КРАТКИЙ ОБЗОР МЕТОДОВ СТАТИСТИЧЕСКОГО АНАЛИЗА КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

МОРДОВСКИЙ ЭДГАР АРТУРОВИЧ

К.М.Н., ДОЦЕНТ

ПЛАН

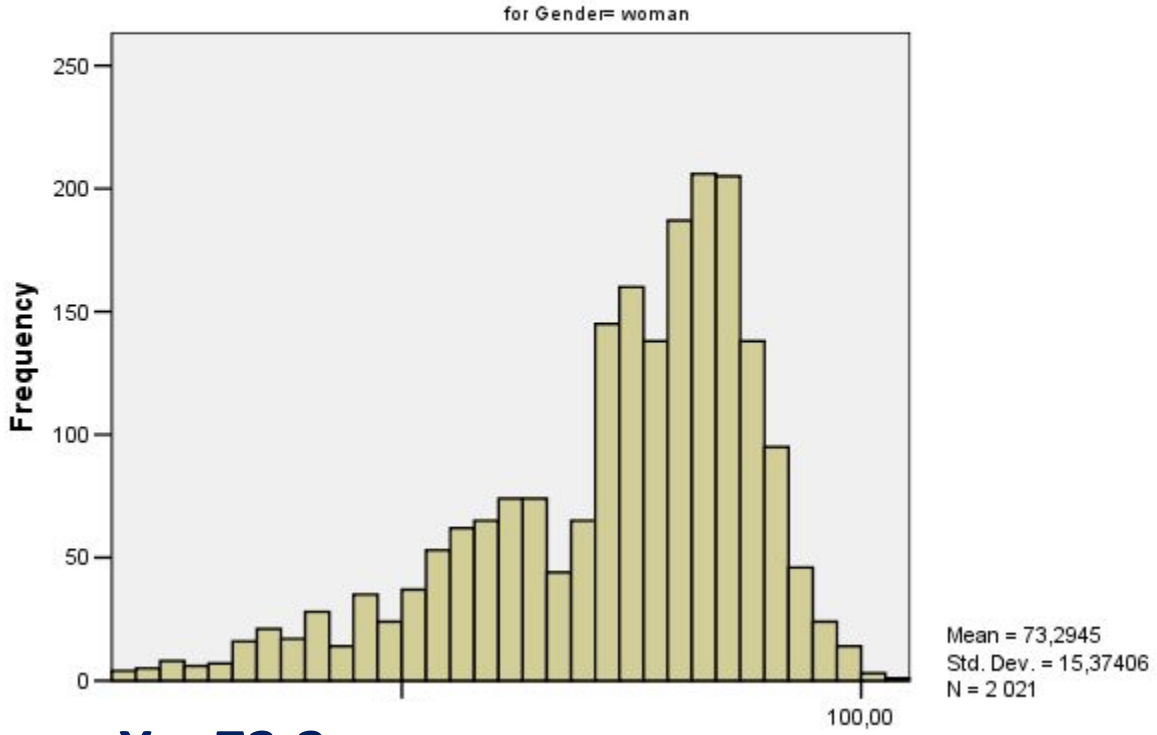
- ОБЩИЕ ТРЕБОВАНИЯ К ВЫПОЛНЕНИЮ СТАТИСТИЧЕСКИХ ТЕСТОВ
- СРАВНЕНИЕ 2-Х СРЕДНИХ ВЕЛИЧИН
- СРАВНЕНИЕ 3-Х И БОЛЕЕ СРЕДНИХ ВЕЛИЧИН
- КОРЕЛЛЯЦИОННЫЙ АНАЛИЗ
- ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

ОБЩИЕ ТРЕБОВАНИЯ К ВЫПОЛНЕНИЮ СТАТИСТИЧЕСКИХ ТЕСТОВ

ТИП ПЕРЕМЕННОЙ ОПРЕДЕЛЯЕТ
НАБОР МЕТОДОВ СТАТИСТИЧЕСКОГО АНАЛИЗА

ПРИМЕР: ФАКТИЧЕСКАЯ СРЕДНЯЯ ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ В ВЫБОРКЕ МУЖЧИН И ЖЕНЩИН, - ЖИТЕЛЕЙ АРХАНГЕЛЬСКОЙ ОБЛАСТИ, УМЕРШИХ В 2012 Г.

ЖЕНЩИНЫ

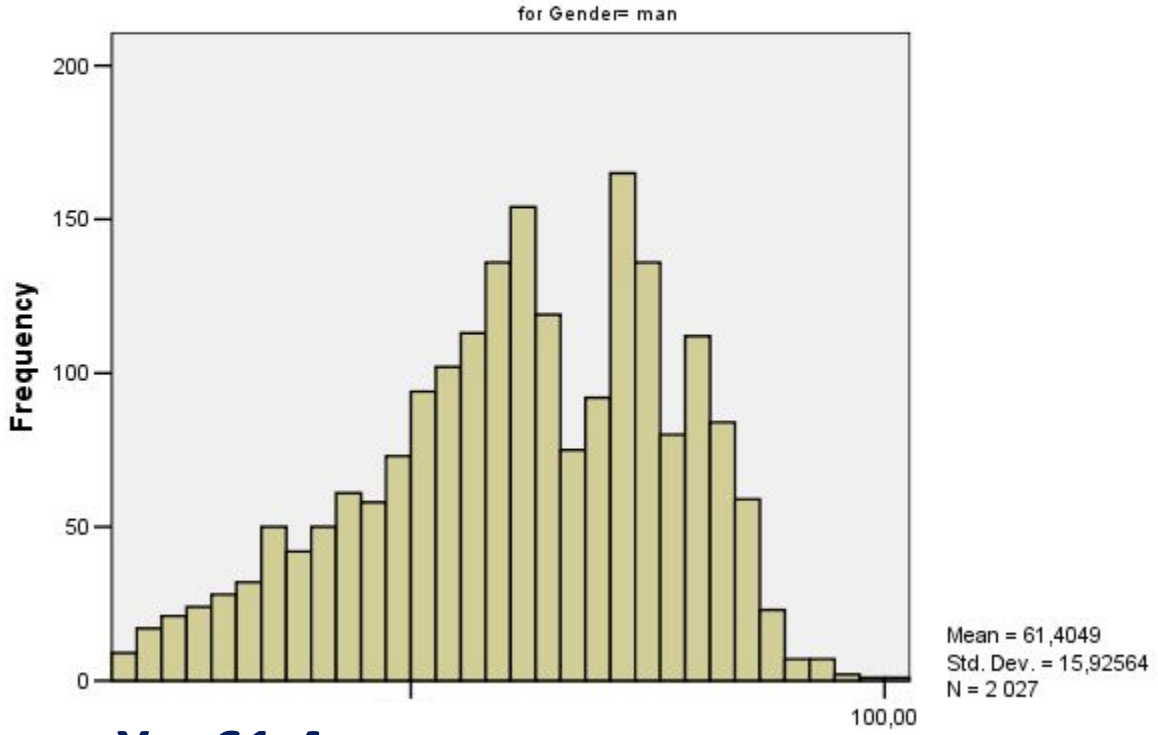


X = 73,3 лет Age

SD = 15,4

N = 2021

МУЖЧИНЫ

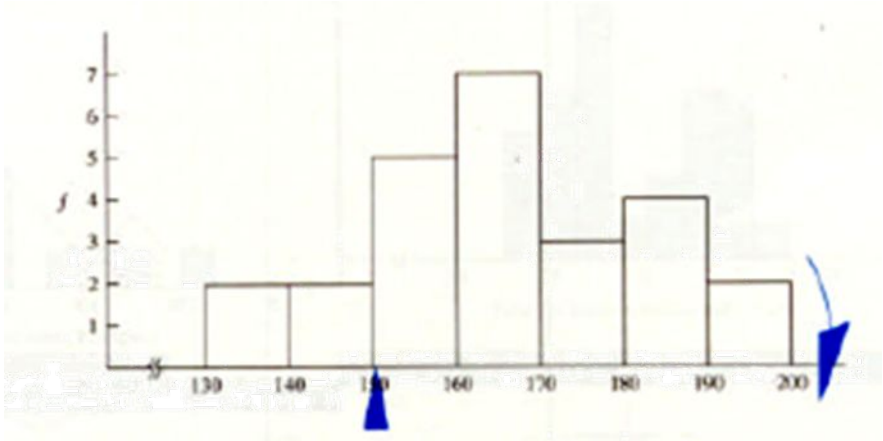


X = 61,4 лет Age

SD = 15,9

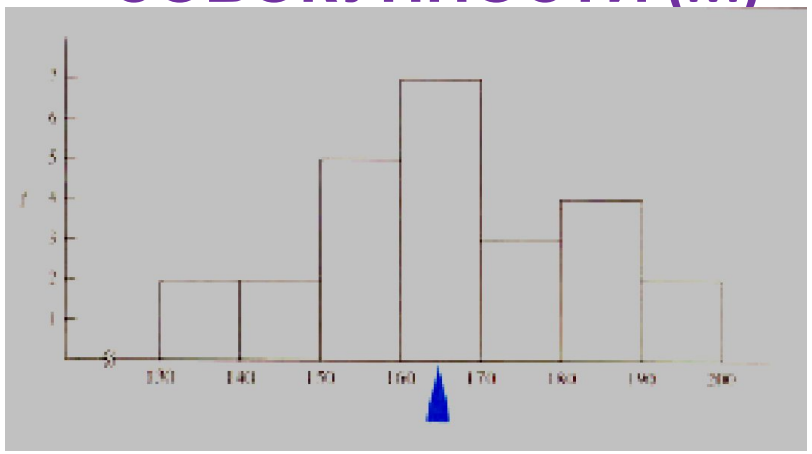
N = 2027

СРЕДНЕЕ АРИФМЕТИЧЕСКИЕ ДЛЯ ВЫБОРКИ (X / m)



$$\overline{X}_{арифм} = \frac{\sum_i x_i}{n},$$

СРЕДНЕЕ АРИФМЕТИЧЕСКИЕ ДЛЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ (м)



$$\mu = \frac{\Sigma X}{N}$$

X (женщины) = 73,3
SD = 15,4
N = 2021

X (мужчины) = 61,4
SD = 15,9
N = 2027

НУЛЕВАЯ И АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА

ГИПОТЕЗА (HYPOTHESIS) – ***предположение*** о свойстве популяции (параметре…)

ФОРМУЛИРУЕМ ДВЕ ВЗАИМОИСКЛЮЧАЮЩИЕ ГИПОТЕЗЫ:

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	Распределение признака **СЛУЧАЙНОЕ** ***(категориальные переменные)*** **НЕТ** отличий в сравниваемых величинах ***(количественные непрерывные переменные)***
Ha (альтернативная гипотеза)	Распределение признака **НЕСЛУЧАЙНОЕ** ***(категориальные переменные)*** **ЕСТЬ** отличия в сравниваемых величинах ***(количественные непрерывные переменные)***

НУЛЕВАЯ И АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	Распределение признака **СЛУЧАЙНОЕ** **НЕТ** отличий в сравниваемых величинах
Ha (альтернативная гипотеза)	Распределение признака **НЕСЛУЧАЙНОЕ** **ЕСТЬ** отличия в сравниваемых величинах

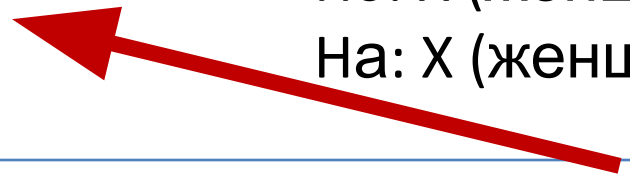
X (женщины) = 73,3 года
SD = 15,4
N = 2021

X (мужчины) = 61,4 года
SD = 15,9
N = 2027

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	**X (женщины) = X (мужчины)** **средняя продолжительность жизни женщин НЕ отличается от средней продолжительности жизни мужчин (т.е. 73,3 = 61,4 в популяции)**
Ha (альтернативная гипотеза)	**X (женщины) ≠ X (мужчины)** **средняя продолжительность жизни женщин ОТЛИЧАЕТСЯ от средней продолжительности жизни мужчин (т.е. 73,3 ≠ 61,4 в популяции)**

2 ВИДА АЛЬТЕРНАТИВНЫХ ГИПОТЕЗ

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
Двусторонняя альтернатива (two-tailed hypothesis)	H0: X (женщины) = X (мужчины) Ha: X (женщины) ≠ X (мужчины)
Односторонняя альтернатива (one-tailed hypothesis)	H0: X (женщины) ≥ X (мужчины) Ha: X (женщины) < X (мужчины)



ЕСТЬ КАКИЕ-ТО ДОПОЛНИТЕЛЬНЫЕ СВЕДЕНИЯ / ГИПОТЕЗЫ (ИСПОЛЬЗУЕТСЯ РЕДКО)

ТЕСТИРОВАНИЕ ГИПОТЕЗ

	ИСТИНА	
	H0 - ВЕРНА	***Ha - ВЕРНА***
МЫ <u>ПРИНИМАЕМ</u> H0	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ** **это чувствительность теста** **(1-α)**	**ОШИБКА 2 ТИПА (β)** **(вероятность НЕ найти то, чего ЕСТЬ)**
МЫ <u>ОТВЕРГАЕМ</u> H0 (ПРИНИМАЕМ Ha)	**ОШИБКА 1 ТИПА (α)** **(уровень значимости – significance (Sig.)** **“p”** **(вероятность найти то,**	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ** **<u>это «мощность теста»</u>** **(1-β)**

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ (методы статистического анализа) – математические расчеты, позволяющие оценить **ВЕРОЯТНОСТЬ ОШИБКИ 1 ТИПА (p / significance (Sig.))**

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ (методы статистического анализа) – математические расчеты, результаты которых позволяют с определенной долей вероятности принять нулевую гипотезу (accept) или ее отвергнуть (reject)

«Приемлемая» вероятность ошибки 1 типа (α-ошибки) = 0.05 (5%)

«КОНСЕНСУС ФИШЕРА»

ЭТО ОТНОСИТЕЛЬНАЯ ВЕЛИЧИНА !!!!!!!!

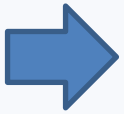
СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

ПОЧЕМУ ВАЖНО ???

А) ПРАВИЛЬНО РАССЧИТАТЬ ОБЪЕМ ВЫБОРКИ ДО НАЧАЛА ИССЛЕДОВАНИЯ ???

ЧТОБЫ МИНИМИЗИРОВАТЬ ВЕРОЯТНОСТЬ ОШИБКИ 1 ТИПА

Б) ПРАВИЛЬНО СФОРМИРОВАТЬ ВЫБОРКУ И ПРАВИЛЬНО ВЫБРАТЬ СТАТИСТИЧЕСКИЙ МЕТОД АНАЛИЗА (СТАТИСТИЧЕСКИЙ КРИТЕРИЙ)



ЧТОБЫ МИНИМИЗИРОВАТЬ ВЕРОЯТНОСТЬ ОШИБКИ 2 ТИПА

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

1 ЭТАП:
ФОРМУЛИРУЕМ H0 и Ha

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	**X (женщины) = X (мужчины)** **средняя продолжительность жизни женщин НЕ отличается от средней продолжительности жизни мужчин**
Ha (альтернативная гипотеза)	**X (женщины) ≠ X (мужчины)** **средняя продолжительность жизни женщин ОТЛИЧАЕТСЯ от средней продолжительности жизни мужчин**

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

2 ЭТАП:
ОПРЕДЕЛЯЕМ УСЛОВИЯ, ПРИ КОТОРЫХ ПРИМЕМ Ha (ОТВЕРГНЕМ H0)

БУДЕМ считать результаты теста «статистически значимыми» (т.е. примем Ha) при вероятности ошибки 1 типа (α-ошибки) менее 0.05 (5%)

«КОНСЕНСУС ФИШЕРА»

p < 0.05 «достаточно», если имеем дело с социологическими исследованиями, «ориентировочными» исследованиями, «пилотными» исследованиями

В клинических испытаниях “p” устанавливается индивидуально (в зависимости от клинической значимости искомого результата) – в т.ч. устанавливается в «SD»

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ ОПРЕДЕЛЯЕТСЯ НАБОРОМ ПАРАМЕТРОВ !!!

СТРОГО ИНДИВИДУАЛЬНО

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

СТАТИСТИЧЕСКИЕ ПРОГРАММЫ
(IBM SPSS, STATA, STATISTICA, PASW, R)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

5 ЭТАП:
ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

ПРИНИМАЕМ H0 / ОТВЕРГАЕМ Ha (если “p” < 0.05)
ПРИНИМАЕМ Ha / ОТВЕРГАЕМ H0 (если “p” ≥ 0.05)

+ ОЦЕНИВАЕМ ВОЗМОЖНОСТЬ ЭКСТРАПОЛЯЦИИ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ НА ГЕНЕРАЛЬНУЮ СОВОКУПНОСТЬ

+ ОЦЕНИВАЕМ СТАТИСТИЧЕСКУЮ МОЩНОСТЬ РЕЗУЛЬТАТА

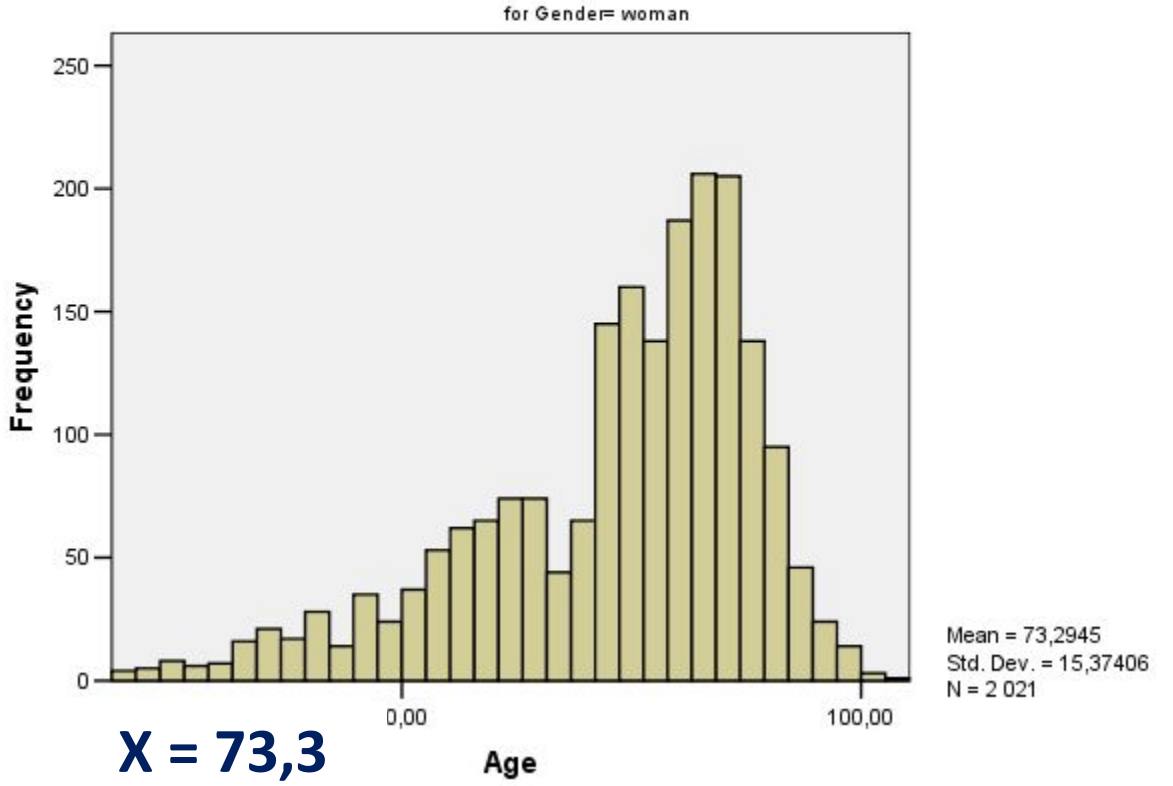
+ ОЦЕНИВАЕМ ПРАКТИЧЕСКУЮ ЗНАЧИМОСТЬ РЕЗУЛЬТАТОВ

СРАВНЕНИЕ 2-Х СРЕДНИХ ВЕЛИЧИН

СРЕДНИЕ ВЕЛИЧИНЫ МОЖНО ВЫЧИСЛИТЬ ТОЛЬКО ДЛЯ КОЛИЧЕСТВЕННЫХ НЕПРЕРЫВНЫХ ВЕЛИЧИН

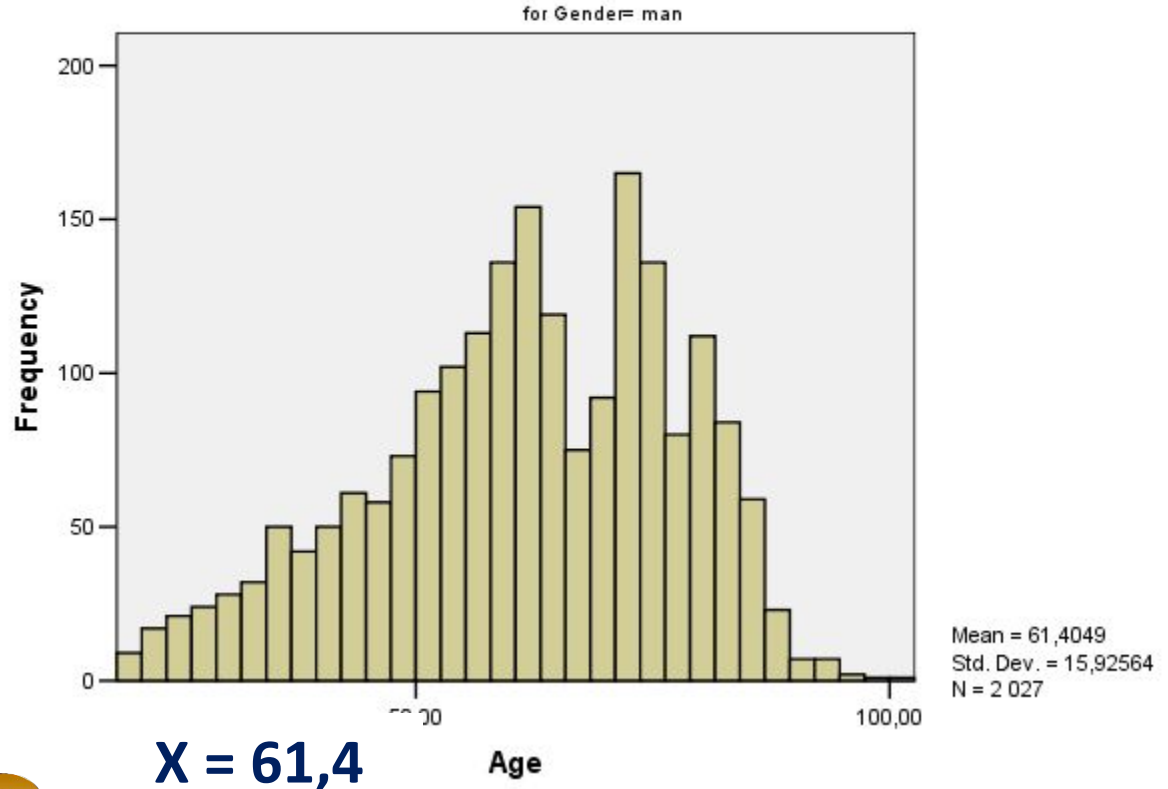
ПРИМЕР: СРЕДНЯЯ ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ В ВЫБОРКЕ МУЖЧИН
И ЖЕНЩИН, - ЖИТЕЛЕЙ АРХАНГЕЛЬСКОЙ ОБЛАСТИ, УМЕРШИХ В

ЖЕНЩИНЫ



X = 73,3
SD = 15,4
N = 2021

МУЖЧИНЫ



X = 61,4
SD = 15,9
N = 2027

ВЫБОР КОНКРЕТНОГО СТАТИСТИЧЕСКОГО МЕТОДА ПРИ СРАВНЕНИИ СРЕДНИХ ВЕЛИЧИН ОПРЕДЕЛЯЕТСЯ:

	УСЛОВИЕ	
1	КОЛИЧЕСТВО СРАВНИВАЕМЫХ ГРУПП	**2 / 3+**
2	РАСПРЕДЕЛЕНИЕ ПРИЗНАКА **В КАЖДОЙ** ИЗ СРАВНИВАЕМЫХ ГРУПП	**нормальное или скошенное**
3	ТИП ВЫБОРКИ	**зависимые выборки («до и после») / независимые выборки (простое сравнение)**

ПОПРАВКА БОНФЕРРОНИ: 2 / 3+ групп

ГОМОГЕННОСТЬ / ГОМОСКЕДАСТИЧНОСТЬ ДИСПЕРСИИ: НЕ КРИТИЧНОЕ ТРЕБОВАНИЕ; ПРИ РАВЕНСТВЕ ОБЪЕМОВ ВЫБОРОК «ПОЧТИ НЕКРИТИЧНОЕ»

СРАВНЕНИЕ 2-Х СРЕДНИХ ВЕЛИЧИН

	НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ В КАЖДОЙ ИЗ СРАВНИВАЕМЫХ ВЫБОРОК (РАЗНИЦЫ ПРИЗНАКА В ПАРАХ ДО-ПОСЛЕ)	СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ В 1 ИЛИ 2 СРАВНИВАЕМЫХ ВЫБОРКАХ (РАЗНИЦЫ ПРИЗНАКА В ПАРАХ ДО-ПОСЛЕ)
НЕЗАВИСИМЫЕ ВЫБОРКИ	Independent Samples T-test (Student T-test) **тест Стьюдента** ***для независимых выборок***	2-Independent Samples test (Mann-Whitney U test) **тест Манна-Уитни** ***для независимых выборок***
ЗАВИСИМЫЕ ВЫБОРКИ (ПОВТОРНЫЕ ИЗМЕРЕНИЯ)	Dependent (Paired Samples) T-test **тест Стьюдента** ***для парных выборок***	2-Related Samples test (Wilcoxon signed-rank test) **тест Вилкоксона** ***для парных выборок***

Independent Samples T-test (Student test)
T-тест Стьюдента

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Нормальное распределение изучаемого признака в каждой из выборок	Test Shapiro-Wilk / Kolmogorov-Smirnov
5. Равенство дисперсий	**Levene's test for Equality of Variances** (sig. (p) ≥ 0,05) **Если дисперсии не равны (p < 0,05)** ***= проблема БЕРЕНСА-ФИШЕРА***

Ho: v 1 = v 2
Ha: v 1 ≠ v 2

2-Independent Samples test (Mann-Whitney U test)
U-тест Манна-Уитни

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Скошенное распределение данных в одной или обеих сравниваемых выборок ***ВНИМАНИЕ: несмотря на то, что распределение скошенное, тест Манна-Уитни оценивает именно СРЕДНИЕ АРИФМЕТИЧЕСКИЕ, А НЕ МЕДИАНЫ !!!***	Test Shapiro-Wilk / Kolmogorov-Smirnov
ДИСПЕРСИЯ НЕ ПРОВЕРЯЕТСЯ	

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

1 ЭТАП:
ФОРМУЛИРУЕМ H0 и Ha

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	**X (женщины) = X (мужчины)** средняя продолжительность жизни женщин **НЕ** отличается от средней продолжительности жизни мужчин
Ha (альтернативная гипотеза)	**X (женщины) ≠ X (мужчины)** средняя продолжительность жизни женщин **ОТЛИЧАЕТСЯ** от средней продолжительности жизни мужчин

2 ЭТАП:
ОПРЕДЕЛЯЕМ УСЛОВИЯ, ПРИ КОТОРЫХ ПРИМЕМ Ha (ОТВЕРГНЕМ H0)

БУДЕМ считать результаты теста «статистически значимыми» (т.е. примем Ha) при вероятности ошибки 1 типа (α-ошибки) менее 0.05 (5%)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender	Statistic	df	Sig.	Statistic	df	Sig.
Age	woman	,118	2021	,000	,923	2021	,000
	man	,066	2027	,000	,978	2027	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ В ВЫБОРКЕ НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ В ВЫБОРКЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

	ИСТИНА	
	Н0 - ВЕРНА	***На - ВЕРНА***
МЫ ПРИНИМАЕМ Н0	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ (= чувствительность теста)**	**ОШИБКА 2 ТИПА (вероятность НЕ найти то, чего ЕСТЬ)**
МЫ ОТВЕРГАЕМ Н0 (ПРИНИМАЕМ На)	**ОШИБКА 1 ТИПА (уровень значимости – significance (Sig.) “p” (вероятность найти то, чего НЕТ)**	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ (= мощность теста)**

p (женщины) < 0,0001
p (мужчины) < 0,0001

т.е. МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender	Statistic	df	Sig.	Statistic	df	Sig.
Age	woman	,118	2021	,000	,923	2021	,000
	man	,066	2027	,000	,978	2027	,000

a. Lilliefors Significance Correction

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Скошенное распределение данных в одной или обеих сравниваемых выборок **ВНИМАНИЕ: несмотря на то, что распределение скошенное, тест Манна-Уитни сравнивает именно СРЕДНИЕ АРИФМЕТИЧЕСКИЕ, А НЕ МЕДИАНЫ !!!**	Test Shapiro-Wilk / Kolmogorov-Smirnov
ДИСПЕРСИЯ НЕ ПРОВЕРЯЕТСЯ	

2-Independent Samples test (Mann-Whitney U test)
U-тест Манна-Уитни

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста Манна-Уитни

Но: m1 = m2 (средняя продолжительность жизни мужчин не отличается от средней продолжительности жизни женщин)

На: m1 ≠ m2 (средняя продолжительность жизни мужчин отличается от средней продолжительности жизни женщин)

Test Statistics[a]

	Age
Mann-Whitney U	1144664
Wilcoxon W	3200042
Z	-24,305
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: Gender

p < 0,0001

т.е. МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

5 ЭТАП:
ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

+ ОЦЕНИВАЕМ ПРАКТИЧЕСКУЮ ЗНАЧИМОСТЬ РЕЗУЛЬТАТОВ

Средняя продолжительность жизни мужчин меньше, чем средняя продолжительность жизни женщин на 11,9 лет

X = 73,3	X = 61,4
SD = 15,4	SD = 15,9
N = 2021	N = 2027

2-Independent Samples test
(Mann-Whitney U test) тест Манна-Уитни

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ («АКАДЕМИЧЕСКАЯ ВЕРСИЯ»)

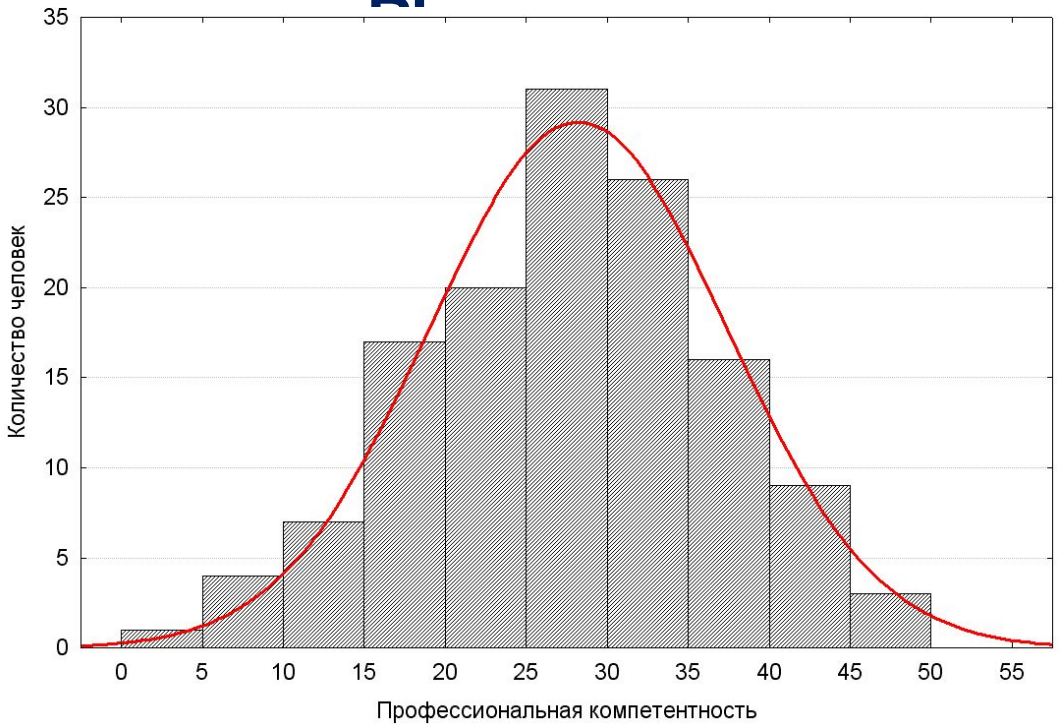
X (мужчины) = 61,4 лет (95% ДИ: 60,7 – 62,1)
X (женщины) = 73,3 лет (95% ДИ: 72,6 – 74,0)

Различия являются статистически значимыми ($p < 0,0001$)

РЕКОМЕНДУЕТСЯ УКАЗЫВАТЬ ТОЧНОЕ ЗНАЧЕНИЕ «p»
(необходимо продемонстрировать вероятность ошибки)

ПРИМЕР: СРЕДНЯЯ ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ В ВЫБОРКЕ МУЖЧИН И ЖЕНЩИН, - ЖИТЕЛЕЙ АРХАНГЕЛЬСКОЙ ОБЛАСТИ, УМЕРШИХ В 2012 Г.

ЖЕНЩИНЫ



X = 73,3

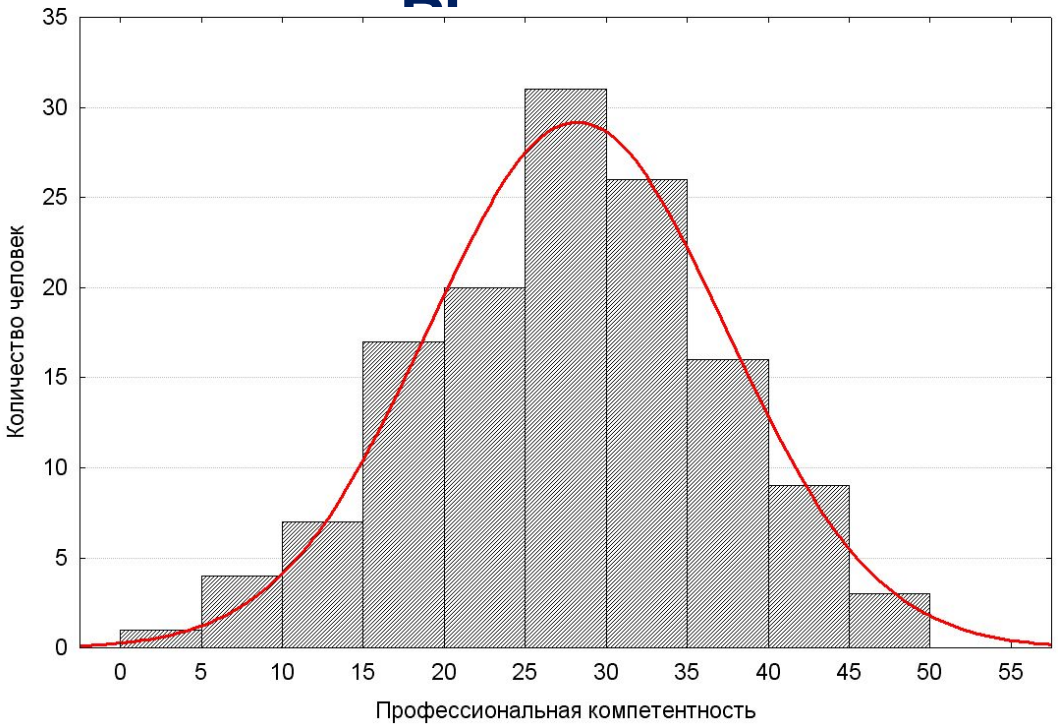
SD = 15,4

N = 2021

?

≠

МУЖЧИНЫ



X = 61,4

SD = 15,9

N = 2027

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender	Statistic	df	Sig.	Statistic	df	Sig.
Age	woman	,118	2021	**,298**	,923	2021	,000
	man	,066	2027	**,345**	,978	2027	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ В ВЫБОРКЕ НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ В ВЫБОРКЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

	ИСТИНА	
	Н0 - ВЕРНА	***На - ВЕРНА***
МЫ ПРИНИМАЕМ Н0	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ (= чувствительность теста)**	**ОШИБКА 2 ТИПА (вероятность НЕ найти то, чего ЕСТЬ)**
МЫ ОТВЕРГАЕМ Н0 (ПРИНИМАЕМ На)	**ОШИБКА 1 ТИПА (уровень значимости – significance (Sig.) “p” (вероятность найти то, чего НЕТ)**	**ПРАВИЛЬНЫЙ РЕЗУЛЬТАТ (= мощность теста)**

p (женщины) = 0,298
p (мужчины) = 0,345

т.е. НЕ МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) = 29,8% и 34,5%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender	Statistic	df	Sig.	Statistic	df	Sig.
Age	woman	,118	2021	**,298**	,923	2021	,000
	man	,066	2027	**,345**	,978	2027	,000

a. Lilliefors Significance Correction

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Нормальное распределение изучаемого признака в каждой из выборок	Test Shapiro-Wilk / Kolmogorov-Smirnov
5. Равенство дисперсий	**Levene's test for Equality of Variances** (sig. (p) ≥ 0,05) **Если дисперсии не равны (p < 0,05)** ***= проблема БЕРЕНСА-ФИШЕРА***

Independent Samples T-test (Student test)
T-тест Стьюдента

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста Стьюдента

Но: m1 = m2 (средняя продолжительность жизни мужчин не отличается от средней продолжительности жизни женщин)

На: m1 ≠ m2 (средняя продолжительность жизни мужчин отличается от средней продолжительности жизни женщин)

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Нормальное распределение изучаемого признака в каждой из выборок	Test Shapiro-Wilk / Kolmogorov-Smirnov
5. Равенство дисперсий	**Levene's test for Equality of Variances** (sig. (p) ≥ 0,05) **Если дисперсии не равны (p < 0,05)** ***= проблема БЕРЕНСА-ФИШЕРА***

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста ЛЕВЕНЕ (тест равенства дисперсий)

Но: σ 1 = σ 2 (дисперсия средней продолжительности жизни мужчин не отличается от дисперсии средней продолжительности жизни женщин)

На: σ 1 ≠ σ 2 (дисперсия средней продолжительности жизни мужчин отличается от дисперсии средней продолжительности жизни женщин)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age	Equal variances assumed	8,286	,004	24,164	4046	,000	11,88957	,49204	10,92491	12,85424
	Equal variances not assumed			24,165	4041,791	,000	11,88957	,49201	10,92496	12,85419

Independent Samples T-test
(Student test) тест Стьюдента

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ

X (мужчины) = 61,4 лет (95% ДИ: 60,7 – 62,1)
X (женщины) = 73,3 лет (95% ДИ: 72,6 – 74,0)

Средняя продолжительность жизни мужчин на 11,9 лет меньше (95% ДИ: 11,9 – 12,9), чем женщин ($p < 0,0001$)

РЕКОМЕНДУЕТСЯ УКАЗЫВАТЬ ТОЧНОЕ ЗНАЧЕНИЕ «p»

(необходимо продемонстрировать вероятность ошибки)

СРАВНЕНИЕ 2-Х СРЕДНИХ ВЕЛИЧИН

	НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ В КАЖДОЙ ИЗ СРАВНИВАЕМЫХ ВЫБОРОК	СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ В 1 ИЛИ 2 СРАВНИВАЕМЫХ ВЫБОРКАХ
НЕЗАВИСИМЫЕ ВЫБОРКИ	**Independent Samples T-test** (Student T-test) **тест Стьюдента**	**2-Independent Samples test** (Mann-Whitney U test) **тест Манна-Уитни**
ЗАВИСИМЫЕ ВЫБОРКИ (ПОВТОРНЫЕ ИЗМЕРЕНИЯ)	**Dependent (Paired Samples) T-test тест Стьюдента *для парных выборок***	**2-Related Samples test (Wilcoxon signed-rank test) тест Вилкоксона *для парных выборок***

Paired Samples T-test
тест Стьюдента для парных выборок

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **зависимыми** (одни и те же участники, но в разное время)	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Нормальное распределение **разности** между значениями изучаемого признака в парах	Test Shapiro-Wilk / Kolmogorov-Smirnov

ДО *(до-после)*	ПОСЛЕ	РАЗНОСТЬ
167	134	-33
156	160	4
177	129	-48
…	…	…

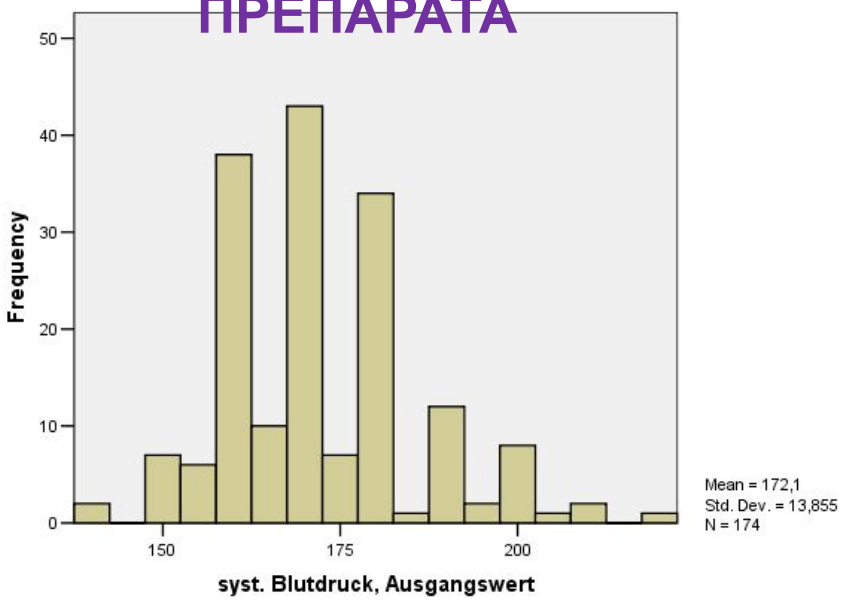
2-Related Samples test (Wilcoxon)
тест Вилкоксона

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **зависимыми** (одни и те же участники в разное время)	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Скошенное распределение **разности** между значениями изучаемого признака	Test Shapiro-Wilk / Kolmogorov-Smirnov

ДО	ПОСЛЕ	РАЗНОСТЬ
167	134	-33
156	160	4
177	129	-48
…	…	…

ПРИМЕР: УРОВЕНЬ АРТЕРИАЛЬНОГО ДАВЛЕНИЯ В ГРУППЕ ПАЦИЕНТОВ, ПРИНИМАЮЩИХ АНТИГИПЕРТЕНЗИВНЫЙ ПРЕПАРАТ

ДО НАЧАЛА ПРИЕМА ПРЕПАРАТА

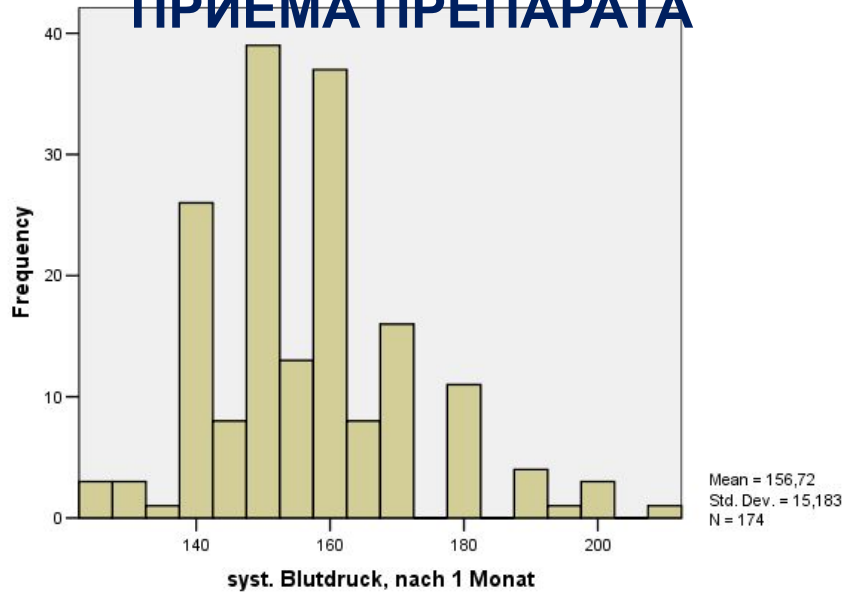


X = 172,1
SD = 13,9
N = 174

?

≠

ЧЕРЕЗ 1 МЕСЯЦ ПОСЛЕ НАЧАЛА ПРИЕМА ПРЕПАРАТА



X = 156,7
SD = 15,2
N = 174

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

1 ЭТАП:
ФОРМУЛИРУЕМ H0 и Ha

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 ***(нулевая гипотеза)***	**X (ДО) = X (ПОСЛЕ)** **средний уровень артериального давления в группе пациентов до начала приема препарата** **НЕ отличается от среднего уровня артериального давления в группе пациентов после начала приема препарата**
Ha ***(альтернативная гипотеза)***	**X (ДО) ≠ X (ПОСЛЕ)** **средний уровень артериального давления в группе пациентов до начала приема препарата ОТЛИЧАЕТСЯ от среднего уровня артериального давления в группе пациентов после начала приема препарата**

2 ЭТАП:
ОПРЕДЕЛЯЕМ УСЛОВИЯ, ПРИ КОТОРЫХ ПРИМЕМ Ha (И ОТВЕРГНЕМ H0)

БУДЕМ считать результаты теста «статистически значимыми» (т.е. примем Ha) при вероятности ошибки 1 типа (α-ошибки) менее 0.01 / 0.05 (1% / 5%)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

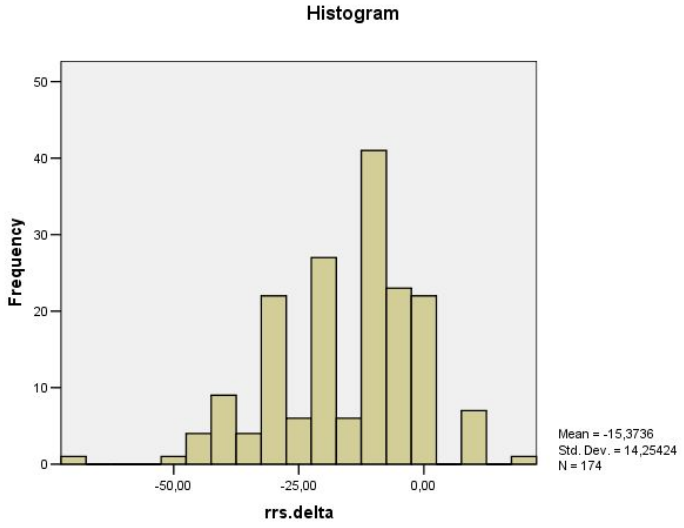
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rrs.delta	,187	174	,000	,954	174	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ РАЗНИЦЫ СРЕДНИХ ВЕЛИЧИН (ДО-ПОСЛЕ) НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ РАЗНИЦЫ СРЕДНИХ ВЕЛИЧИН (ДО-ПОСЛЕ) В ВЫБОРКЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО



p (женщины) < 0,0001
p (мужчины) < 0,0001

т.е. МОЖЕМ принять На вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rrs.delta	,187	174	,000	,954	174	,000

a. Lilliefors Significance Correction

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **зависимыми** (одни и те же участники в разное время)	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Скошенное распределение **разности** между значениями изучаемого признака	Test Shapiro-Wilk / Kolmogorov-Smirnov

2-Related Samples test (Wilcoxon)
тест Вилкоксона

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста Вилкоксона

Но: m1 = m2 (среднее АД до начала приема препарата **не отличается** от среднего АД через 1 месяц после начала приема препарата)

Ha: m1 ≠ m2 (среднее АД до начала приема препарата **отличается** от среднего АД через 1 месяц после начала приема препарата)

Test Statistics[b]

	syst. Blutdruck, nach 1 Monat - syst. Blutdruck, Ausgangswert
Z	-9,970[a]
Asymp. Sig. (2-tailed)	,000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

p < 0,0001

т.е. МОЖЕМ принять Ha
вероятность ошибки 1 типа (ошибочно принять Ha - найти то, чего нет) < 0,1%

2-Related Samples test (Wilcoxon)
тест Вилкоксона

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ («АКАДЕМИЧЕСКАЯ ВЕРСИЯ»)

M (до) = 172,1 мм рт.ст.

M (после) = 156,7 мм рт.ст.

Различия являются статистически значимыми

($p < 0,0001$)

РЕКОМЕНДУЕТСЯ УКАЗЫВАТЬ ТОЧНОЕ ЗНАЧЕНИЕ «p»

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

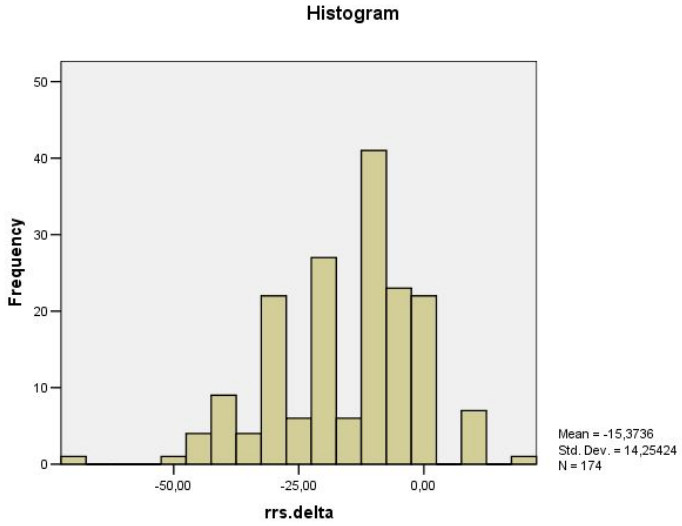
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rrs.delta	,187	174	**,298**	,954	174	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ РАЗНИЦЫ СРЕДНИХ ВЕЛИЧИН (ДО-ПОСЛЕ) НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ РАЗНИЦЫ СРЕДНИХ ВЕЛИЧИН (ДО-ПОСЛЕ) В ВЫБОРКЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО



p (мужчины) = 0,298

т.е. НЕ МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rrs.delta	,187	174	**,298**	,954	174	,000

a. Lilliefors Significance Correction

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **зависимыми** (одни и те же участники в разное время)	см. характеристики собранных данных
3. Количественный непрерывный тип данных в каждой из сравниваемых выборок	см. тип данных
4. Нормальное распределение **разности** между значениями изучаемого признака в парах (до-после)	Test Shapiro-Wilk / Kolmogorov-Smirnov

Paired Samples T-test
тест Стьюдента для парных выборок

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для парного теста Стьюдента

Но: m1 = m2 (среднее АД до начала приема препарата **не отличается** от среднего АД через 1 месяц после начала приема препарата)

На: m1 ≠ m2 (среднее АД до начала приема препарата **отличается** от среднего АД через 1 месяц после начала приема препарата)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	syst. Blutdruck, Ausgangswert - syst. Blutdruck, nach 1 Monat	15,374	14,254	1,081	13,241	17,506	14,227	173	,000

Paired Samples T-test
тест Стьюдента для парных выборок

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ («АКАДЕМИЧЕСКАЯ ВЕРСИЯ»)

M (до) = 172,1 мм рт.ст.

M (после) = 156,7 мм рт.ст.

Различия являются статистически значимыми

($p < 0,0001$)

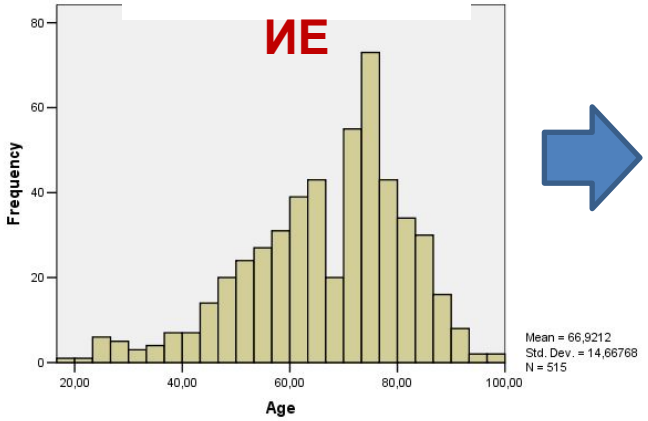
РЕКОМЕНДУЕТСЯ УКАЗЫВАТЬ ТОЧНОЕ ЗНАЧЕНИЕ «p»

СРАВНЕНИЕ 3-Х И БОЛЕЕ СРЕДНИХ ВЕЛИЧИН

СРАВНЕНИЕ 2-х СРЕДНИХ ВЕЛИЧИН
СРАВНЕНИЕ 3-х И БОЛЕЕ СРЕДНИХ ВЕЛИЧИН

ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

ВЫСШЕЕ ОБРАЗОВАНИЕ

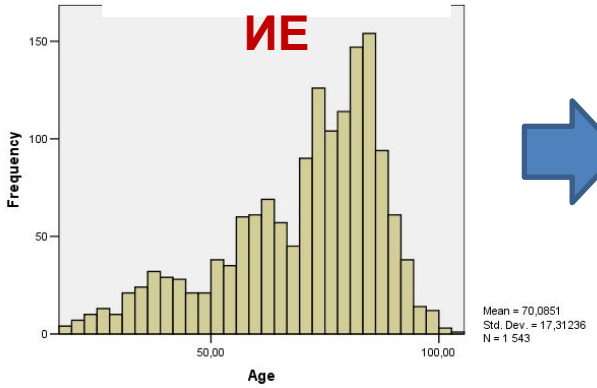


X = 66,9
SD = 14,7
N = 515

СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ

X = 65,9
SD = 16,1
N = 1627

СРЕДНЕЕ ОБРАЗОВАНИЕ



X = 70,1
SD = 17,3
N = 1543

СРАВНЕНИЕ 2-х СРЕДНИХ ВЕЛИЧИН
СРАВНЕНИЕ 3-х И БОЛЕЕ СРЕДНИХ ВЕЛИЧИН

ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

ВЫСШЕЕ ОБРАЗОВАНИЕ		СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ		СРЕДНЕЕ ОБРАЗОВАНИЕ
X = 66,9	?	X = 65,9	?	X = 70,1
SD = 14,7		SD = 16,1		SD = 17,3
N = 515		N = 1627		N = 1543

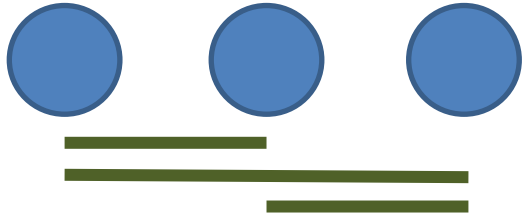
Почему нельзя сравнить группы попарно с помощью *t*-критерия Стьюдента?

ЭФФЕКТ МНОЖЕСТВЕННЫХ СРАВНЕНИЙ

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

$$P_{ошибки} = 1-(1-0,05)^k$$

$$P_{ошибки} = 1-(1-0,05)^k = 1-(1-0,05)^3 = 14,3\%$$



ВЫПОЛНЯЯ СЕРИЮ ПОПАРНЫХ СРАВНЕНИЙ, В КАЖДОМ СЛУЧАЕ МЫ УМЕНЬШАЕМ ОБЪЕМ

СРАВНЕНИЕ 3-Х И БОЛЕЕ СРЕДНИХ ВЕЛИЧИН

		НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ В КАЖДОЙ ИЗ СРАВНИВАЕМЫХ ВЫБОРОК	СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ В 1 ИЛИ БОЛЕЕ СРАВНИВАЕМЫХ ВЫБОРОК
НЕЗАВИСИМЫЕ ВЫБОРКИ	**РУС. ВЕРСИЯ**	**ONE-WAY ANOVA** **ДИСПЕРСИОННЫЙ АНАЛИЗ**	**K-Independent Samples test** (Kruskall-Wallis H test) **Тест Крускелла-Уоллиса**
ЗАВИСИМЫЕ ВЫБОРКИ (ПОВТОРНЫЕ ИЗМЕРЕНИЯ)	**РУС. ВЕРСИЯ**	**REPEATED MEASURES ANOVA (GLM-4)** **Дисперсионный анализ для повторных измерений**	**Friedman's test (Friedman's ANOVA)** **Дисперсионный анализ Фридмана**

НО ! Считается, что нарушение нормальности распределения не оказывает существенного влияния на результаты)

НО ! Считается, что нарушение равенства дисперсии выборок оказывает значимое влияние в том случае, если сравниваемые выборки отличаются по численности)

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 3 и более выборки	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. **Количественный непрерывный тип данных** в каждой из сравниваемых выборок	см. тип данных
4. **Нормальное распределение** изучаемого признака в сравниваемых группах	Test Shapiro-Wilk / Kolmogorov-Smirnov
5. **Равенство дисперсий** изучаемого признака в сравниваемых группах (гомоскедастичность)	Levene's test for Equality of Variances (Sig. (p) ≥ 0,05) Если дисперсии не равны (p < 0,05) **= поправки Brown-Forsythe / Welch**

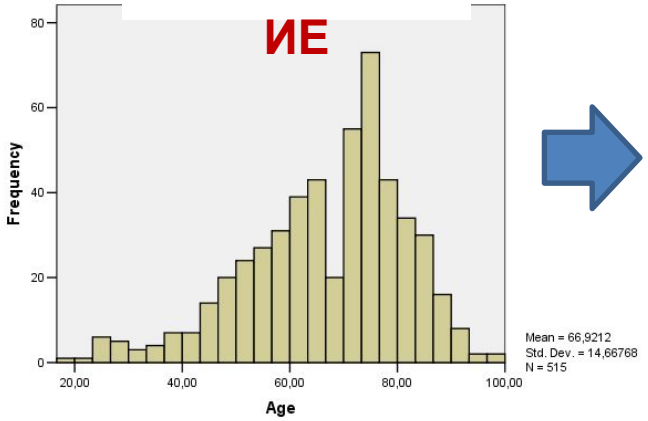
K-Independent Samples test (Kruskall-Wallis H test)
Тест Краскелла-Уоллиса

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 3 и более выборок	см. характеристики собранных данных
2. Выборки д.б. независимыми	см. характеристики собранных данных
3. **Количественный непрерывный тип данных** в каждой из сравниваемых выборок	см. тип данных
4. **Скошенное распределение** данных хотя бы в одной из сравниваемых выборок	Test Shapiro-Wilk / Kolmogorov-Smirnov

ДИСПЕРСИЯ НЕ ПРОВЕРЯЕТСЯ

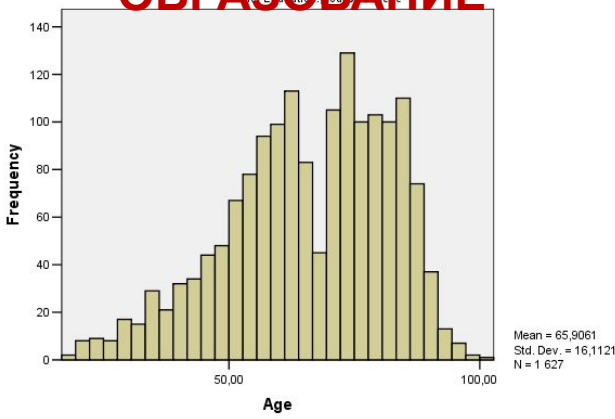
ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

ВЫСШЕЕ ОБРАЗОВАНИЕ



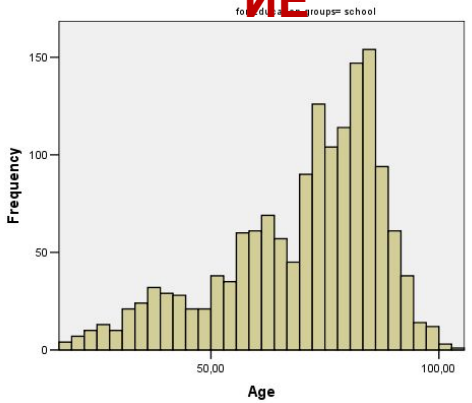
X = 66,9
SD = 14,7
N = 515

СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ



X = 65,9
SD = 16,1
N = 1627

СРЕДНЕЕ ОБРАЗОВАНИЕ



X = 70,1
SD = 17,3
N = 1543

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

1 ЭТАП:
ФОРМУЛИРУЕМ H0 и Ha

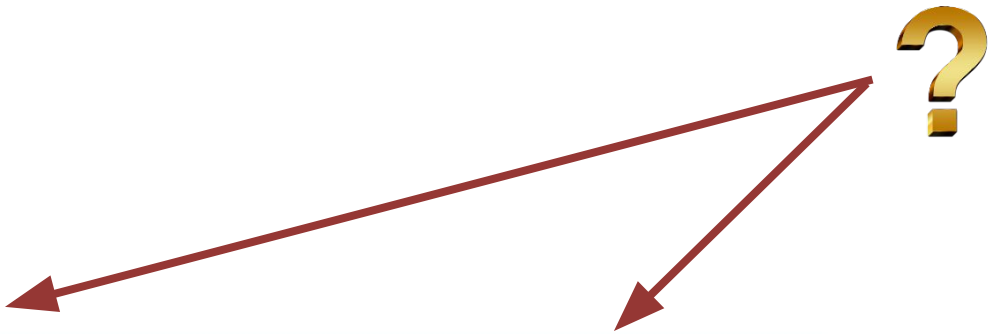
ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	**X (высшее) = X (ср.спец.) = X (среднее)** **средняя продолжительность жизни не зависит от уровня образования**
Ha (альтернативная гипотеза)	**X (высшее) ≠ X (ср.спец.)** **X (высшее) ≠ X (среднее)** **X (ср.спец.) ≠ X (среднее)** **мы отвергаем H0 гипотезу если верна хотя бы одна из частных Ha**

2 ЭТАП:
ОПРЕДЕЛЯЕМ УСЛОВИЯ, ПРИ КОТОРЫХ ПРИМЕМ Ha (ОТВЕРГНЕМ H0)

БУДЕМ считать результаты теста «статистически значимыми» (т.е. примем Ha) при вероятности ошибки 1 типа (α-ошибки) менее 0.05 (5%)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА



		НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ В КАЖДОЙ ИЗ СРАВНИВАЕМЫХ ВЫБОРОК	СКОШЕННОЕ РАСПРЕДЕЛЕНИЕ В 1 ИЛИ БОЛЕЕ СРАВНИВАЕМЫХ ВЫБОРОК
НЕЗАВИСИМЫЕ ВЫБОРКИ	РУС.ВЕРСИЯ	ONE-WAY ANOVA ДИСПЕРСИОННЫЙ АНАЛИЗ	K-Independent Samples test (Kruskall-Wallis H test) Тест Краскелла-Уоллиса
ЗАВИСИМЫЕ ВЫБОРКИ (ПОВТОРНЫЕ ИЗМЕРЕНИЯ)	РУС.ВЕРСИЯ	REPEATED MEASURES ANOVA (GLM-4) Дисперсионный анализ для повторных измерений	Friedman's test (Friedman's ANOVA) тест Фридмана

НО ! Считается, что нарушение нормальности распределения не оказывает существенного влияния на результаты)	НО ! Считается, что нарушение равенства дисперсии выборок оказывает значимое влияние в том случае, если сравниваемые выборки отличаются по численности)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Education groups	Statistic	df	Sig.	Statistic	df	Sig.
Age	university	,098	515	,000	,965	515	,000
	college	,074	1627	,000	,972	1627	,000
	school	,118	1543	,000	,931	1543	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ ПРИЗНАКА В ГРУППАХ НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ В ГРУППАХ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

p (высшее) < 0,0001
p (сред.спец.) < 0,0001
p (среднее) < 0,0001

т.е. МОЖЕМ принять На
вероятность ошибки 1 типа < 0,1%
(ошибочно принять На - найти то, чего нет)

K-Independent Samples test
(Kruskall-Wallis H test)
Тест Краскелла-Уоллиса

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста Краскелла-Уоллиса

Но: m1 = m2 = m3

На: m1 ≠ m2 / m1 ≠ m3 / m2 ≠ m3

Test Statistics[a,b]

	Age
Chi-Square	79,561
df	2
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: Education.groups

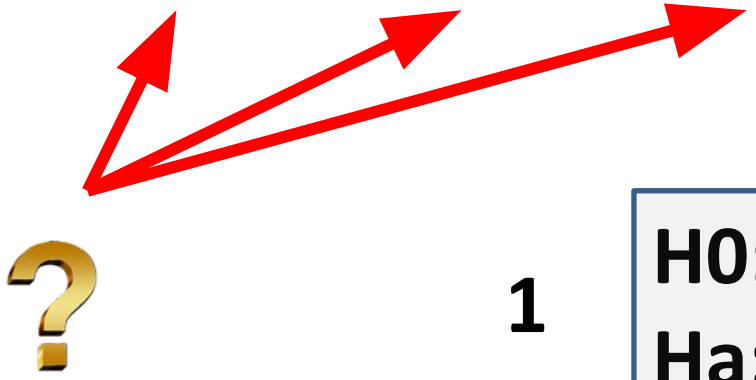
p < 0,0001

т.е. МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

Ha: m1 ≠ m2 / m1 ≠ m3 / m2 ≠ m3



ДАЛЕЕ НЕОБХОДИМА СЕРИЯ ПРОЦЕДУР
ТЕСТА МАННА-УИТНИ

1
H0: m1 = m2
Ha: m1 ≠ m2

2
H0: m1 = m3
Ha: m1 ≠ m3

3
H0: m2 = m3
Ha: m2 ≠ m3

ПОПРАВКА БОНФЕРРОНИ: критический уровень “p” < 0.05/3 = < 0.017

Test Statistics[a]

	Age
Mann-Whitney U	407457,0
Wilcoxon W	1731835
Z	-,940
Asymp. Sig. (2-tailed)	,347

a. Grouping Variable: Education.groups

H0: m1 = m2

Средняя продолжительность жизни лиц с высшим образованием не отличается от средней продолжительности жизни лиц со средним специальным образованием

Test Statistics[a]

	Age
Mann-Whitney U	331066,0
Wilcoxon W	463936,0
Z	-5,674
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: Education.groups

Ha: m1 ≠ m3

Средняя продолжительность жизни лиц с высшим образованием отличается от средней продолжительности жизни лиц со средним образованием

Test Statistics[a]

	Age
Mann-Whitney U	1038009
Wilcoxon W	2362387
Z	-8,434
Asymp. Sig. (2-tailed)	,000

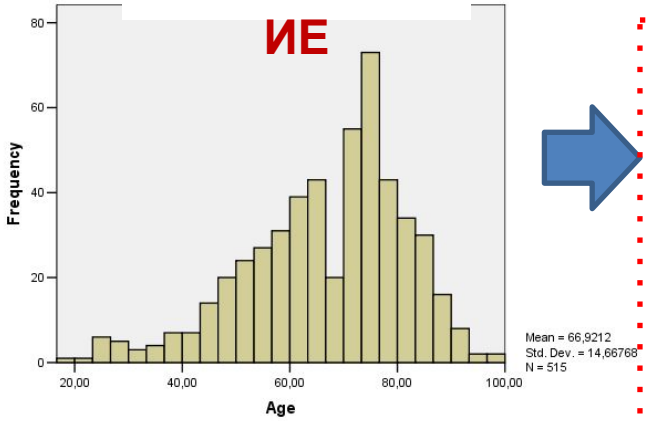
a. Grouping Variable: Education.groups

Ha: m2 ≠ m3

Средняя продолжительность жизни лиц со средним специальным образованием отличается от средней продолжительности жизни лиц со средним образованием

ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

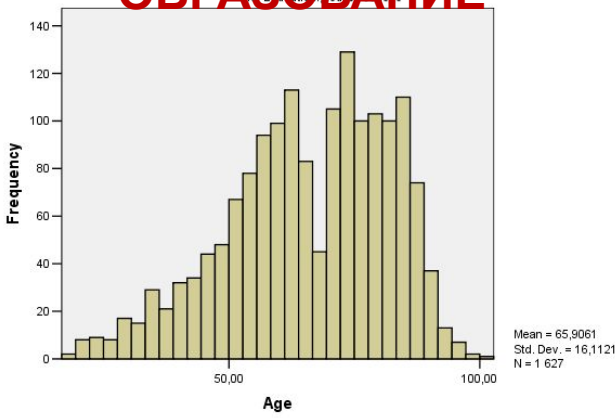
ВЫСШЕЕ ОБРАЗОВАНИЕ



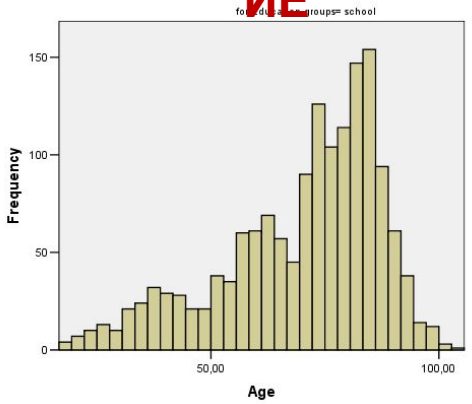
X = 66,9
SD = 14,7
N = 515

X = 65,9
SD = 16,1
N = 1627

СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ



СРЕДНЕЕ ОБРАЗОВАНИЕ



X = 70,1
SD = 17,3
N = 1543

K-Independent Samples test (Kruskall-Wallis H test)
Тест Краскелла-Уоллиса

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ («АКАДЕМИЧЕСКАЯ ВЕРСИЯ»)

m1 = 66,9 (95% ДИ: 65,7 – 68,2)

m2 = 65,9 (95% ДИ: 65,1 – 66,7)

m3 = 70,1 (95% ДИ: 69,2 – 70,9)

«…средняя продолжительность жизни зависит от уровня образования человека ($H = 79,6$; $p < 0,0001$). Продолжительность жизни лиц, имевших среднее образование, была статистически значимо выше, чем у лиц, имевших высшее и среднее специальное образование; средняя продолжительность жизни лиц, имевших высшее и среднее специальной

ПОПРАВКА БОНФЕРРОНИ:
ОШИБКА 1 ТИПА: $\alpha / n = 0.05/3 = 0,017$

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

Tests of Normality

	Education groups	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Age	university	,098	515	**,298**	,965	515	,000
	college	,074	1627	**,345**	,972	1627	,000
	school	,118	1543	**,455**	,931	1543	,000

a. Lilliefors Significance Correction

Но: РАСПРЕДЕЛЕНИЕ ПРИЗНАКА В ГРУППАХ НЕ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

На: РАСПРЕДЕЛЕНИЕ ПРИЗНАКА В ГРУППАХ ОТЛИЧАЕТСЯ ОТ НОРМАЛЬНОГО

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

p (высшее) = 0,298
p (сред.спец.) < 0,345
p (среднее) < 0,455

т.е. ОТКЛОНЯЕМ На
вероятность ошибки 1 типа > 5%

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

- **ЦЕЛЬ:** с помощью ***ДА*** исследуют влияние ***одной (одномерный анализ)*** или нескольких ***(многомерный анализ)*** **независимых переменных** на ***одну зависимую*** переменную или на ***несколько зависимых*** переменных
- Независимые переменные ***КАК ПРАВИЛО*** принимают только дискретные значения (относятся к номинальной или порядковой шкале) - это **ФАКТОРНЫЙ АНАЛИЗ**
- Если независимые переменные принадлежат к интервальной шкале или к шкале отношений, то их называют ***ковариациями*** - это **КОВАРИАЦИОННЫЙ АНАЛИЗ**

ДИСПЕРСИОННЫЙ АНАЛИЗ: ОСНОВНАЯ ИДЕЯ

ВЫСШЕЕ	СРЕД.СПЕЦ.	СРЕДНЕЕ
34	32	43
56	44	56
76	57	43
46	87	35
89	91	53
51	43	47
60	74	48
67	73	40
76	68	44
43	35	46
54	63	56
71	49	80
80	21	16
24	67	37
59	78	50
...	...	...
66,9	***65,9***	***70,1***

- $SD = \sigma =$ СТАНДАРТНОЕ ОТКЛОНЕНИЕ
- $SD^2 =$ ДИСПЕРСИЯ (**VARIANCE**)

Оценка общей дисперсии по разбросу МЕЖДУ группами

средние в каждой группе

общее среднее

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

размер групп ы

число групп

MS_B – mean square between groups
оценка расстояния между средними в группах

ВЫСШЕЕ	СРЕД.СПЕЦ.	СРЕДНЕЕ
34	32	43
56	44	56
76	57	43
46	87	35
89	91	53
51	43	47
60	74	48
67	73	40
76	68	44
43	35	46
54	63	56
71	49	80
80	21	16
24	67	37
59	78	50
…	…	…
66,9	***65,9***	***70,1***

ДИСПЕРСИОННЫЙ АНАЛИЗ: ОСНОВНАЯ ИДЕЯ

- $SD = \sigma =$ СТАНДАРТНОЕ ОТКЛОНЕНИЕ
- $SD^2 =$ ДИСПЕРСИЯ (**VARIANCE**)

Оценка общей дисперсии по разбросу ВНУТРИ групп

сумма квадратов стандартных отклонений внутри групп

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

число групп

$$df_W = n_G - k$$

ВЫСШЕЕ	СРЕД.СПЕЦ.	СРЕДНЕЕ
34	32	43
56	44	56
76	57	43
46	87	35
89	91	53
51	43	47
60	74	48
67	73	40
76	68	44
43	35	46
54	63	56
71	49	80
80	21	16
24	67	37
59	78	50
…	…	…
66,9	***65,9***	***70,1***

ДИСПЕРСИОННЫЙ АНАЛИЗ: ОСНОВНАЯ ИДЕЯ

- $SD = \sigma =$ СТАНДАРТНОЕ ОТКЛОНЕНИЕ
- $SD^2 =$ ДИСПЕРСИЯ (**VARIANCE**)

Расчет F-статистики

ANOVA

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

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

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для теста ЛЕВЕНЕ (тест равенства дисперсий)

Но: $\sigma_1 = \sigma_2 = \sigma_3$ (дисперсии средней продолжительности жизни в группах лиц в зависимости от уровня образования равны между собой)

На: $\sigma_1 \neq \sigma_2 \neq \sigma_3$ (дисперсии средней продолжительности жизни в группах лиц в зависимости от уровня образования НЕ равны между собой)

Test of Homogeneity of Variances

Age

Levene Statistic	df1	df2	Sig.
9,963	2	3682	,000

NB:

НЕОБХОДИМА ПОПРАВКА БРОУНА-ФОРСИТА / УЭЛЧА

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для ANOVA

Ho: m1 = m2 = m3

Ha: m1 ≠ m2 / m1 ≠ m3 / m2 ≠ m3

Robust Tests of Equality of Means

Age

	Statistic[a]	df1	df2	Sig.
Welch	25,418	2	1508,262	,000
Brown-Forsythe	28,339	2	2562,151	,000

a. Asymptotically F distributed.

p < 0,0001

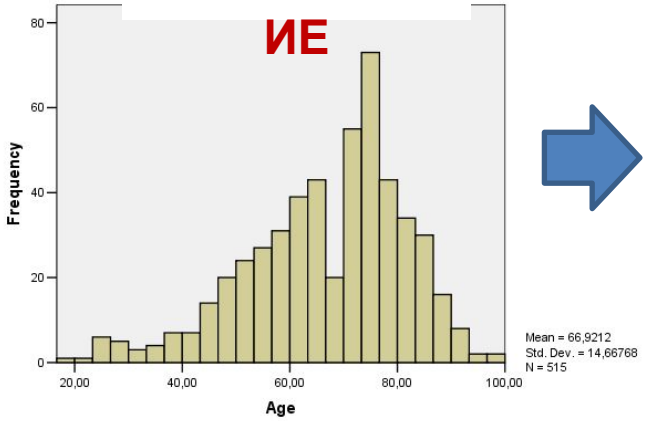
т.е. МОЖЕМ принять Ha вероятность ошибки 1 типа (ошибочно принять Ha - найти то, чего нет) < 0,1%

ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

Но: m1 = m2 = m3

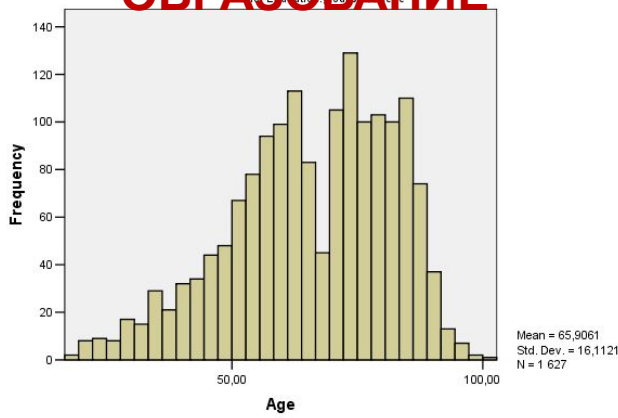
На: m1 ≠ m2 / m1 ≠ m3 / m2 ≠ m3

ВЫСШЕЕ ОБРАЗОВАНИЕ



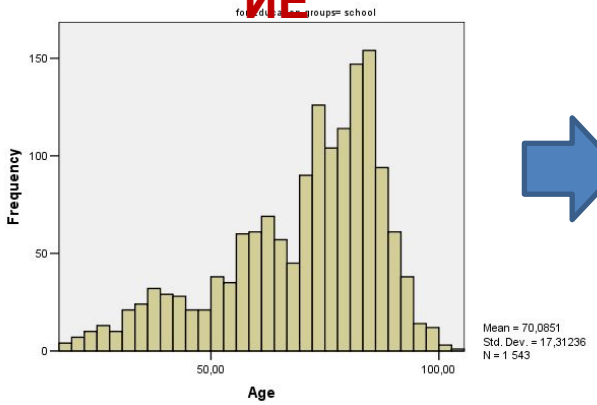
X = 66,9
SD = 14,7
N = 515

СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ



X = 65,9
SD = 16,1
N = 1627

СРЕДНЕЕ ОБРАЗОВАНИЕ



X = 70,1
SD = 17,3
N = 1543

В КАКОЙ ИМЕННО ПАРЕ СРЕДНЯЯ ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ ОТЛИЧАЕТСЯ ???

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

Ha: m1 ≠ m2 / m1 ≠ m3 / m2 ≠ m3

ДАЛЕЕ НЕОБХОДИМА СЕРИЯ
POST HOC тестов

1
H0: m1 = m2
Ha: m1 ≠ m2

2
H0: m1 = m3
Ha: m1 ≠ m3

3
H0: m2 = m3
Ha: m2 ≠ m3

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

ДАЛЕЕ НЕОБХОДИМА СЕРИЯ
POST HOC тестов

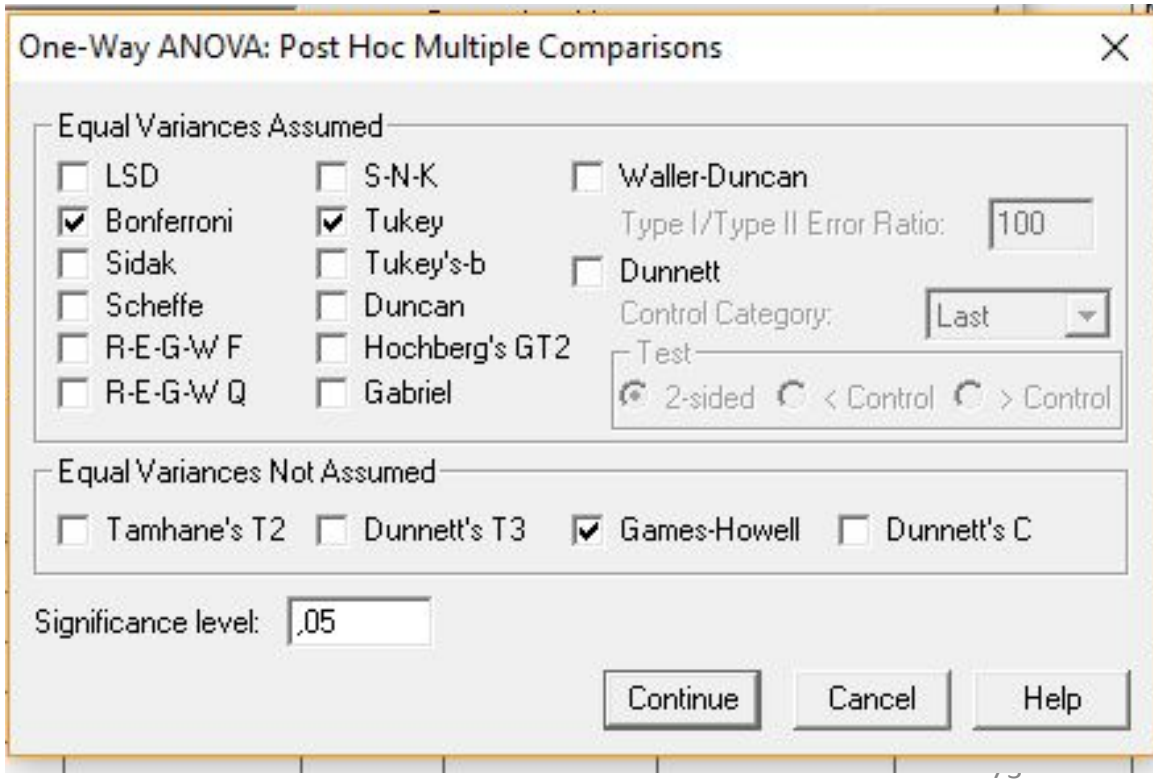
УСЛОВИЕ О РАВЕНСТВЕ ДИСПЕРСИЙ СОБЛЮДЕНО

Bonferroni – если число групп не более 5

Tukey – если число групп более 5

УСЛОВИЕ О РАВЕНСТВЕ ДИСПЕРСИЙ НЕ СОБЛЮДЕНО

Games-Howell – если группы равны, большие группы



ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

ПРОБЛЕМА БОНФЕРРОНИ:
НЕОБХОДИМО ВНЕСТИ ПОПРАВКУ НА КОЛИЧЕСТВО ГРУПП

ОШИБКА 1 ТИПА: $\alpha / n = 0.05/3 = 0,017$

Depend

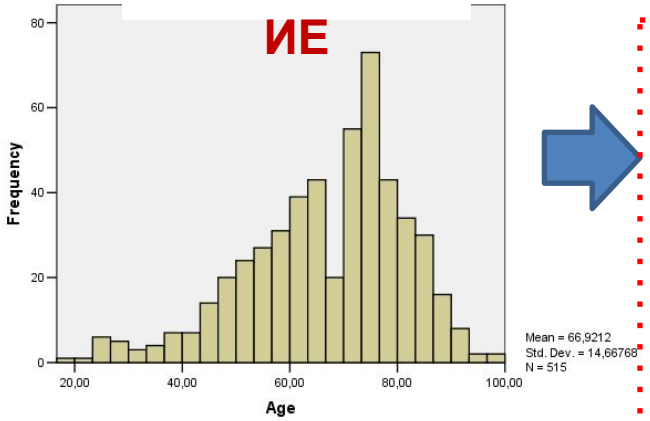
Tukey H

			Difference			95% Confidence Interval	
	school	university	3,16388*	,83652	,000	1,2025	5,1252
		college	4,17901*	,58411	,000	2,8095	5,5485
Bonferroni	university	college	1,01513	,83109	,666	-,9754	3,0057
		school	-3,16388*	,83652	,000	-5,1674	-1,1604
	college	university	-1,01513	,83109	,666	-3,0057	,9754
		school	-4,17901*	,58411	,000	-5,5780	-2,7800
	school	university	3,16388*	,83652	,000	1,1604	5,1674
		college	4,17901*	,58411	,000	2,7800	5,5780
Games-Howell	university	college	1,01513	,75981	,376	-,7685	2,7987
		school	-3,16388*	,78230	,000	-5,0000	-1,3277
	college	university	-1,01513	,75981	,376	-2,7987	,7685
		school	-4,17901*	,59481	,000	-5,5737	-2,7843
	school	university	3,16388*	,78230	,000	1,3277	5,0000
		college	4,17901*	,59481	,000	2,7843	5,5737

*. The mean difference is significant at the .05 level.

ИССЛЕДОВАТЕЛЬСКИЙ ВОПРОС:
УРОВЕНЬ ОБРАЗОВАНИЯ ВЛИЯЕТ НА ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ?

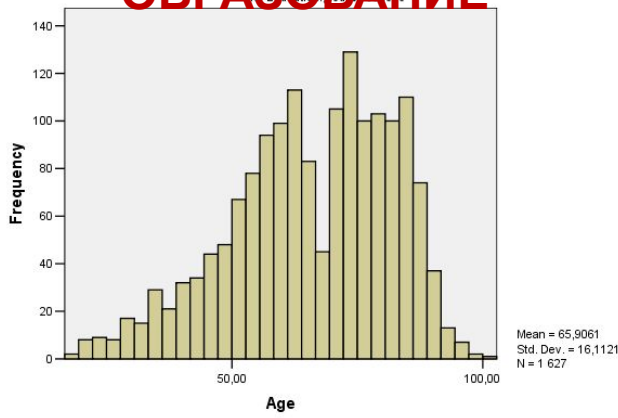
ВЫСШЕЕ ОБРАЗОВАНИЕ



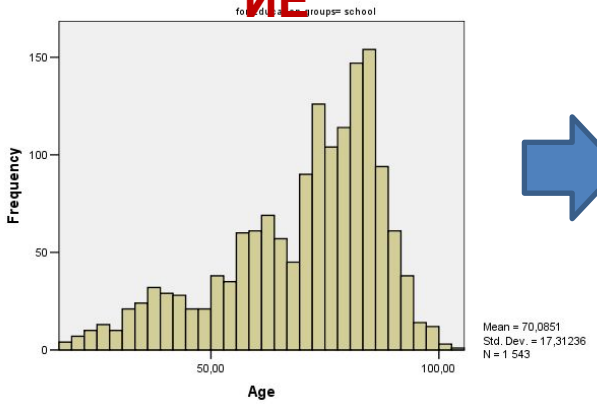
X = 66,9
SD = 14,7
N = 515

X = 65,9
SD = 16,1
N = 1627

СРЕДНЕЕ СПЕЦИАЛЬНОЕ ОБРАЗОВАНИЕ



СРЕДНЕЕ ОБРАЗОВАНИЕ



X = 70,1
SD = 17,3
N = 1543

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

КАК ПРЕДСТАВИТЬ РЕЗУЛЬТАТ («АКАДЕМИЧЕСКАЯ ВЕРСИЯ»)

m1 = 66,9 (95% ДИ: 65,7 – 68,2)

m2 = 65,9 (95% ДИ: 65,1 – 66,7)

m3 = 70,1 (95% ДИ: 69,2 – 70,9)

«…средняя продолжительность жизни зависит от уровня образования человека (F = 25,4 (Welch); p < 0,0001). Продолжительность жизни лиц, имевших среднее образование, была статистически значимо выше, чем у лиц, имевших высшее и среднее специальное образование»; средняя продолжительность жизни лиц, имевших высшее и среднее специальной образование, была равной

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

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

SS - суммы квадратов отклонений (sum of squares):

SS_B - средних в группах от общего среднего = **Effect**

SS_W – измерений от средних в группах = **Error**

«доля объяснённой вариабельности»

R^2 = 0.01 – «незначительный» эффект

R^2 = 0.06 – «средний» эффект

R^2 = 0.14 – «значительный» эффект

5 ЭТАП:
ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ / оценка практической значимости

ONE-WAY ANOVA
ДИСПЕРСИОННЫЙ АНАЛИЗ

$$f = \frac{s_{\overline{X}}}{\sqrt{MS_W}}$$

Общая дисперсия по разбросу ВНУТРИ групп

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

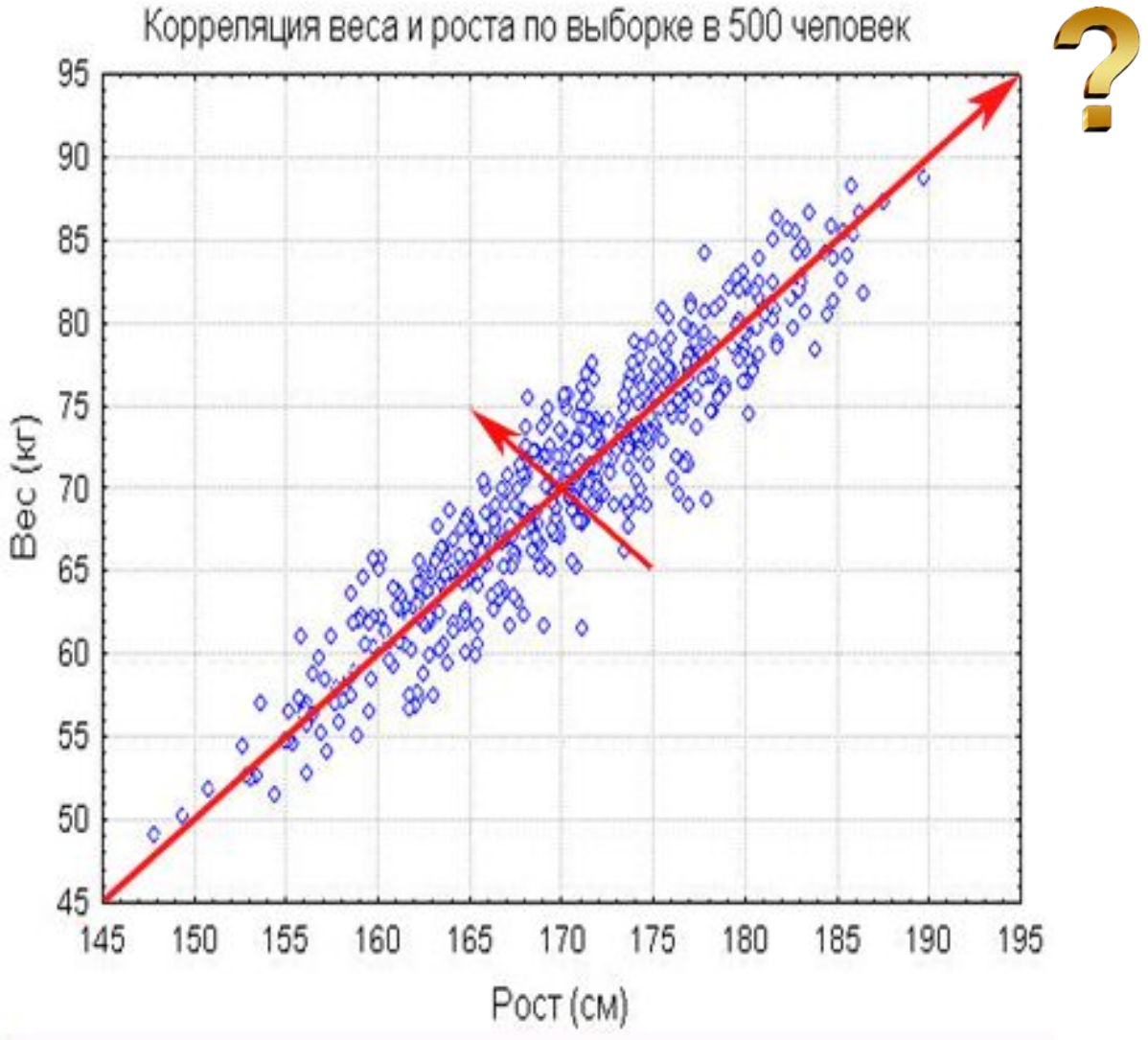
«практическая значимость» результата:

f = 0,1 – «незначительный» эффект
f = 0.25 – «средний» эффект
f = 0.4 – «значительный» эффект

5 ЭТАП:
ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ / оценка практической значимости

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Математическая зависимость величин



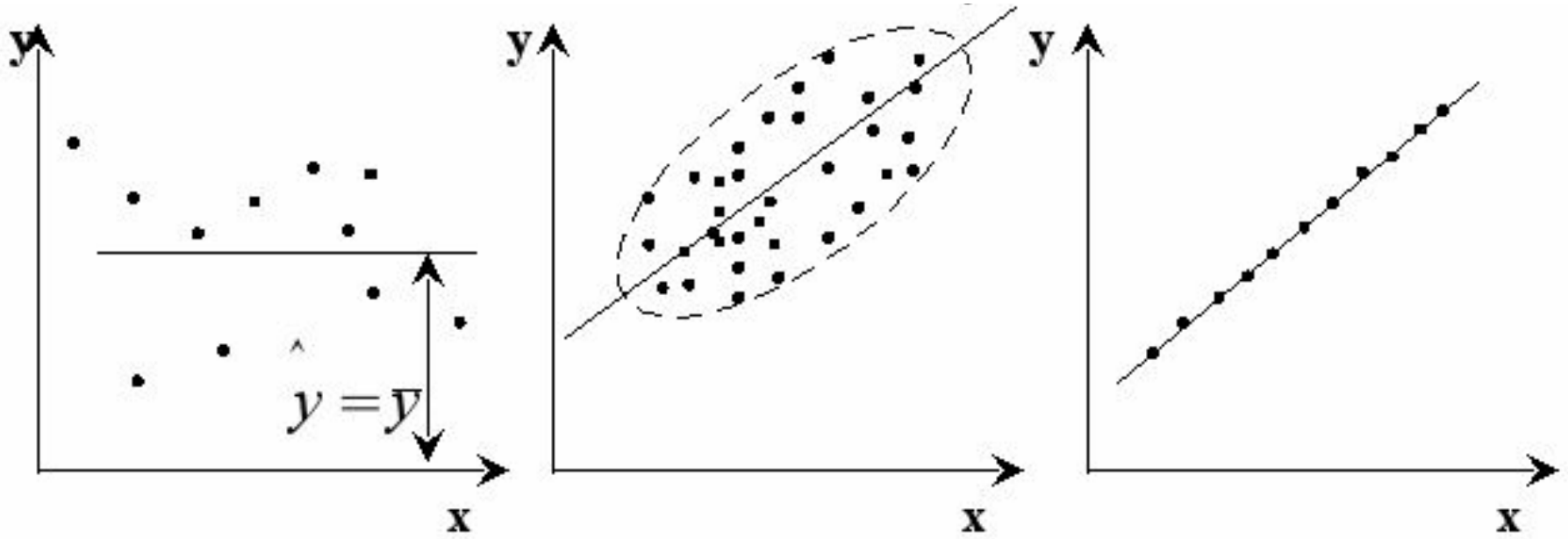
НАПРАВЛЕНИЕ ЗАВИСИМОСТИ:

- Положительная
- Отрицательная

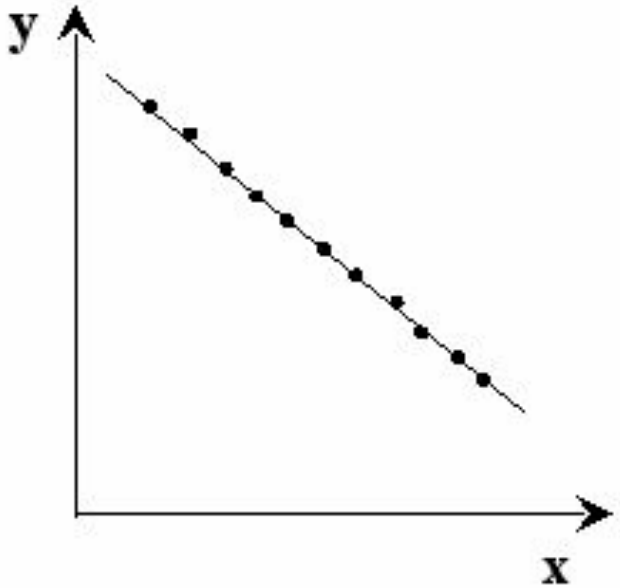
СИЛА ЗАВИСИМОСТИ:

- Отсутствует
- Слабая
- Средняя
- Сильная
- Абсолютная

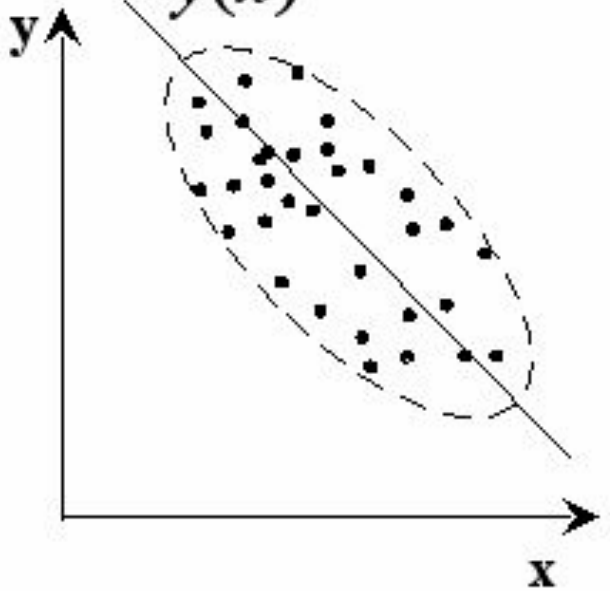
Наличие математической зависимости / корреляции НЕ ОЗНАЧАЕТ наличия ПРИЧИННО-СЛЕДСТВЕННОЙ взаимосвязи между переменными



$\tau_{Y,X} = -1$



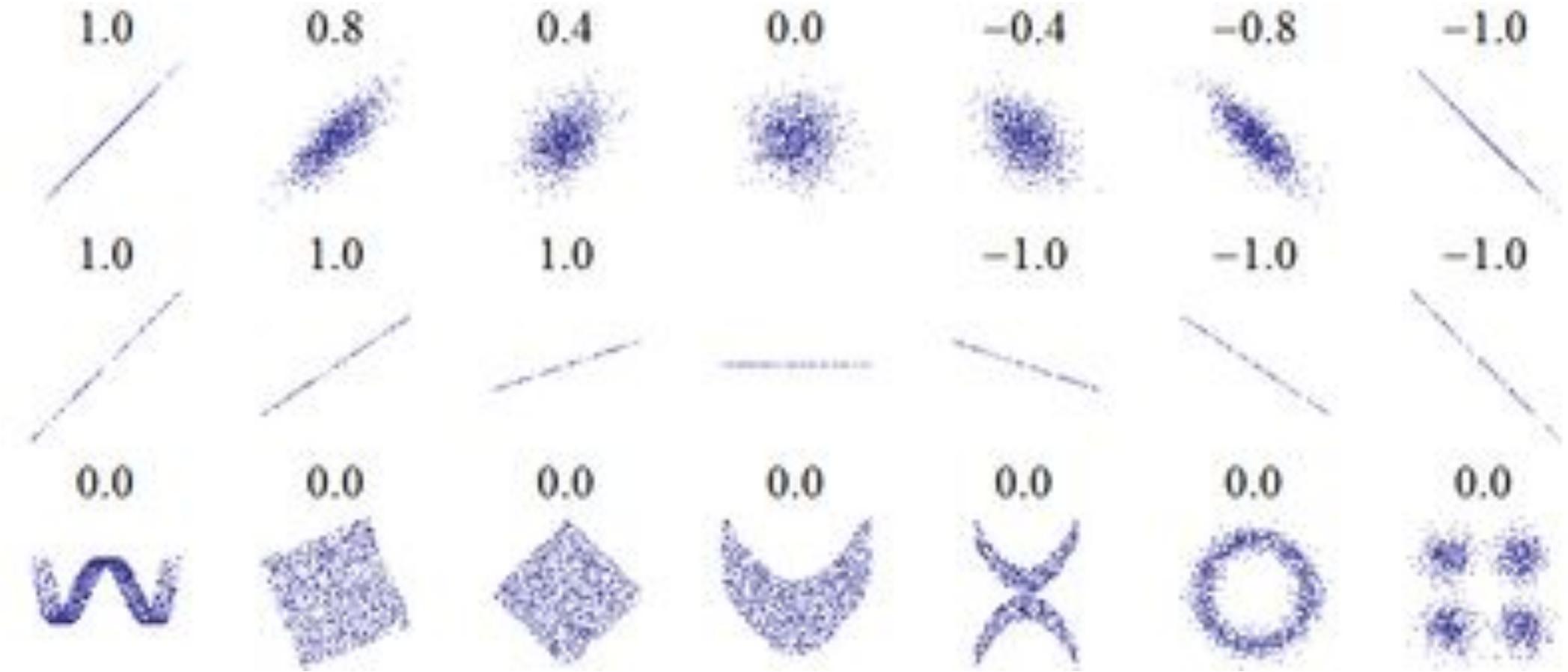
$-1 < \tau_{Y,X} < 0$



ЗАДАНИЕ:

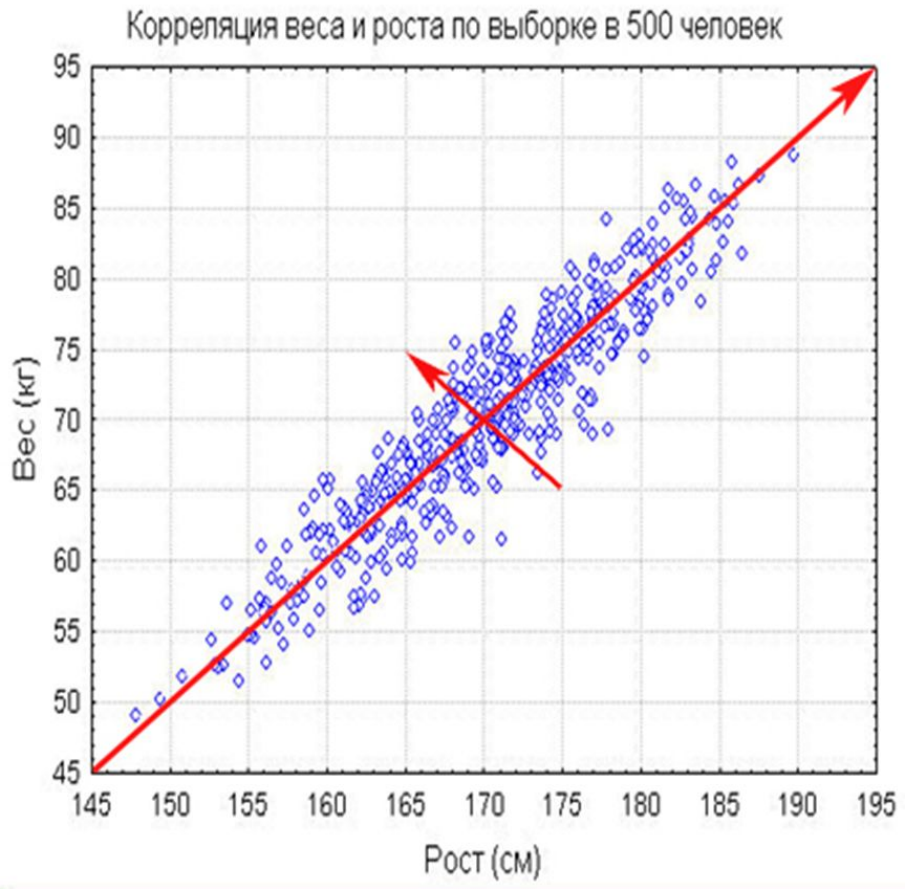
ОПРЕДЕЛИТЬ НАПРАВЛЕНИЕ И СИЛУ ЗАВИСИМОСТИ ПЕРЕМЕННЫХ

МНОЖЕСТВО КОРРЕЛЯЦИОННЫХ ПОЛЕЙ



Множество корреляционных полей.
https://ru.wikipedia.org/wiki/Корреляция

Как можно количественно выразить математическую зависимость 2-х величин ?



КОВАРИАЦИЯ

КОВАРИАЦИЯ – это степень согласованности отклонений двух переменных

$cov(x,y) = \Sigma[(x - среднее x)(y - среднее y)]$

Смысл: если 1 варианта отклоняется от средней, можно ожидать, что 2-я отклонится в ту же сторону

КОРРЕЛЯЦИЯ

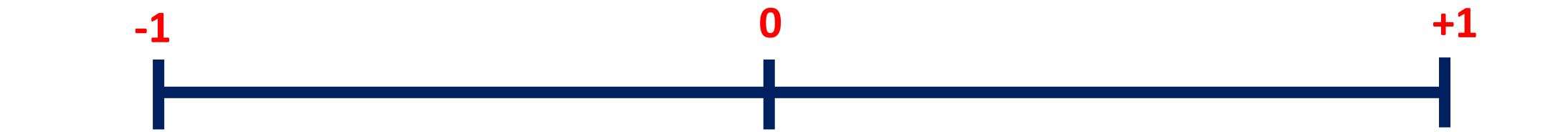
КОРРЕЛЯЦИЯ – это ковариация стандартизованных переменных

$r = cov(x,y) / SDxy$

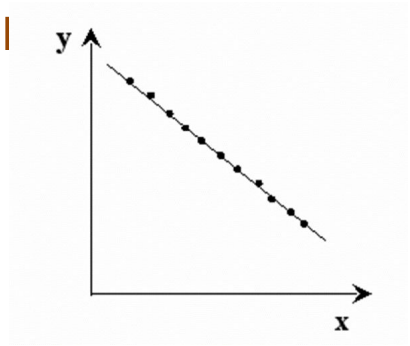
Смысл: отношение наблюдаемой ковариации двух стандартизованных переменных к максимально возможной ковариации

КОРРЕЛЯЦИЯ

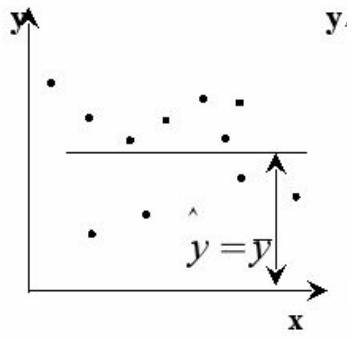
КОРРЕЛЯЦИЯ – это двумерное измерение силы и направления математической взаимосвязи между двумя переменными



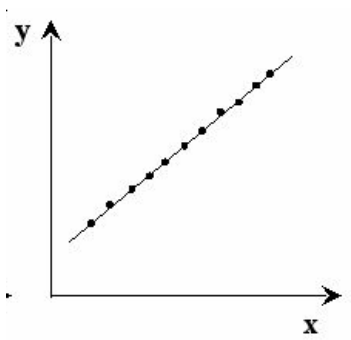
абсолютная негативная линейная связь



случайная связь



абсолютная положительная линейная связь



КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

Единственный параметрический критерий

	Непрерывные	Порядковые	Дихотомические
Непрерывные	Pearson's r Spearman's rho Kendall's tau	Spearman's rho Kendall's tau Polyserial correlation	Polyserial correlation Point-biserial correlation (истинная дихотомия) Biserial correlation (ложная дихотомия)
Порядковые	Spearman's rho Kendall's tau Polyserial correlation	Spearman's rho Kendall's tau Polychoric correlation	Rank biserial correlation
Дихотомические	Polyserial correlation Point-biserial correlation (истинная дихотомия) Biserial correlation (ложная дихотомия)	Rank biserial correlation	Polychoric correlation (tetrachoric correlation) phi

Пример расчета коэффициента корреляции Пирсона

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)
1.	951	83
2.	874	76
3.	957	84
4.	1084	89
5.	903	79

1 ЭТАП. Расчет суммы значений переменных X и Y:

Σ(X) = 951 + 874 + 957 + 1084 + 903 = 4769
Σ(Y) = 83 + 76 + 84 + 89 + 79 = 441

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)
1.	951	83
2.	874	76
3.	957	84
4.	1084	89
5.	903	79

2 ЭТАП. Расчет средних арифметических для X и Y:

Mx = Σ(X) / n = 4769 / 5 = 953.8

My = Σ(Y) / n = 441 / 5 = 82.2

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	*Отклонение содержания тестостерона от среднего значения (d_x)*	*Отклонение % мышечной массы от среднего значения (d_y)*
1.	951	83	*-2.8*	*0.8*
2.	874	76	*-79.8*	*-6.2*
3.	957	84	*3.2*	*1.8*
4.	1084	89	*130.2*	*6.8*
5.	903	79	*-50.8*	*-3.2*

3 ЭТАП. Расчет для каждого значения сопоставляемых показателей величину отклонения от среднего арифметического

dx = X - Mx

dy = Y - My

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения (d_x)	Отклонение % мышечной массы от среднего значения (d_y)	d_x^2	d_y^2
1.	951	83	-2.8	0.8	*7.84*	*0.64*
2.	874	76	-79.8	-6.2	*6368.04*	*38.44*
3.	957	84	3.2	1.8	*10.24*	*3.24*
4.	1084	89	130.2	6.8	*16952,04*	*46.24*
5.	903	79	-50.8	-3.2	*2580,64*	*10.24*

4 ЭТАП. Возвести в квадрат каждое значение отклонения dx и dy

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения (d_x)	Отклонение % мышечной массы от среднего значения (d_y)	d_x^2	d_y^2	$d_x \times d_y$
1.	951	83	-2.8	0.8	7.84	0.64	*-2.24*
2.	874	76	-79.8	-6.2	6368.04	38.44	*494.76*
3.	957	84	3.2	1.8	10.24	3.24	*5.76*
4.	1084	89	130.2	6.8	16952,04	46.24	*885.36*
5.	903	79	-50.8	-3.2	2580,64	10.24	*162.56*

5 ЭТАП. Расчет для каждой пары анализируемых значений произведение отклонений dx x dy:

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

6 ЭТАП. Расчет значения суммы квадратов отклонений $\Sigma(d_x^2)$ и $\Sigma(d_y^2)$

$\Sigma(d_x^2)$ **= 25918.8**

$\Sigma(d_y^2)$ **= 98.8**

7 ЭТАП. Расчет значения суммы произведений отклонений $\Sigma(d_x \times d_y)$

$\Sigma(d_x \times d_y)$ = 1546.2

8 ЭТАП. Расчет значения коэффициента корреляции Пирсона

r_{xy}

$$r_{xy} = \frac{\sum(d_x \times d_y)}{\sqrt{(\sum d_x^2 \times \sum d_y^2)}} = \frac{1546.2}{\sqrt{(25918.8 \times 98.8)}} = 0.966$$

http://medstatistic.ru/theory/pirson.html

Пример расчета коэффициента корреляции Пирсона

9 ЭТАП. Оценка достоверности результата – расчет t-критерия

$$t_r = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.97\sqrt{5-2}}{\sqrt{1-0.97^2}} = 7.0$$

Критическое значение **t-критерия можно найти по специальной статистической таблице**

http://medstatistic.ru/theory/pirson.html

УСЛОВИЯ ПРИМЕНЕНИЯ КОРРЕЛЯЦИИ ПИРСОНА

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	**КАК ПРОВЕРИТЬ?**
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **независимыми**	см. характеристики собранных данных
3. **Количественный непрерывный тип данных** в каждой из сравниваемых выборок	см. тип данных
4. **Нормальное распределение** изучаемого признака в каждой из выборок	Test Shapiro-Wilk / Kolmogorov-Smirnov
5. **Гомоскедастичность** - предполагается, что дисперсия ошибки остается той же самой в любой точке на протяжении всей линейной связи (иначе коэффициент корреляции будет завышаться или, наоборот, занижаться)	***обычно не проверяется***
6. **Линейная связь**	Graphs – Scatter/Dot (точечный график)
7. **Отсутствие «выбросов»**	

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА

Значение коэффициента корреляции r	Интерпретация
0 < r <= 0,2	Очень слабая корреляция
0,2 < r <= 0,5	Слабая корреляция
0,5 < r <= 0,7	Средняя корреляция
0,7 < r <= 0,9	Сильная корреляция
0,9 < r <= 1	Очень сильная корреляция

Корреляция является симметричной, поэтому она не может говорить о направлении каузальной связи

Коэффициент детерминации R^2

R^2 - коэффициент детерминации - доля дисперсии переменной X, объясняемая вариабельностью переменной Y

$$r_{xy} = 0,5$$

$$R^2 = 0,25$$

Таким образом, вариабельность переменной X объясняет 25% вариабельности переменной Y

УСЛОВИЯ ПРИМЕНЕНИЯ КОЭФФИЦИЕНТА РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА, КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ КЕНДАЛЛА (тау)

ASSUMPTIONS / УСЛОВИЯ ПРИМЕНЕНИЯ	КАК ПРОВЕРИТЬ?
1. Сравниваем 2 выборки	см. характеристики собранных данных
2. Выборки д.б. **независимыми**	см. характеристики собранных данных
3. **Количественный непрерывный / порядковый тип данных** в каждой из сравниваемых выборок	см. тип данных
4. **Нормальное / скошенное распределение** изучаемого признака	***можно не проверять***

ОСНОВНОЙ НЕДОСТАТОК КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Коэффициент корреляции демонстрирует

А) направление взаимосвязи переменных

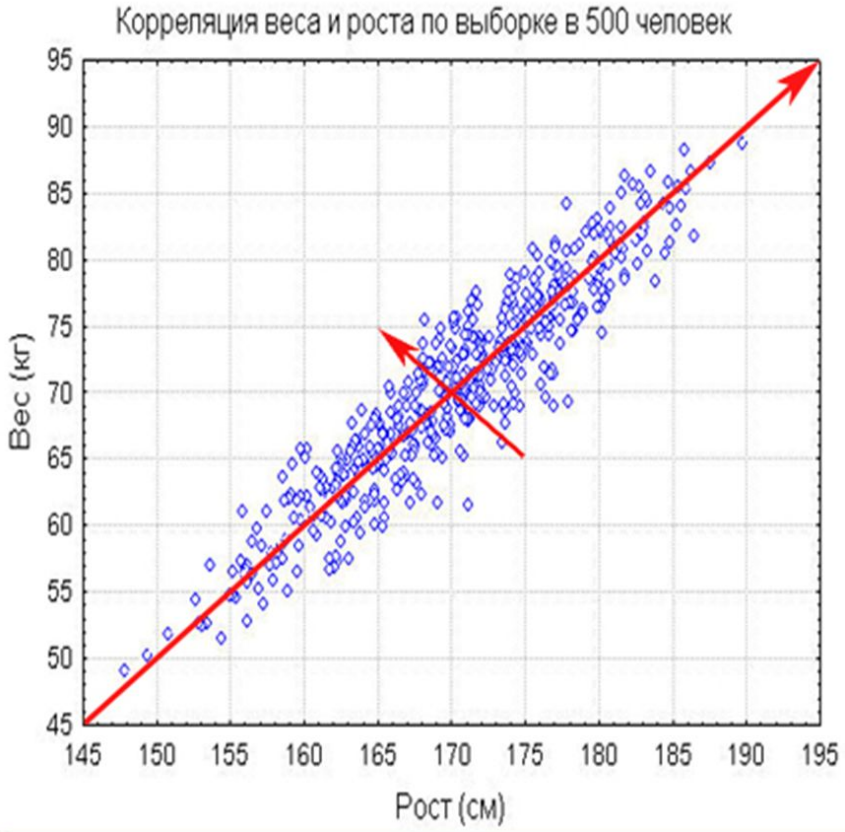
Б) силу взаимосвязи переменных

НО коэффициент корреляции бесполезен, если мы хотим ПРЕДСКАЗАТЬ значение переменной X по значению переменной Y

РЕГРЕССИОННЫЙ АНАЛИЗ

ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

КОРРЕЛЯЦИЯ VS. РЕГРЕССИЯ



МЕЖДУ ПЕРЕМЕННЫМИ ЕСТЬ ЗАВИСИМОСТЬ?

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ – демонстрирует лишь направление взаимосвязи переменных и силу взаимосвязи переменных

ИССЛЕДОВАТЕЛЯ МОГУТ ДОПОЛНИТЕЛЬНО ИНТЕРЕСОВАТЬ ВОПРОСЫ:

1) как сильно влияет на зависимую (1) переменную

А) другая (1) независимая переменная?

Б) одновременно 2 и > независимых переменных?

2) какие именно переменные влияют на зависимую переменную (отсеять из набора переменных «лишние»)?

3) какие именно переменные влияют одновременно на 2 и более зависимых переменных из набора?

4) можно ли по значениям одной (нескольких) переменных ПРЕДСКАЗАТЬ значение другой (других) переменных

РЕГРЕССИЯ: ОСНОВНАЯ ИДЕЯ

$Y = f(X)$

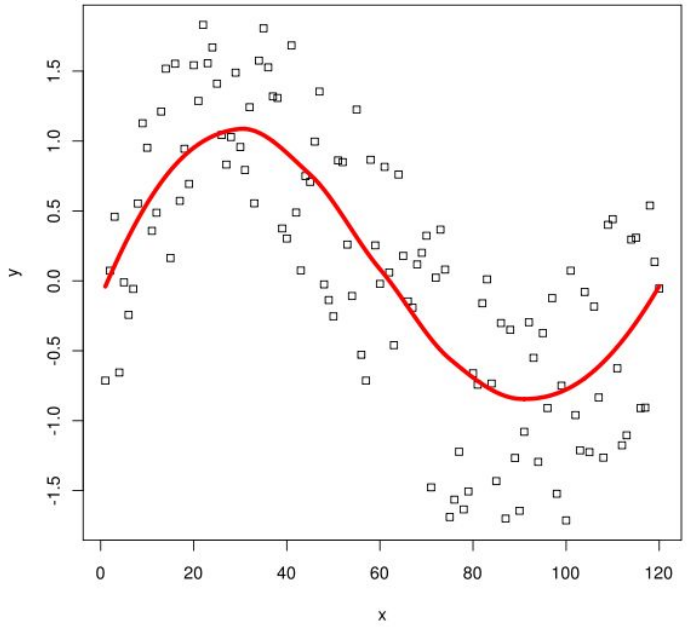
Зависимость между переменными может быть выражена *УРАВНЕНИЕМ*

ОСНОВНАЯ ИДЕЯ РЕГРЕССИИОННОГО АНАЛИЗА:

математически рассчитать параметры УРАВНЕНИЯ РЕГРЕССИИ

(с какой силой / в каком направлении переменные влияют на зависимую переменную)

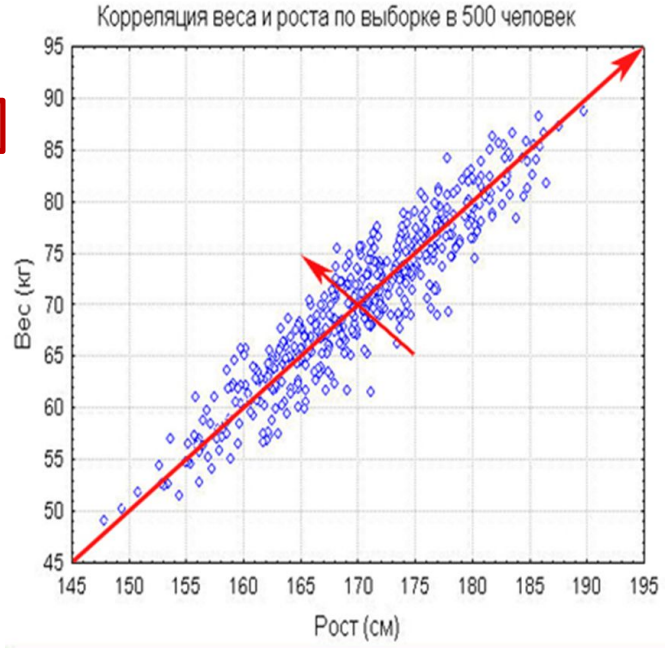
РЕГРЕССИЯ: ОСНОВНАЯ ПРОБЛЕМА



нелинейная зависимость

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

НЕЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ



линейная зависимость

КАКАЯ ФОРМА ЗАВИСИМОСТИ ОДНОЙ ПЕРЕМЕННОЙ ОТ ДРУГОЙ ПЕРЕМЕННОЙ?

КАКАЯ ФОРМА ЗАВИСИМОСТЬ ОДНОЙ ПЕРЕМЕННЫХ ОТ НЕСКОЛЬКИХ ПЕРЕМЕННЫХ?

ВЫБОР МОДЕЛИ РЕГРЕССИОННОГО АНАЛИЗА

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_N X_N + E$$

ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

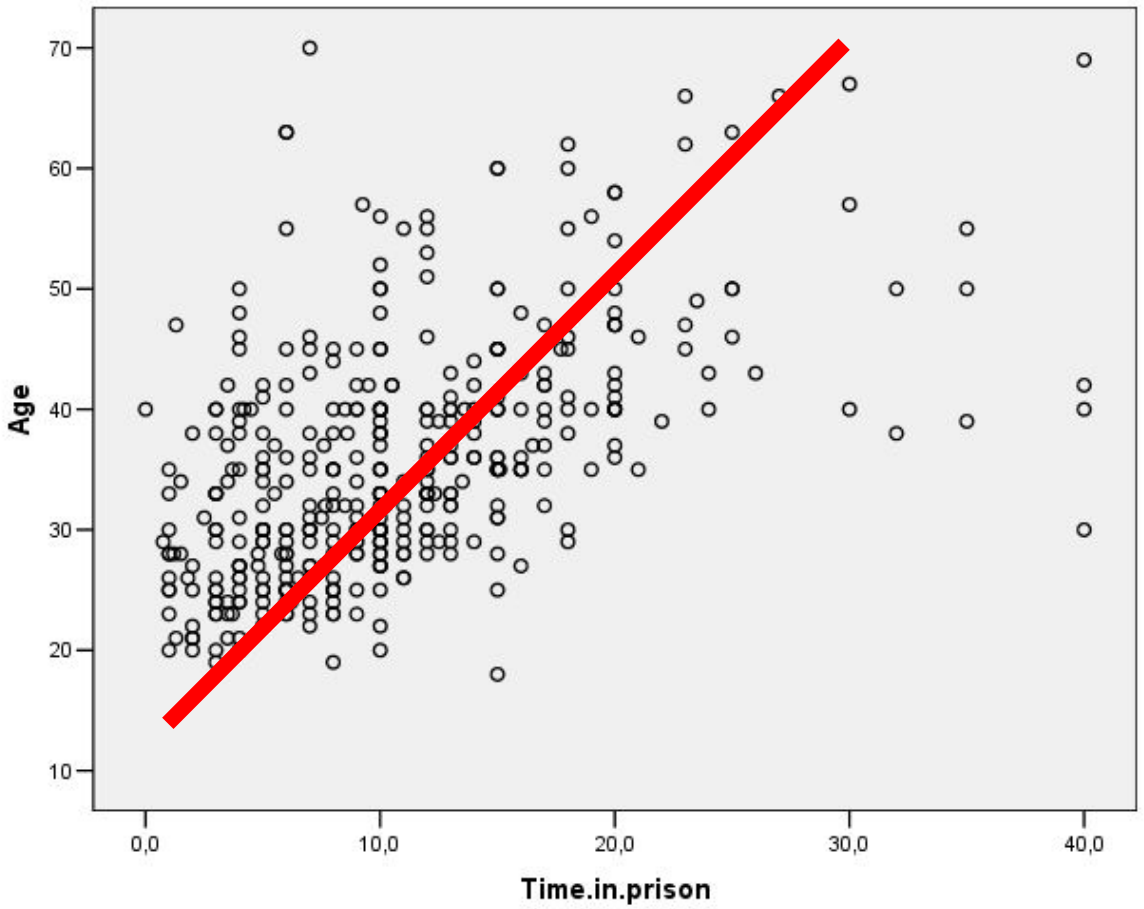
Y – зависимая переменная / переменная отклика

B0 – константа

Bn – коэффициент регрессии / градиент

E - ошибка

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

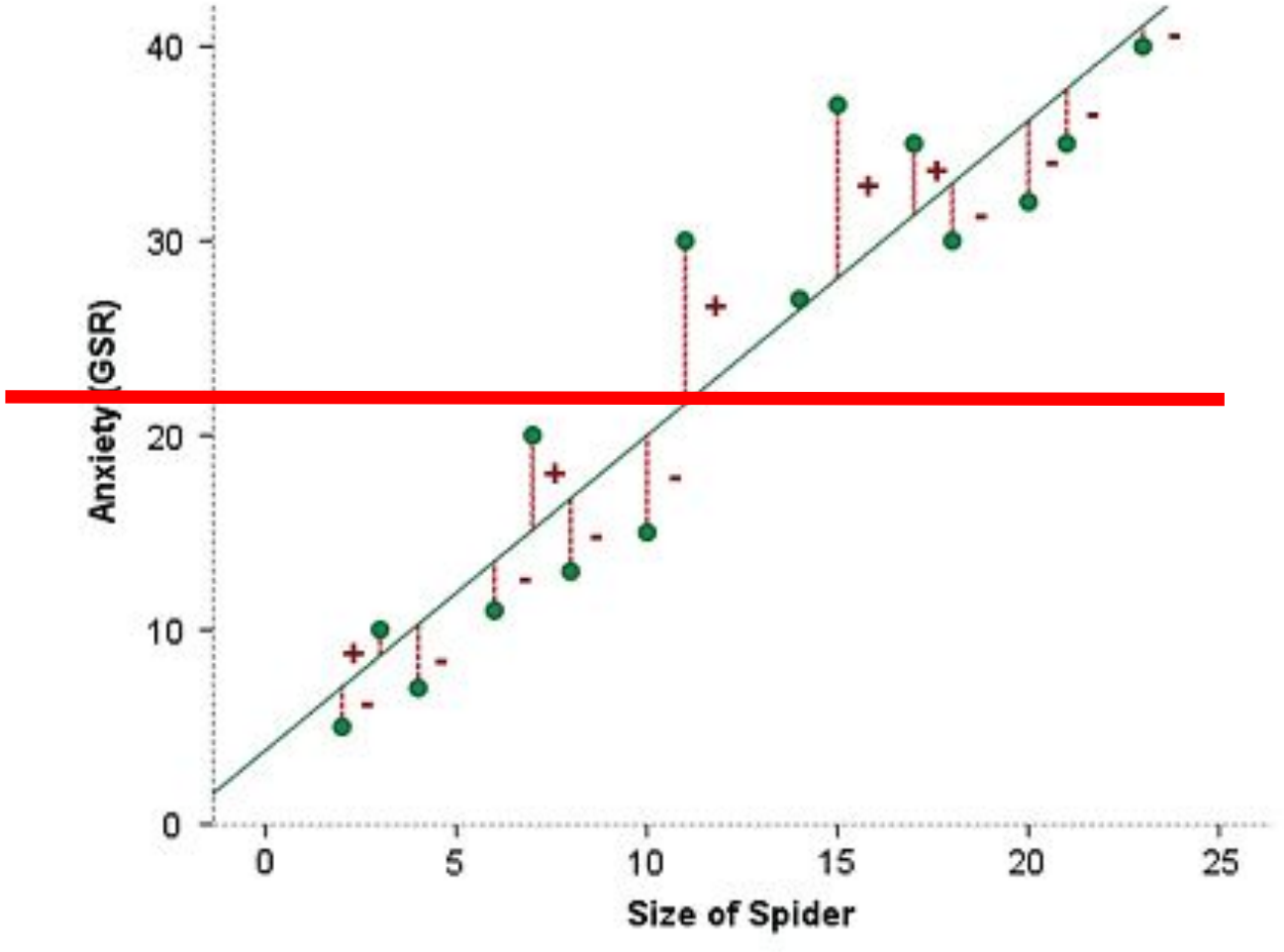


Идея LRA: построить прямую, наиболее точно предсказывающую значение зависимой переменной от предиктора (-ов) (и рассчитать ее параметры, т.е. ФОРМУЛУ)

– «линейный» анализ

В ЭТОМ «МИНУС» ЛРА – в природе нет линейной зависимости (тем более 1 зависимой переменной от нескольких)

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ



H0 (LRA):

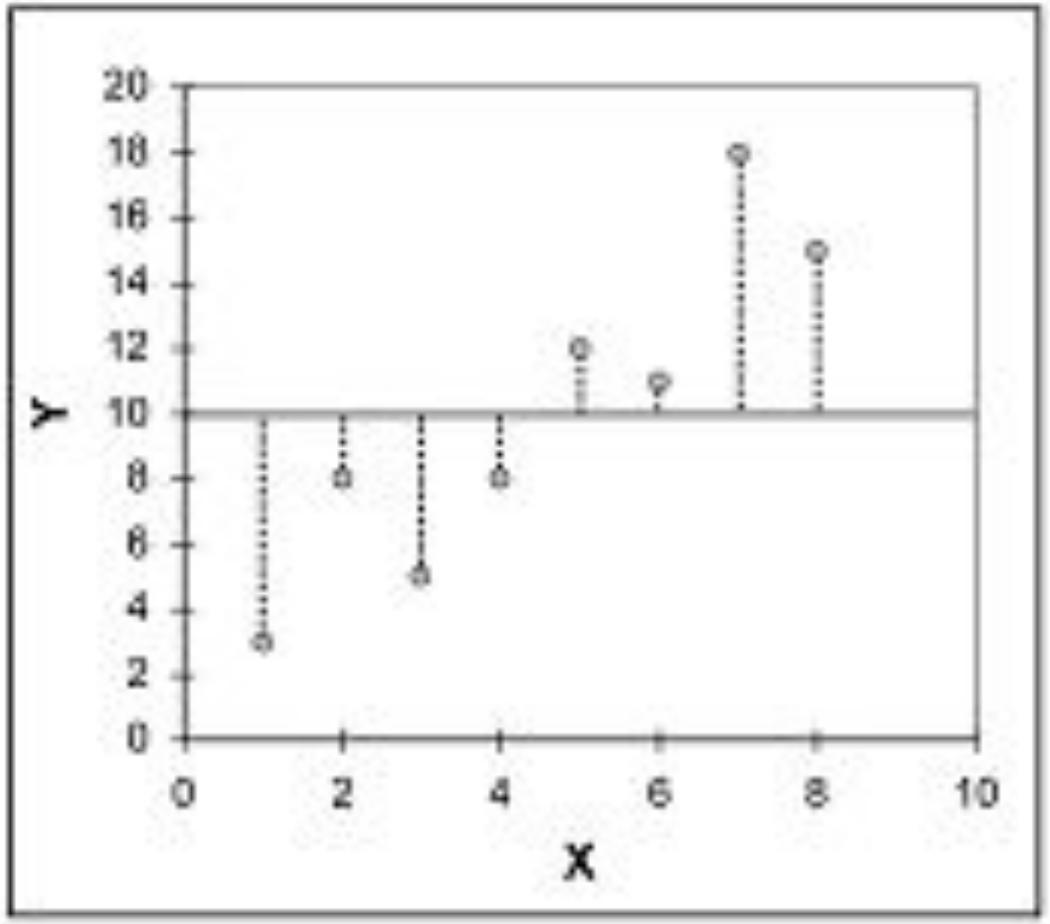
- Зависимая переменная лучше всего описывается средней арифметической

Ha (LRA):

- Зависимая переменная лучше всего описывается некоторой линейной моделью

Далее программа (по методу «наименьших квадратов») «подбирает» линию (модель), которая наилучшим образом «предсказывает» зависимую переменную по значению независимого предиктора

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ



СНАЧАЛА ПРОГРАММА АНАЛИЗИРУЕТ,

НАСКОЛЬКО ХОРОШО СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ (H0) ПРЕДСКАЗЫВАЕТ ЗАВИСИМУЮ ПЕРЕМЕННУЮ:

SSт

общая сумма различий между фактическими данными и средней арифметической

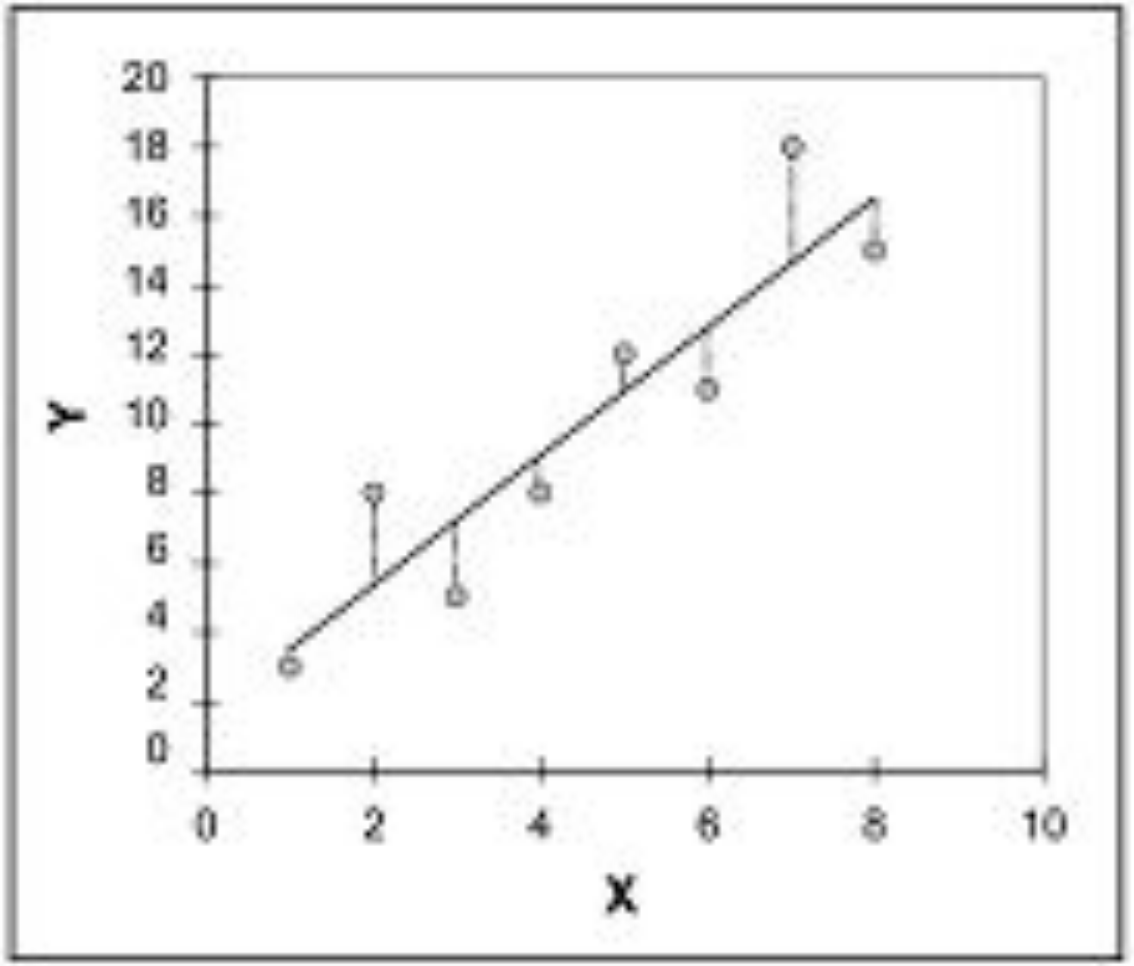
ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

ДАЛЕЕ ПРОГРАММА АНАЛИЗИРУЕТ,

НАСКОЛЬКО ХОРОШО МОДЕЛЬ (На) ПРЕДСКАЗЫВАЕТ ЗАВИСИМУЮ ПЕРЕМЕННУЮ

SS_R

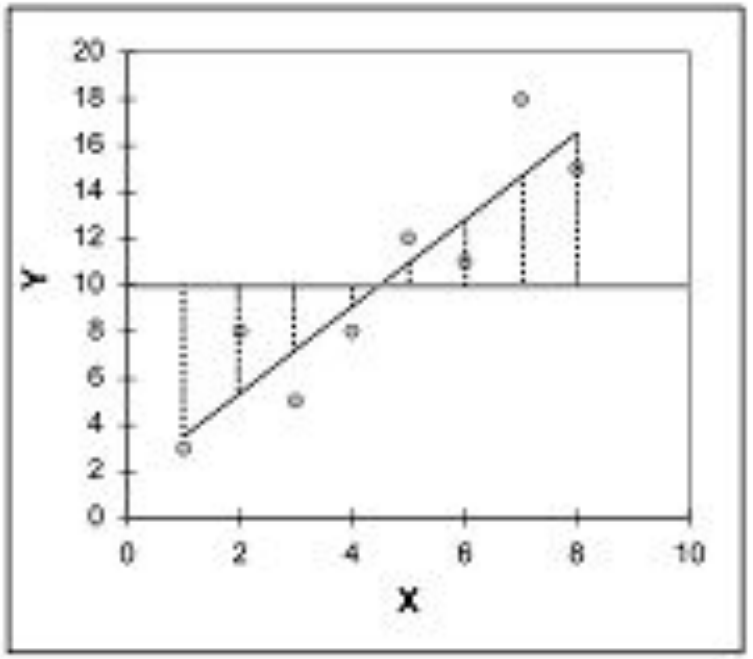
общая сумма различий между фактическими данными и моделью



ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

ДАЛЕЕ ПРОГРАММА АНАЛИЗИРУЕТ,

НАСКОЛЬКО ХОРОШО МОДЕЛЬ (на) ПРЕДСКАЗЫВАЕТ ЗАВИСИМУЮ ПЕРЕМЕННУЮ В СРАВНЕНИИ С ПРОСТОЙ СРЕДНЕЙ АРИФМЕТИЧЕСКОЙ (н0)



$$SS_M = SS_T - SS_R$$

ПОКАЗЫВАЕТ УЛУЧШЕНИЕ В ПРЕДСКАЗАТЕЛЬНОЙ СИЛЕ МОДЕЛИ В СРАВНЕНИИ С ПРОСТОЙ СРЕДНЕЙ АРИФМЕТИЧЕСКОЙ

$$RR^2 = \frac{SS(M)}{SS(T)}$$

RR^2 - показывает количество дисперсии, которая объясняется моделью

$$1 - RR^2 = \text{ERROR}$$

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

ПРОГРАММА РАССЧИТЫВАЕТ СТАТИСТИКУ РЕГРЕССИОННОЙ МОДЕЛИ (F – TEST)

СПОСОБНОСТЬ МОДЕЛИ УЛУЧШАТЬ ПРЕДСКАЗАНИЕ ЗАВИСИМОЙ ПЕРЕМЕННОЙ В СРАВНЕНИИ СО СРЕДНЕЙ АРИФМЕТИЧЕСКОЙ

$$F_{test} = \frac{MS(M)}{MS(R)}$$

$$MS(M) = \frac{SS(M)}{df(M)}$$

$$MS(R) = \frac{SS(R)}{df(R)}$$

p (F – test) < 0,05
МОДЕЛЬ «РАБОТАЕТ», т.е.
предсказывает зависимую переменную лучше, чем средняя арифметическая
(H_0)

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_N X_N + E$$

СТАТИСТИКА РЕГРЕССИОННОЙ МОДЕЛИ (F – TEST)
демонстрирует статистическую значимость всего уравнения регрессии

Bn – коэффициент регрессии / градиент - демонстрирует изменение значения зависимой переменной (Y) при изменении предиктора (Xn) на “1” (единицу)

Статистическую значимость каждого коэффициента регрессии необходимо оценить

H0: b1 = 0
Ha: b1 ≠ 0

УСЛОВИЯ ПРИМЕНЕНИЯ (ASSUMPTIONS) ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА

А) ЗАВИСИМАЯ ПЕРЕМЕННАЯ: **количественная непрерывная (неограниченная)**

Б) НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ (ПРЕДИКТОРЫ): **количественные непрерывные и дихотомические (0;1)**

В) ЛИНЕЙНАЯ СВЯЗЬ: Graphs – Scatter/Dot (можно проверить для простой регрессии)

Г) ГОМОСКЕДАСТИЧНОСТЬ - предполагается, что дисперсия ошибки остается той

же самой в любой точке на протяжении всей линейной связи

Д) НЕЗАВИСИМЫЕ НАБЛЮДЕНИЯ (DURBIN-WATSON ≈ 2 (DW ϵ [1;3]))

Е) НОРМАЛЬНО РАСПРЕДЕЛЕННЫЕ ОСТАТКИ (residuals)

Ж) НЕ Д.Б. МУЛЬТИКОЛЛИНЕАРНОСТИ (R > 0,8 – проблема; VIF > 10 - проблема)

УСЛОВИЯ ПРИМЕНЕНИЯ (ASSUMPTIONS) ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА

OUTLIER: случаи, значительно влияющие на тренд (>2,58 – проблема)

INFLUENTIAL CASES: случаи, заметно влияющие на модель (ее значимость)

COOK'S DISTANCE – д.б. < 1 – мера влияния случая на модель

MAHALANOBIS – разница м/д случаем и средней арифметической

N = 500 – д.б. < 25

N = 100 – д.б. < 15

N = 30 – д.б. < 11

ПРИМЕР ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА

ВЕЛИЧИНА РАСХОДОВ ПАЦИЕНТОВ НА МЕДИКАМЕНТЫ

ПОЛ ПАЦИЕНТА

ВОЗРАСТ ПАЦИЕНТА

ДОХОД ПАЦИЕНТА

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

1 ЭТАП:
ФОРМУЛИРУЕМ H0 и Ha

ГИПОТЕЗЫ	ФОРМУЛИРОВКА
H0 (нулевая гипотеза)	**простая средняя арифметическая предсказывает исход лучше, чем модель регрессии**
Ha (альтернативная гипотеза)	**модель регрессии предсказывает исход лучше, чем простая средняя арифметическая**

2 ЭТАП:
ОПРЕДЕЛЯЕМ УСЛОВИЯ, ПРИ КОТОРЫХ ПРИМЕМ Ha (ОТВЕРГНЕМ H0)

БУДЕМ считать результаты теста «статистически значимыми» (т.е. примем Ha) при вероятности ошибки 1 типа (α-ошибки) менее 0.05 (5%)

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

3 ЭТАП:
ВЫБОР СТАТИСТИЧЕСКОГО КРИТЕРИЯ / МЕТОДА

ПОДХОДИТ МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

зависимая переменная: количественная непрерывная
ВЕЛИЧИНА РАСХОДОВ ПАЦИЕНТОВ НА МЕДИКАМЕНТЫ

предикторы: количественная непрерывная / дихотомическая
ПОЛ ПАЦИЕНТА: дихотомическая
ВОЗРАСТ ПАЦИЕНТА: количественная непрерывная
BMI ПАЦИЕНТА: количественная непрерывная

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha

Но: F-статистика модели стат.незначима

На: F-статистика модели стат.значима

МОДЕЛЬ РЕГРЕССИИ «РАБОТАЕТ»
(описывает данные лучше, чем средняя арифметическая)

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4E+008	3	116900381,9	30,579	,000[a]
	Residual	4E+009	1024	3822872,805		
	Total	4E+009	1027			
2	Regression	3E+008	2	173363139,4	45,347	,000[b]
	Residual	4E+009	1025	3823021,092		
	Total	4E+009	1027			

a. Predictors: (Constant), BMI, Gender_new, Age

b. Predictors: (Constant), Gender_new, Age

c. Dependent Variable: Drug_spendings

p < 0,0001

т.е. МОЖЕМ принять На
вероятность ошибки 1 типа (ошибочно принять На - найти то, чего нет) < 0,1%

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

МОДЕЛЬ ОБЪЯСНЯЕТ 8,1% ДИСПЕРСИИ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,287[a]	,082	,080	1955,217	,082	30,579	3	1024	,000	
2	,285[b]	,081	,079	1955,255	-,001	1,040	1	1024	,308	1,997

a. Predictors: (Constant), BMI, Gender_new, Age

b. Predictors: (Constant), Gender_new, Age

c. Dependent Variable: Drug_spendings

ПОРЯДОК ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4 ЭТАП:
МАТЕМАТИЧЕСКИЕ РАССЧЕТЫ

формулируем H0 и Ha для t-статистики коэффициентов b

Ho: t-статистика b стат.незначима
Ha: t-статистика b стат.значима

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	288,542	344,777		,837	,403	-388,009	965,093		
	Age	29,726	3,986	,235	7,457	,000	21,903	37,548	,901	1,109
	Gender_new	-470,182	123,483	-,115	-3,808	,000	-712,490	-227,874	,977	1,023
	BMI	12,029	11,797	,032	1,020	,308	-11,120	35,179	,909	1,100
2	(Constant)	551,149	229,234		2,404	,016	101,328	1000,970		
	Age	30,885	3,821	,244	8,083	,000	23,387	38,383	,981	1,019
	Gender_new	-478,291	123,229	-,117	-3,881	,000	-720,100	-236,481	,981	1,019

a. Dependent Variable: Drug_spendings

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	288,542	344,777		,837	,403	-388,009	965,093		
	Age	29,726	3,986	,235	7,457	,000	21,903	37,548	,901	1,109
	Gender_new	-470,182	123,483	-,115	-3,808	,000	-712,490	-227,874	,977	1,023
	BMI	12,029	11,797	,032	1,020	,308	-11,120	35,179	,909	1,100
2	(Constant)	551,149	229,234		2,404	,016	101,328	1000,970		
	Age	30,885	3,821	,244	8,083	,000	23,387	38,383	,981	1,019
	Gender_new	-478,291	123,229	-,117	-3,881	,000	-720,100	-236,481	,981	1,019

a. Dependent Variable: Drug_spendings

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_N X_N$$

Y = расходы на медикаменты
B0 = CONSTANT = 551,1
B1 = ВОЗРАСТ = 30,9
B2 = ПОЛ = -478,3 (для мужчин)

ДЛЯ 50-ЛЕТНЕГО МУЖЧИН ВЕЛИЧИНА РАСХОДОВ НА МЕДИКАМЕНТЫ

РАСХОДЫ = 551,1 + 30,9 × 50 – 478,3 = 1617,8 руб. + ERROR

КРАТКИЙ ОБЗОР МЕТОДОВ СТАТИСТИЧЕСКОГО АНАЛИЗА КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ