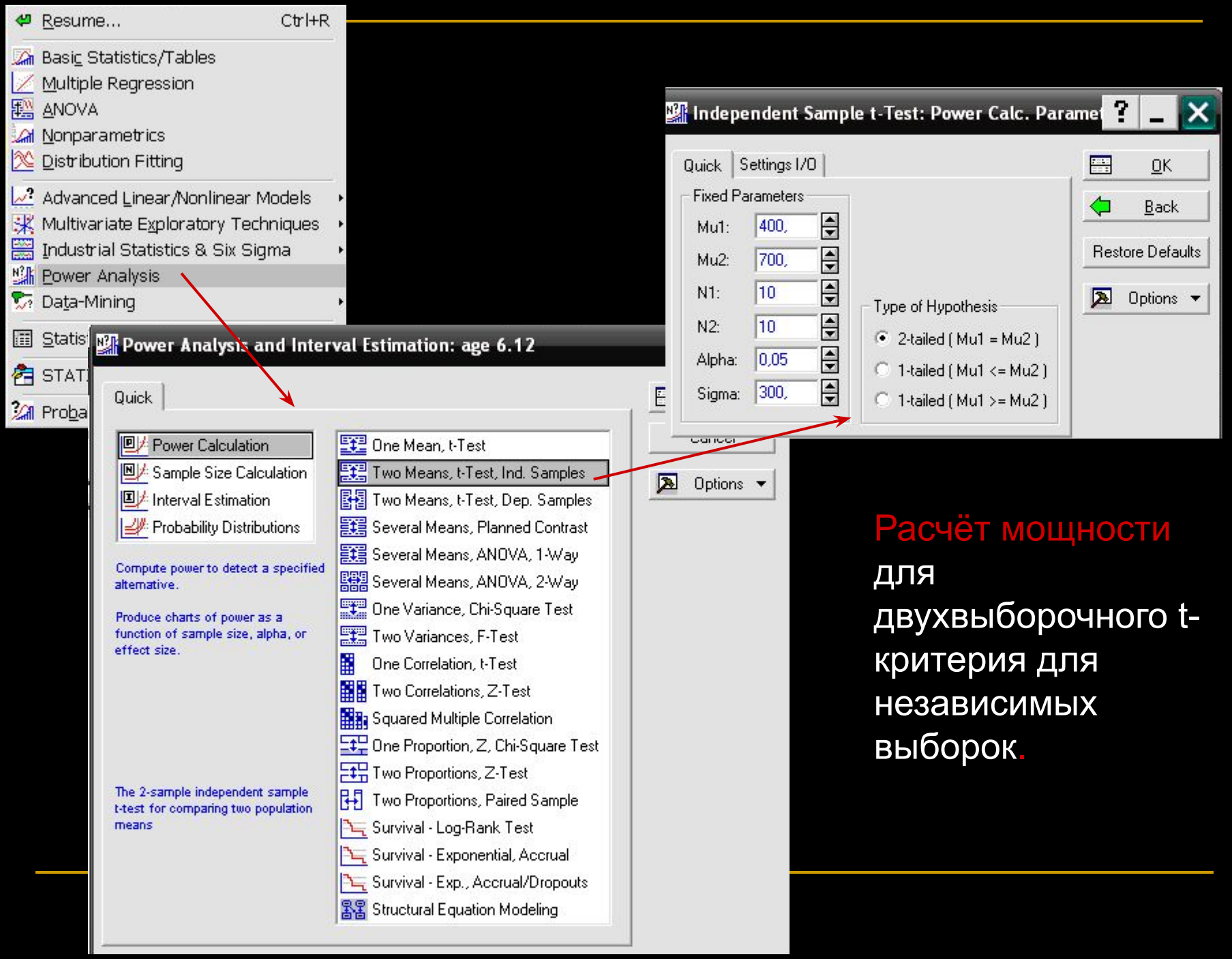
Здесь стандартное отклонение уменьшили вдвое, и мощность теста тоже стала 64%.



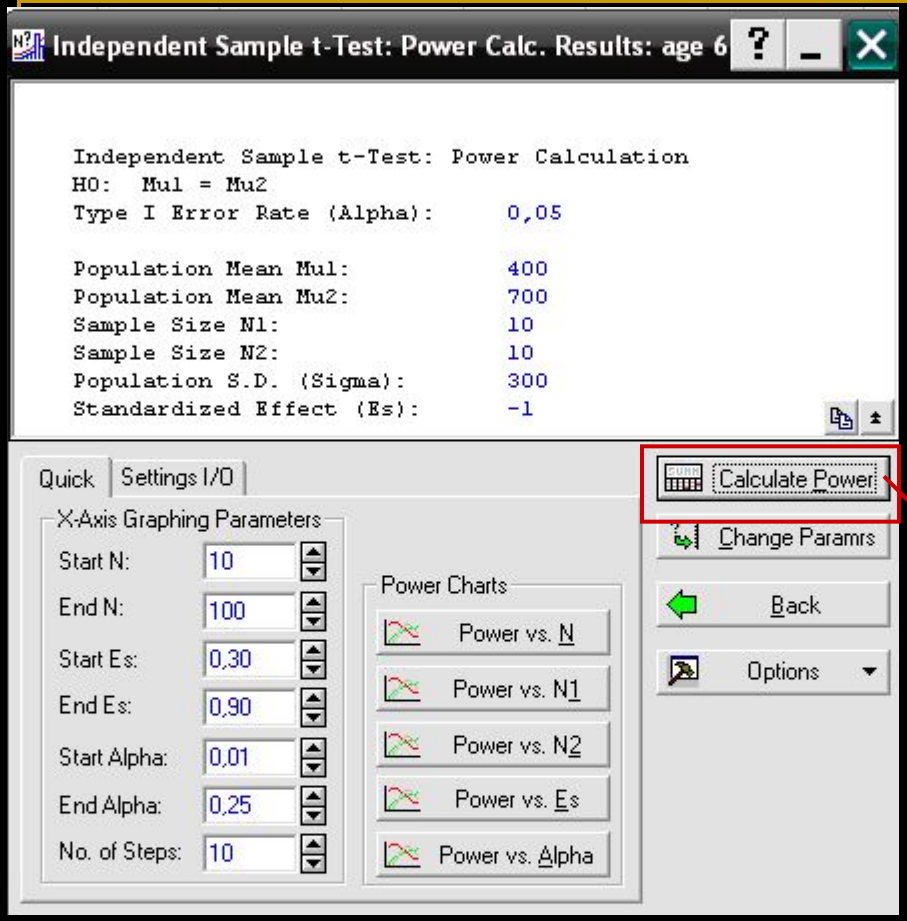
Как использовать понятие мощности критерия:

При **планировании** исследования мы можем рассчитать размер выборки, необходимый для того, чтобы «разглядеть» предполагаемые различия между выборками.
(Реальные различия нам, очевидно, неизвестны, но можно задать минимальные, имеющие биологическое значение).

Ещё мы можем **после** проведения теста (в котором мы не отвергли H_0) оценить вероятность ошибки (2-го рода).



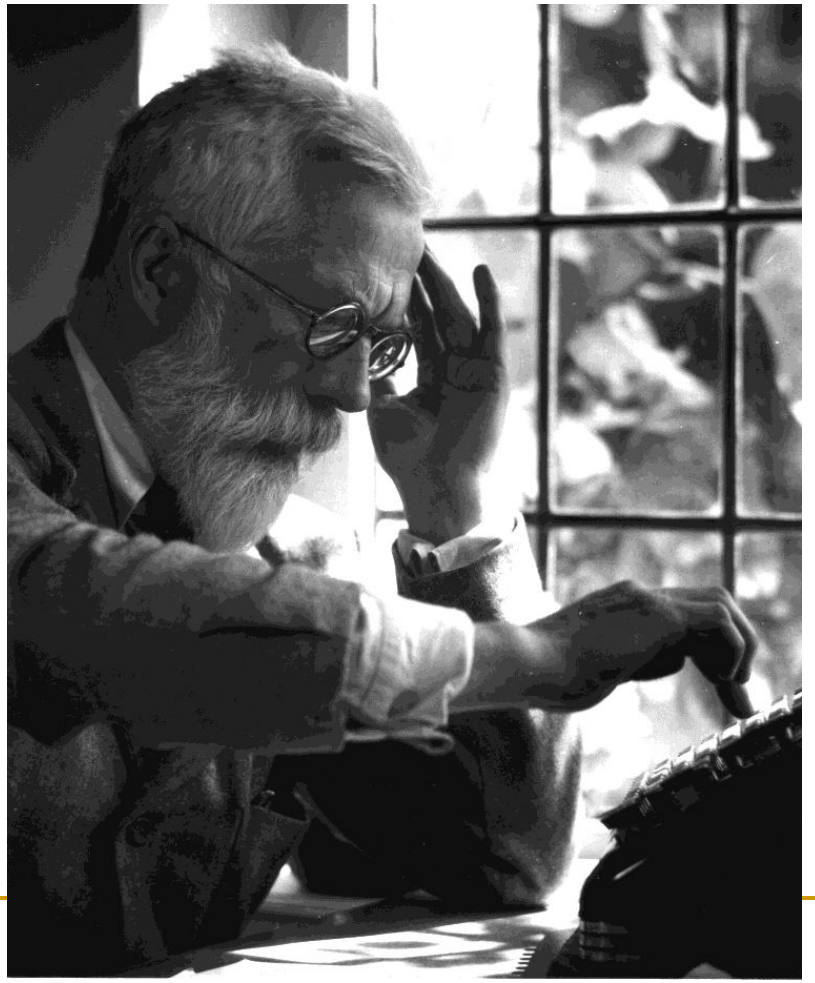
Расчёт мощности для двухвыборочного t-критерия для независимых выборок.



	Power Calculation (age 6.12) Two Means, t-Test, Ind. Samples H0: Mu1 = Mu2
	Value
Population Mean Mu1	400,0000
Population Mean Mu2	700,0000
Population S.D. (Sigma)	300,0000
Standardized Effect (Es)	-1,0000
Sample Size N1	10,0000
Sample Size N2	10,0000
Type I Error Rate (Alpha)	0,0500
Critical Value of t	2,1009
Power	0,5620

Сравнение **ДВУХ** И **БОЛЕЕ** групп

Дисперсионный анализ ANOVA (analysis of variance)

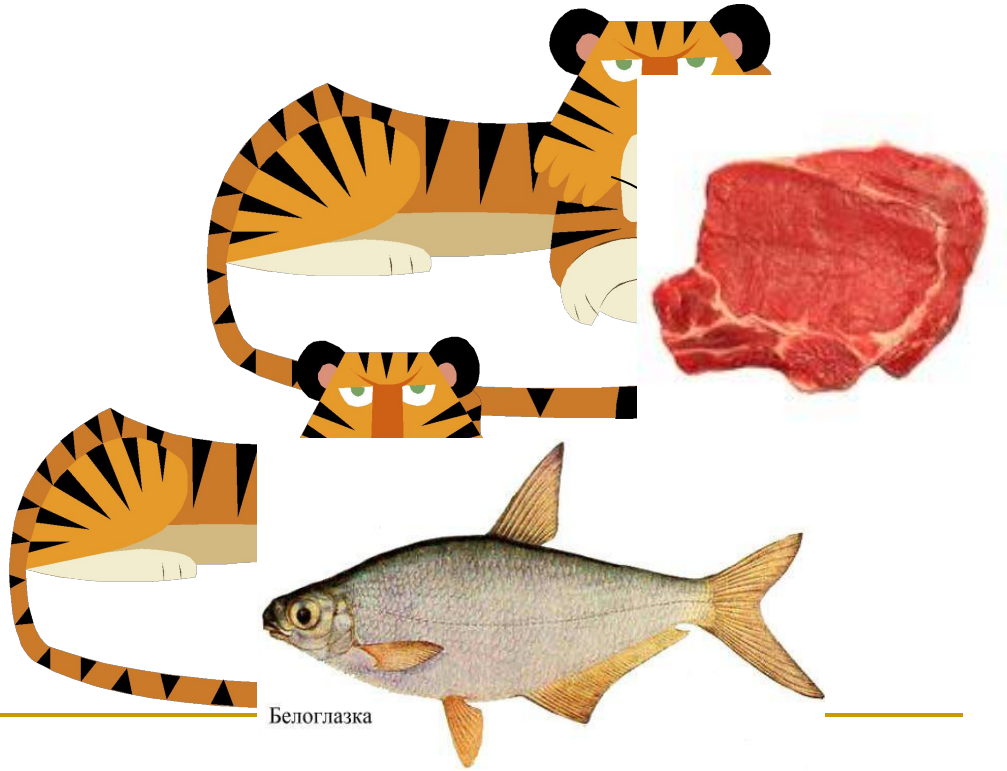


Sir Ronald Aylmer
FISHER

ANOVA

Мы тестировали гипотезы о среднем значении для одной и двух выборок.
Как быть, если выборок **три или больше**?

Предположим, у нас 4 группы тигров, которых кормят по-разному. Различается ли средняя масса тигра в этих группах?



Формулируем гипотезу H_0:

Тигров кормили:

1. овощами;
2. фруктами;
3. рыбой;
4. мясом.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Это сложная гипотеза (omnibus hypothesis). Она включает в себя много маленьких гипотез (для 3-х групп – 3, для 4-х – 12 …):

$$H_{01} : \mu_1 = \mu_2$$
$$H_{02} : \mu_1 = \mu_4$$
$$H_{03} : \mu_1 = \mu_3$$
$$H_{04} : \mu_2 = \mu_3$$
$$H_{05} : \mu_2 = \mu_4$$
$$H_{06} : \mu_3 = \mu_4$$

Парные (pairwise) нулевые гипотезы

$$H_{07} : \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$$

$$H_{08} : \mu_1 = \frac{\mu_2 + \mu_3 + \mu_4}{3}$$

…

Комплексные (complex) нулевые гипотезы

Зависимая переменная: масса;
Независимая (группирующая) – тип еды.

Формулируем альтернативную гипотезу:

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \quad ?$$

НЕВЕРНО!

H_1: $\mu_1 \neq \mu_2$ или $\mu_1 \neq \mu_3$ или $\mu_1 \neq \mu_4$ …

Мы отвергаем общую H_0 гипотезу если верна хотя бы одна из маленьких частных альтернативных гипотез (парных или комплексных)!
Какая именно – ANOVA не говорит.

ANOVA

Почему бы не сравнить группы попарно *t*-критерием? (Ошибка использования критерия Стьюдента)

1. мы таким образом проверяем не все гипотезы, которые содержатся в сложной гипотезе;
2. резко увеличивается вероятность найти различия, где их нет!! (общая вероятность ошибки 1-го рода).

Эффект МНОЖЕСТВЕННЫХ СРАВНЕНИЙ (при попарном сравнении нескольких групп).

При уровне значимости α=0,05 вероятность ошибиться в хотя бы в одном из k сравнений примерно равна:

$$P_{ошибки}=1-(1-0,05)^k$$

Например, для попарного сравнения 4-х групп k=6, т.е., $P_{ошибки}=1-(1-0,05)^6 = 0,265$

($P_{ошибки}$~0,05k)

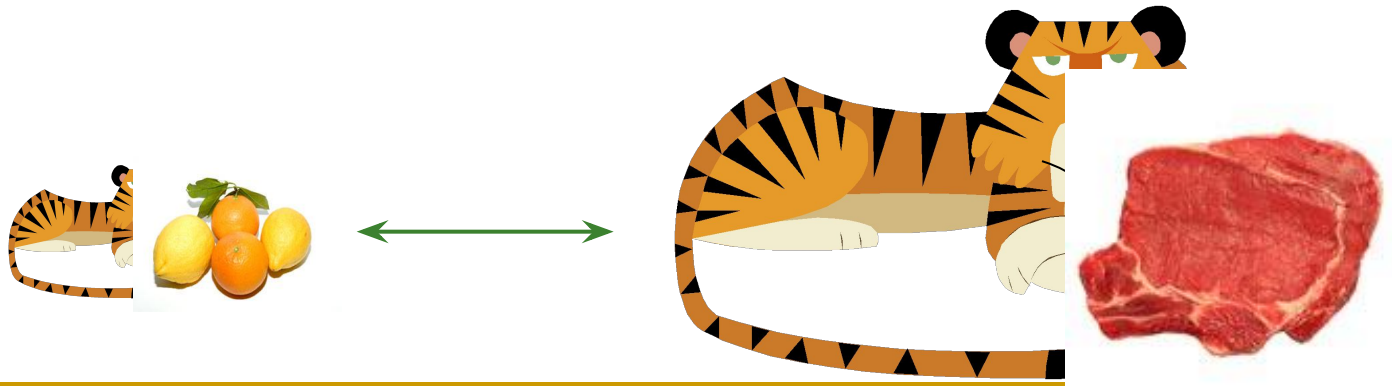
Общая логика ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$ (т.е., средние в 4-х популяциях равны)

Формируем 4 независимых случайных выборки и считаем выборочные средние для каждой из них (они оценивают популяционные средние).

Если H_0 верна, выборочные средние должны быть **примерно** (насколько примерно?) одинаковы.

Чем дальше друг от друга отстоят средние значения в группах, тем меньше вероятность, что верна H_0



В t-тесте сходство выборочных средних оценить легко – просто посчитать разность. Но с 3-мя (4, 5...) группами так не получится!

ANOVA

Пусть все группы будут одинакового размера (для простоты объяснения).

Если H_0 верна, то 4 наших группы получены из ОДНОЙ популяции с конкретными средним μ и дисперсией σ^2.
Получим 2 независимые оценки σ^2 и сравним их!
На этой идее основана АНОВА.

овощи	фрукты	мясо	рыба
151	108	147	130
135	94	138	128
137	84	143	140
118	87	135	142
132	82	153	139
135	79	137	145
131	74	148	144
137	73	140	140
121	67	144	141
140	78	146	140
152	63	151	142
133	90	145	137
151	81	146	148
132	96	147	142
139	83	150	143
96	89	144	140
133,7	***83***	***144,6***	***140,1***

1. Оценка общей дисперсии по разбросу МЕЖДУ группами

средние в каждой группе

общее среднее

$$MS_B = s^2_{\bar{X}} n = \frac{\sum \left(\bar{X}_j - \bar{X}_G\right)^2}{k-1} n$$

$df_B = k\text{-}1$

число групп -1 (3 - 1 = 2)

размер группы

MS_B – **m**ean **s**quare between groups, оценка расстояния между средними в группах.

различия большие - H_0 не верна

ANOVA

овощи	фрукты	мясо	рыба
151	108	147	130
135	94	138	128
137	84	143	140
118	87	135	142
132	82	153	139
135	79	137	145
131	74	148	144
137	73	140	140
121	67	144	141
140	78	146	140
152	63	151	142
133	90	145	137
151	81	146	148
132	96	147	142
139	83	150	143
96	89	144	140
133,7	***83***	***144,6***	***140,1***

2. Оценка общей дисперсии по разбросу ВНУТРИ групп

сумма квадратов стандартных отклонений внутри групп

$$MS_W = \frac{s_1^2 + s_2^2 + s_3^2 + \ldots + s_k^2}{k}$$

число групп

$$df_W = n_G - k$$

статистика:

$$F = \frac{MS_B}{MS_W}$$

Статистика критерия: F

$$F = \frac{\text{оценка дисперсии } \textbf{между} \text{ группами}}{\text{оценка дисперсии } \textbf{внутри} \text{ групп}} \qquad F = \frac{MS_B}{MS_W}$$

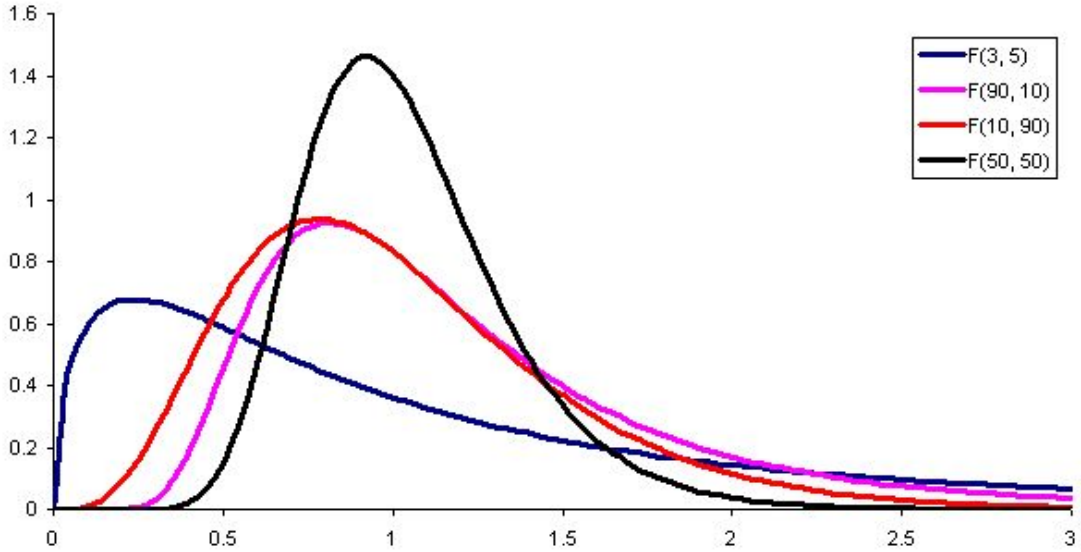
✔Не соответствует общей формуле

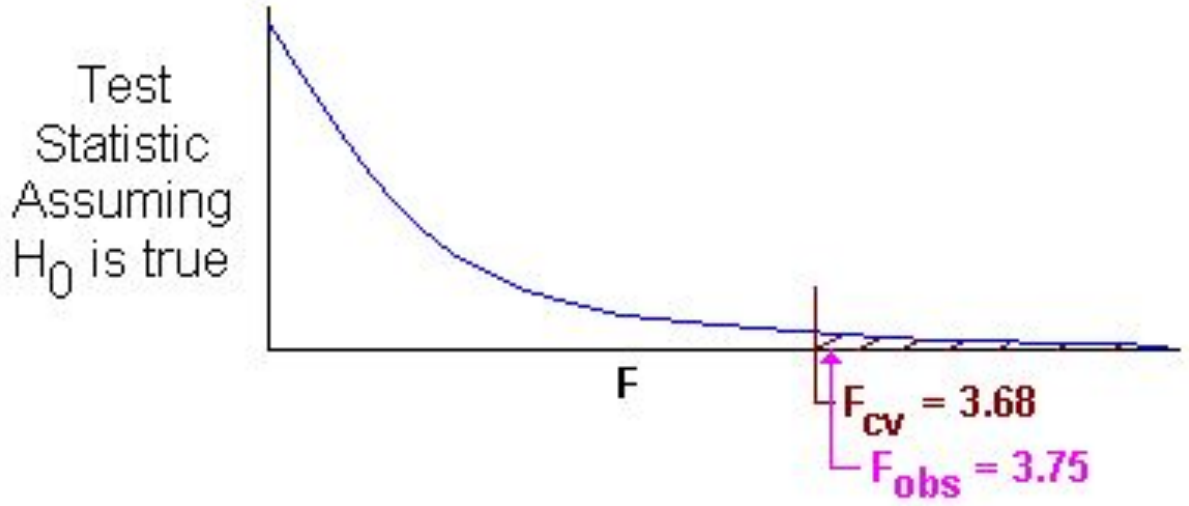
$$\text{Статистика} = \frac{\text{параметр } \textbf{выборки} - \text{параметр } \textbf{популяции}}{\text{стандартная } \textbf{ошибка} \text{ параметра выборки}}$$

✔Приводится как F_{df_B, df_W} т.е., например, $F_{3,60}$

Статистика критерия: F

Принципиально ненаправленный (двусторонний) тест





ANOVA table

источник изменчивости	***SS***	***df***	***MS***	***F***
между	SS_B	df_B	MS_B	F
внутри	SS_W	df_W	MS_W	
общее	SS_T	df_T		

$$F = \frac{MS_B}{MS_W}$$

SS это суммы квадратов отклонений (sum of squares) :
SS_B - средних в группах от общего среднего = **Effect**;
SS_W – измерений от средних в группах = **Error**.

$$SS_T = \sum (X_{ij} - \overline{X_G})^2 = \sum (X_{ij} - \overline{X_j})^2 + \sum (\overline{X_j} - \overline{X_G})^2 = SS_W + SS_B$$

$$df_T = df_W + df_B$$

$$MS_B = \frac{SS_B}{df_B} \qquad MS_W = \frac{SS_W}{df_W}$$

ANOVA effect size

«Практическая значимость» результата:

1. $$f = \frac{s_{\bar{X}}}{\sqrt{MS_W}}$$

f = 0.1 – маленький эффект
f = 0.25 – средний эффект
f = 0.4 – большой эффект

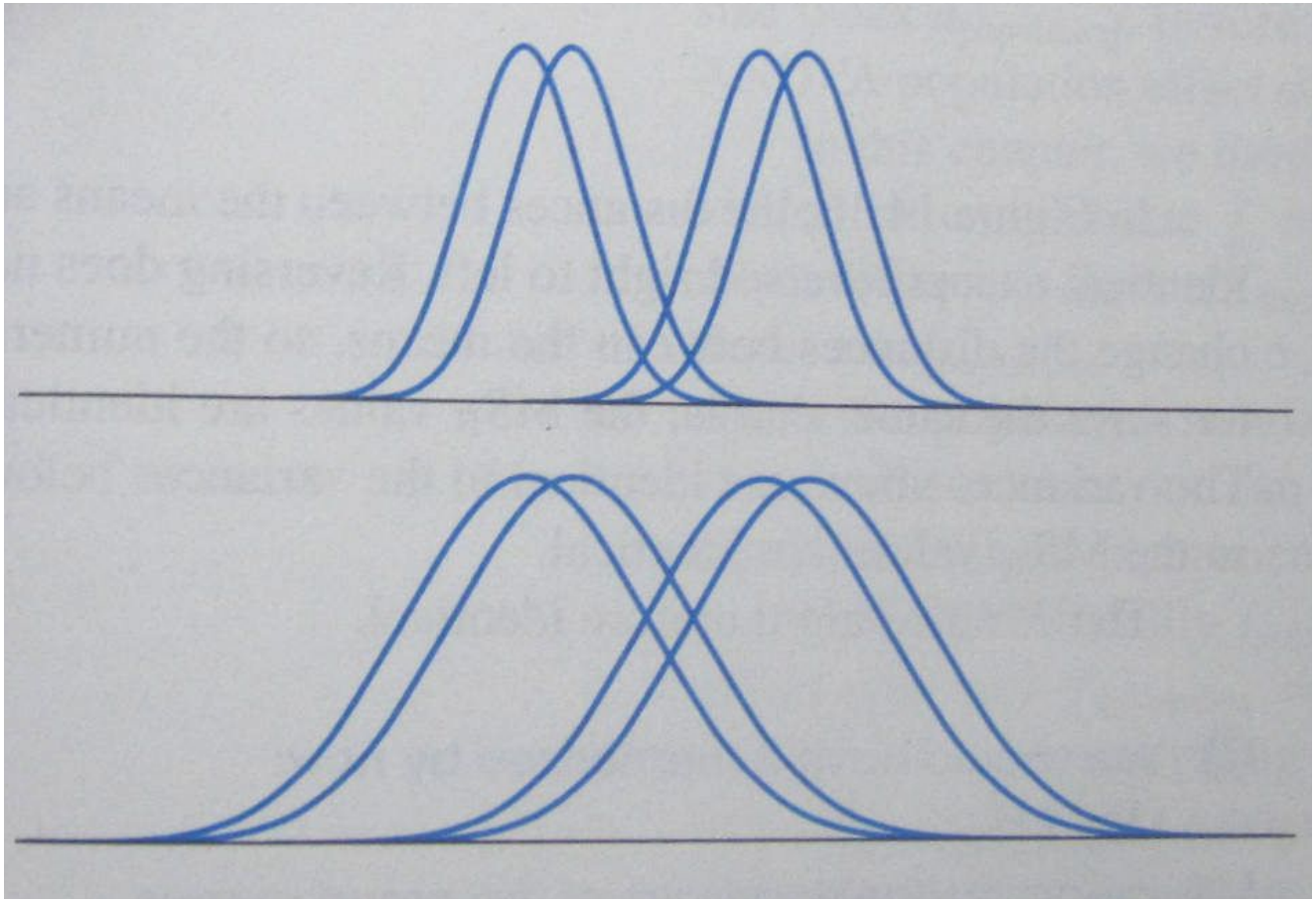
2. «доля объяснённой изменчивости»

$$R^2 = \frac{SS_B}{SS_T}$$

R^2 = 0.01 – маленький эффект
R^2 = 0.06 – средний эффект
R^2 = 0.14 – большой эффект

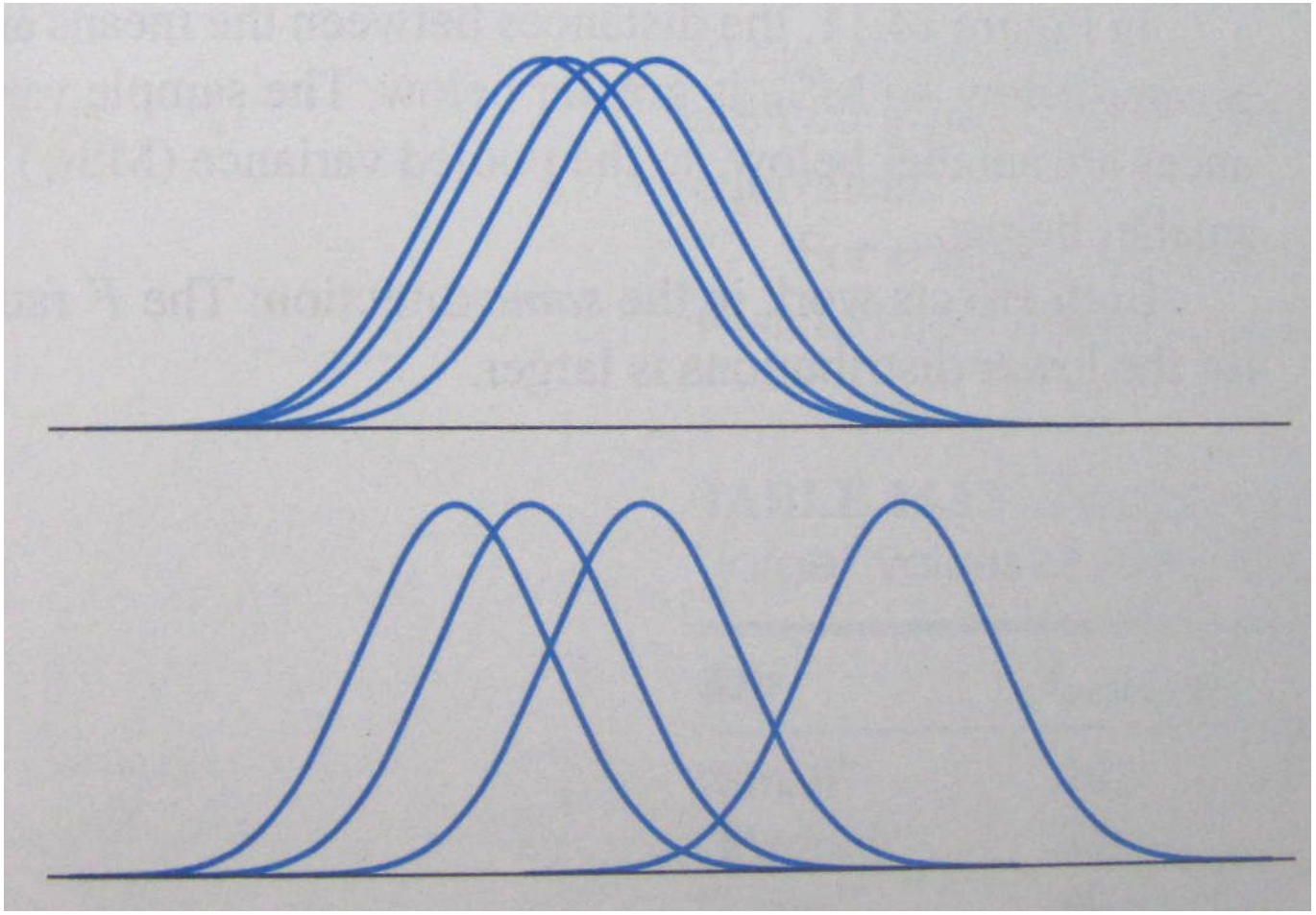
ANOVA

В каком случае значение F-статистики будет больше?



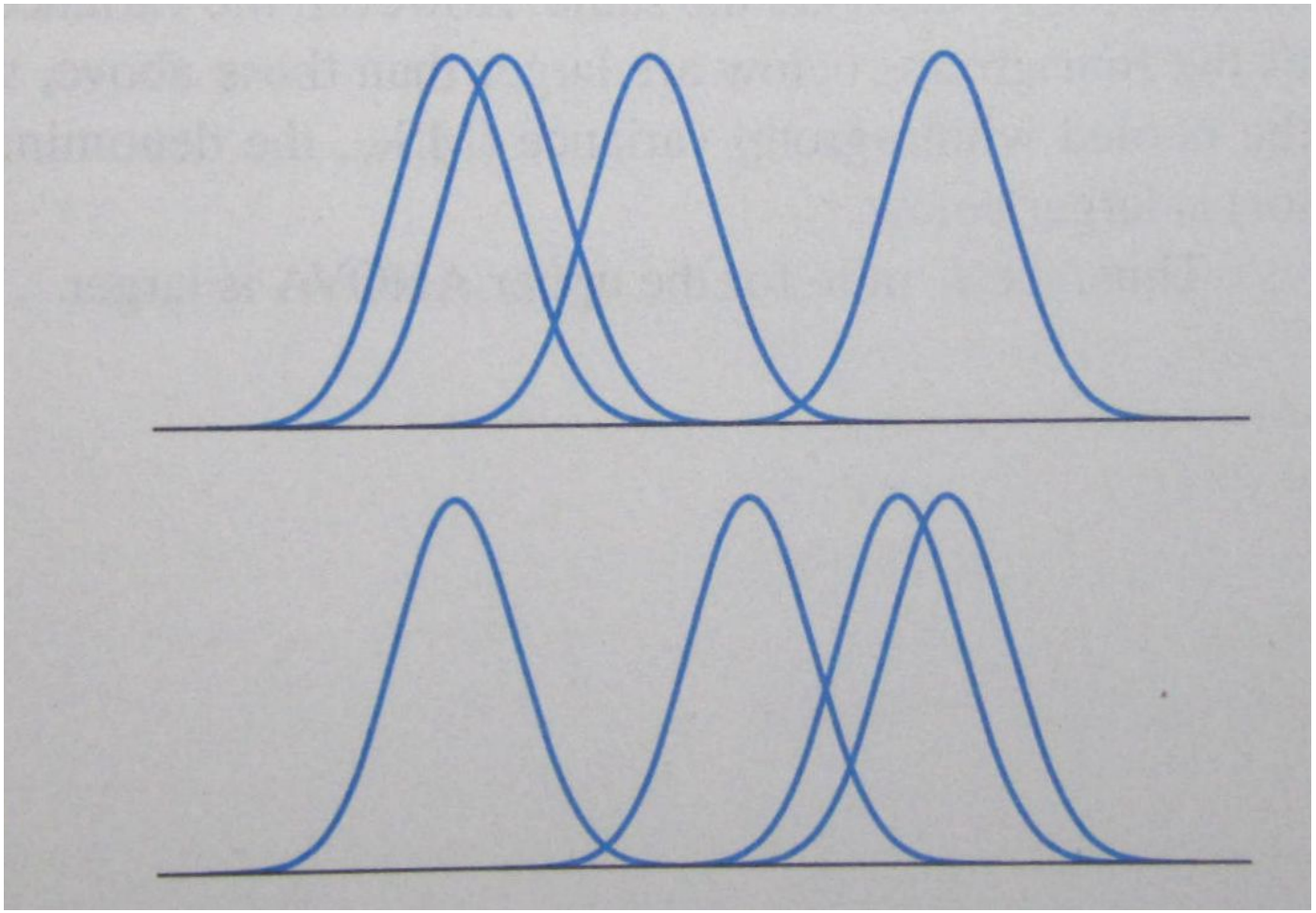
ANOVA

В каком случае значение F-статистики будет больше?



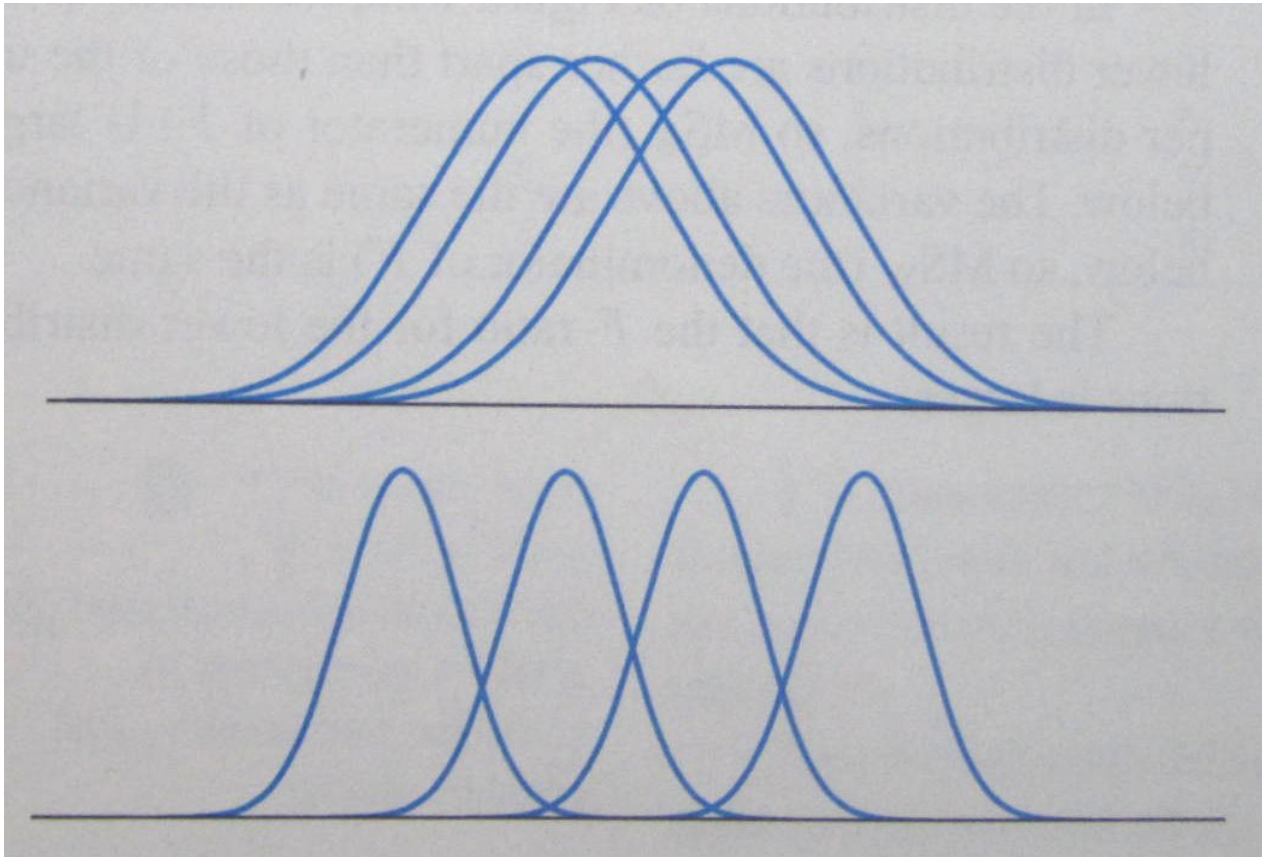
ANOVA

В каком случае значение F-статистики будет больше?



ANOVA

В каком случае значение F-статистики будет больше?

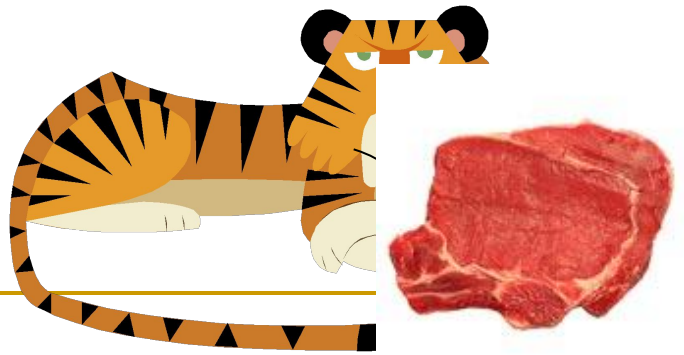


ANOVA

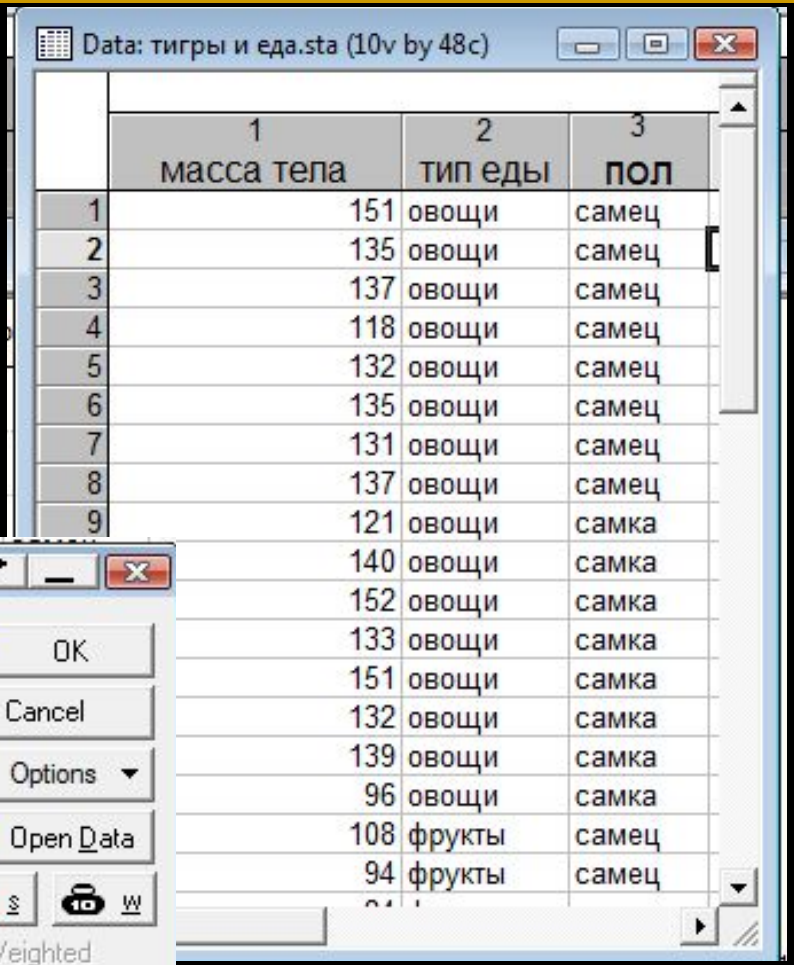
У нас только одна независимая (группирующая) переменная.
Такой анализ называется

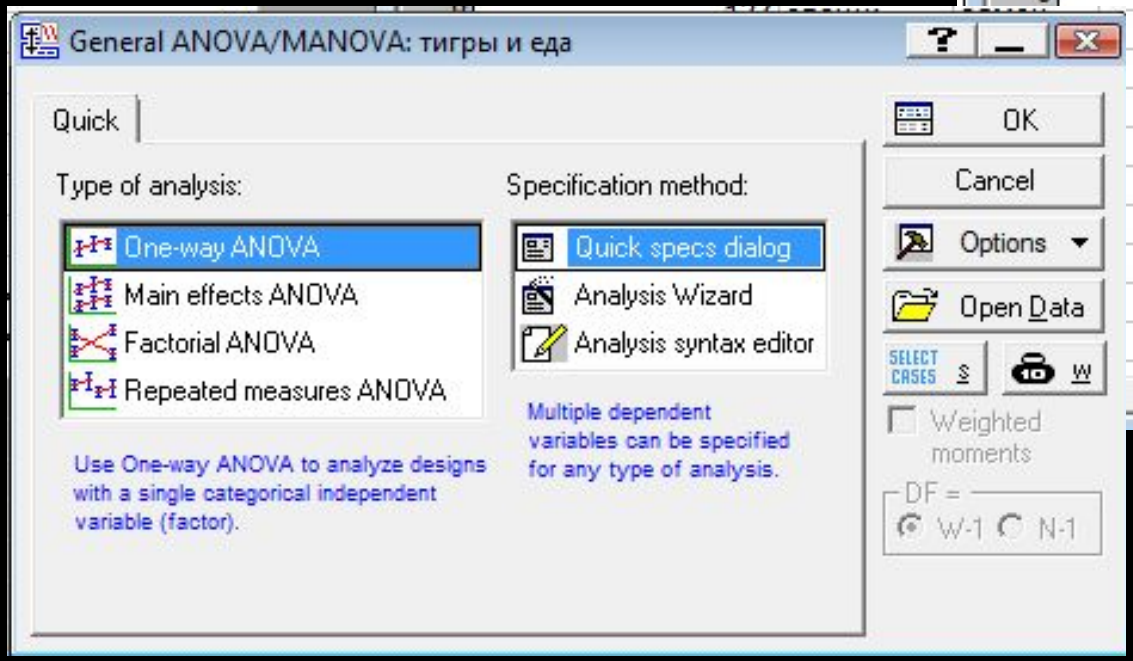
One-way ANOVA

требования и ограничения – как в критерии Стьюдента

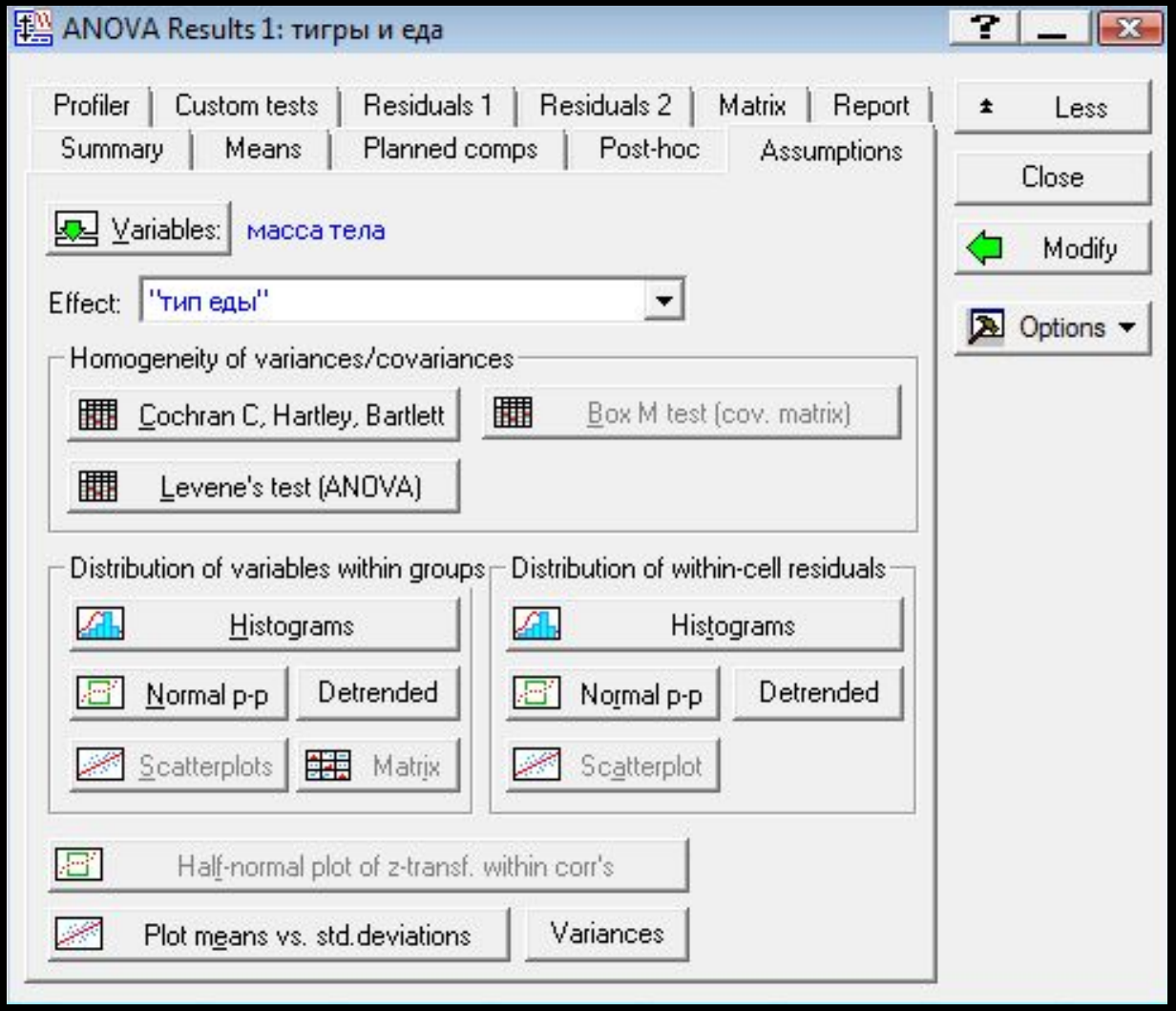


One-way ANOVA

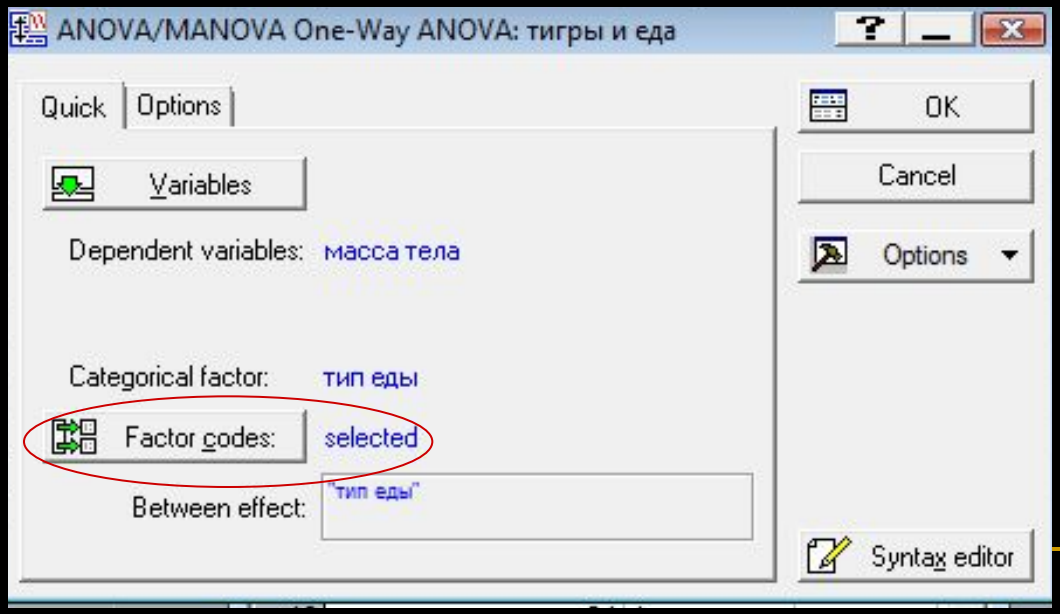


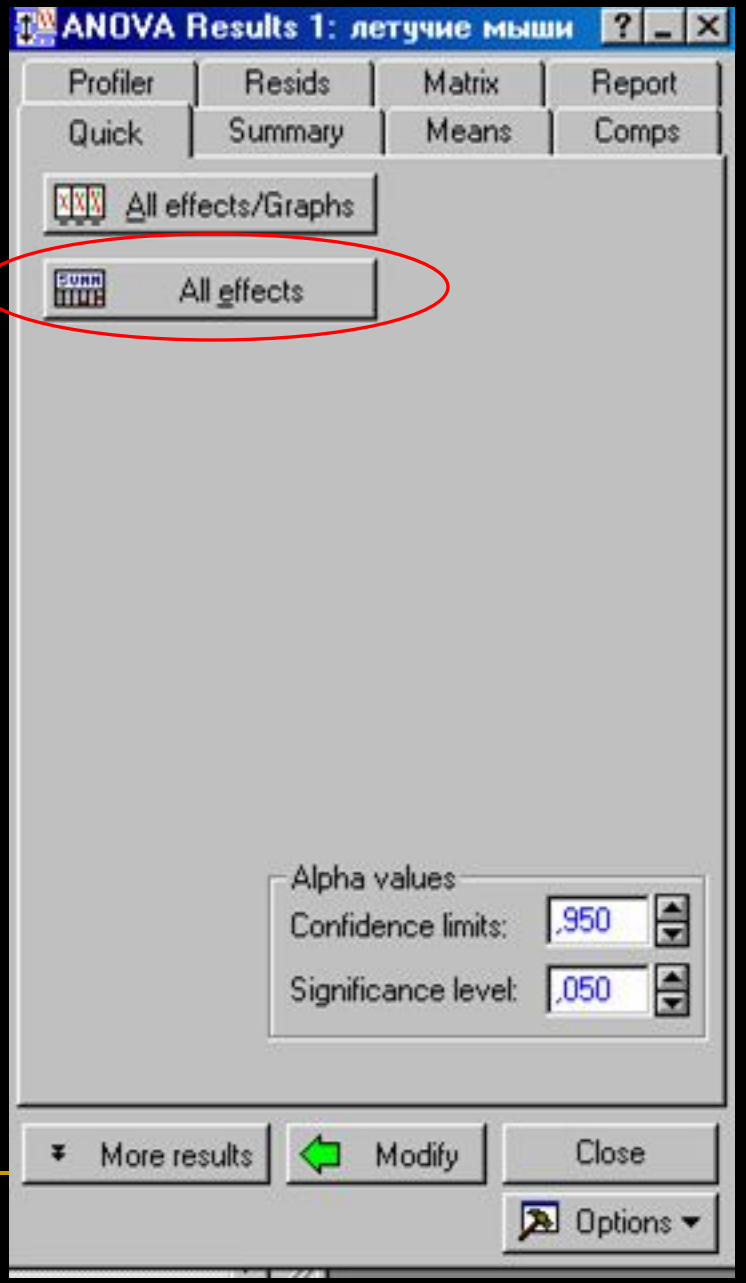


assumptions: нормальность, гомогенность



One-way ANOVA





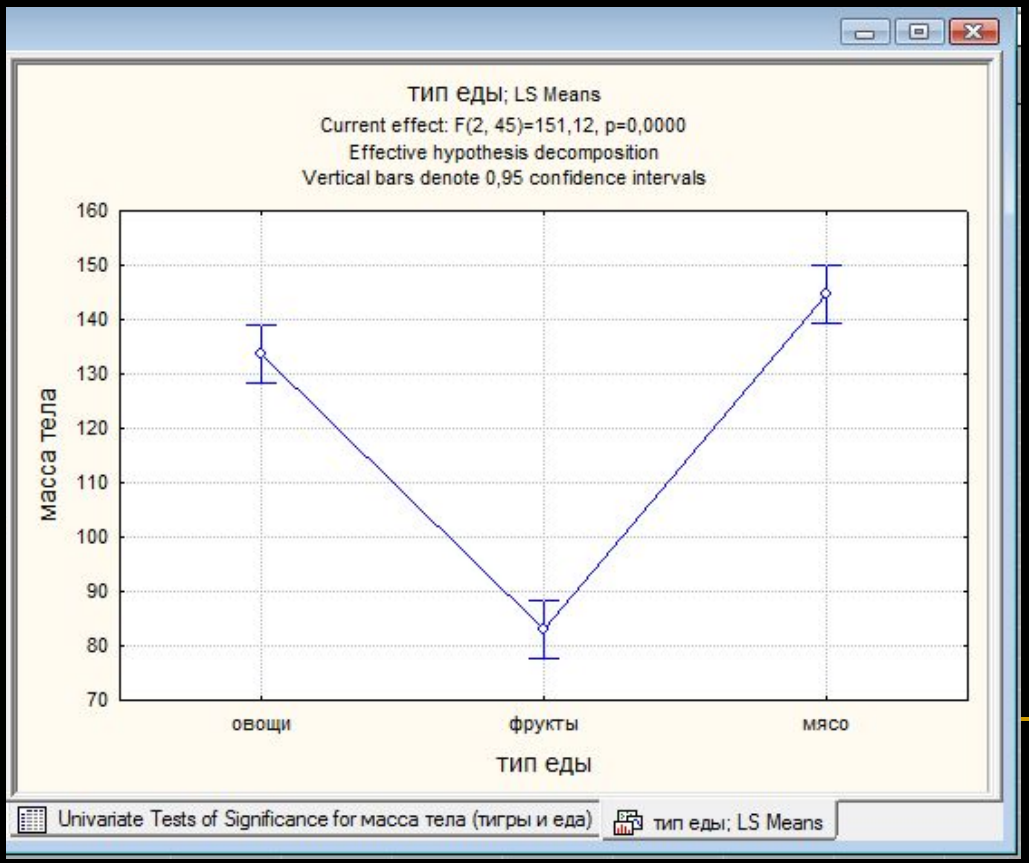
между группами

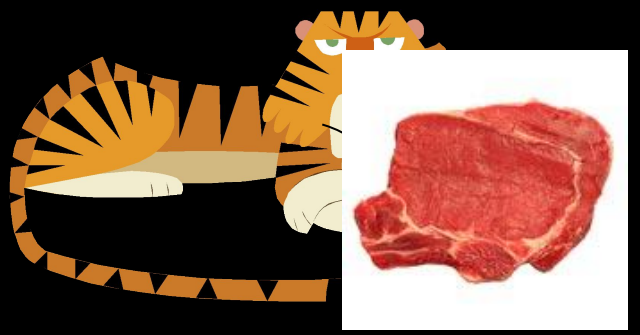
внутри групп

ignificance for масса тела (тигры и еда)

Univariate Tests of Significance for масса тела (тигры и еда)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	696490,1	1	696490,1	6080,228	0,00
тип еды	34621,2	2	17310,6	151,118	0,00
Error	5154,8	45	114,6		





мы отвергаем H_0.
тип еды влиял на массу тигров

ANOVA post hoc tests

Сложная «омнибусная» гипотеза АНОВЫ:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$$

Похожа на стрельбу из дробовика: не нужно особенно точно целиться, **НО** непонятно, какая дробинка попала в какую мишень!

<u>Какая же из отдельных гипотез не верна?</u>

Ответить поможет апостериорный (post hoc) тест!

ANOVA post hoc tests

Если у нас 3 и более групп:

1. Сначала сравнить ВСЕ группы между собой с помощью ANOVA
2. Если различия есть, использовать методы множественного сравнения (группы сравнивают попарно, но вводят поправки)
3. Если различий нет, мы НЕ ИМЕЕМ ПРАВА ПРЕДПРИНИМАТЬ ДАЛЬНЕЙШИЙ АНАЛИЗ!

ANOVA post hoc tests

Поправка Бонферрони (*Bonferroni correction* для небольших k)

если мы хотим обеспечить уровень значимости α, то в каждом из k сравнений нужно принять уровень значимости α/k

Простейшая поправка, но очень грубая!
Не работает при большом числе групп – с увеличением их числа очень сильно падает мощность теста.

Сегодня почти не используется.

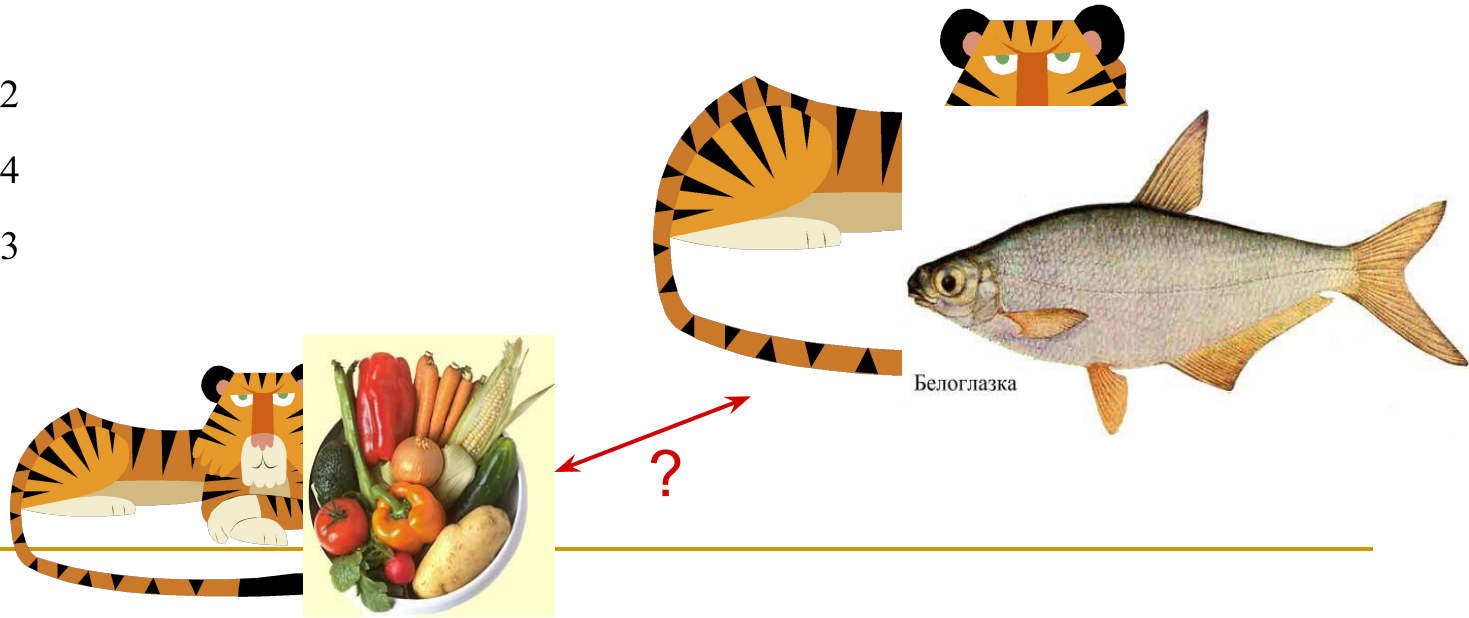
Тест Тьюки (Tukey HSD test)

Наиболее распространённый и рекомендуемый в литературе тест.
Рекомендуется для близких по размеру групп.

Проверяет только ПАРНЫЕ (но не комплексные) гипотезы.

$$H_{01} : \mu_1 = \mu_2$$
$$H_{02} : \mu_1 = \mu_4$$
$$H_{03} : \mu_1 = \mu_3$$

…



Другие апостериорные тесты

1. Критерий Ньюмена-Кейлса (*Newman-Keuls test*) - наименее строгий. Все средние упорядочивают по возрастанию и вычисляют критерий; начинают от сравнения наибольшего с наименьшим.
2. Критерий Шеффе (Scheffe test) – поверяет не только парные гипотезы, но и комплексные.
3. Критерий Даннетта (*Dunnett test*) – используется для сравнения нескольких групп с контрольной группой.

Поправки для множественных сравнений и сравнений с контрольной группой