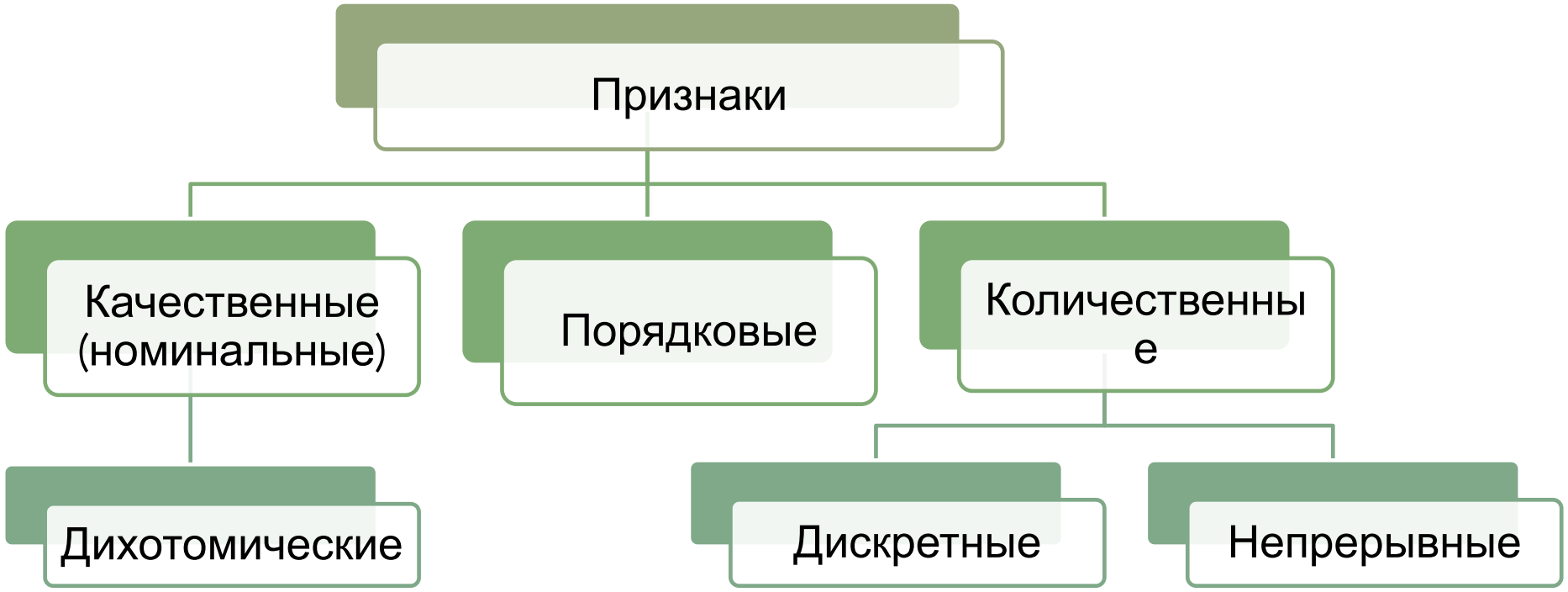
СТАТИСТИКА

ПОДГОТОВИЛИ:
СТУДЕНТКА 6 КУРСА
ЧЕТВЕРТАКОВА СВЕТЛАНА
И СТУДЕНТКА 5 КУРСА
ГАЛУС АННА

ПРИЗНАКИ

– это единицы совокупности, обладающие определенными свойствами и качествами.



О.Ю. Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера, 2002. – 312 с.

КАЧЕСТВЕННЫЕ ПРИЗНАКИ (НОМИНАЛЬНЫЕ)

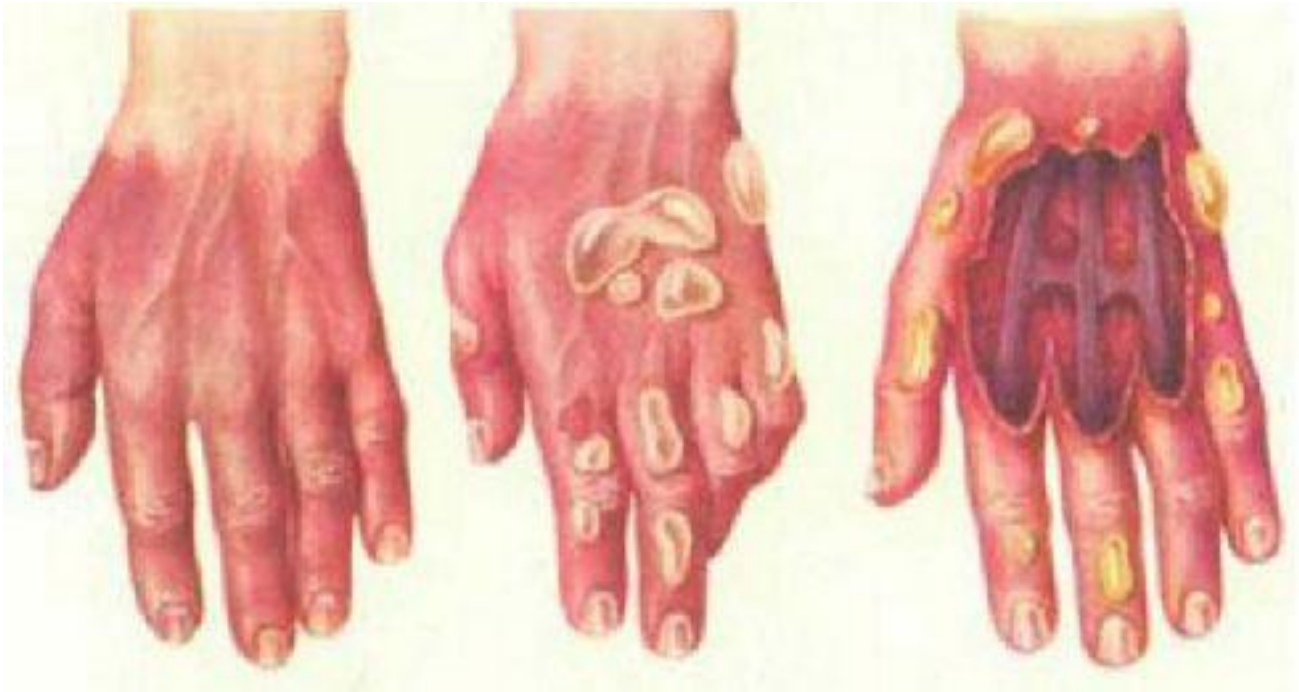
- это такие признаки, которые не поддаются непосредственному измерению.

КАЧЕСТВЕННЫЕ ПРИЗНАКИ

Разновидностью качественных признаков, которые могут быть отнесены только к двум противоположным категориям «да – нет», принимающие одно из двух значений называются **дихотомическими.**

ПОРЯДКОВЫЕ ПРИЗНАКИ

- это признаки, которые можно расположить в естественном порядке (ранжировать), но при этом отсутствует количественная мера расстояния между величинами.



КОЛИЧЕСТВЕННЫЕ ПРИЗНАКИ

– признаки, количественная мера которых четко определена.

Количественные признаки

Непрерывные признаки

Дискретные признаки

НЕПРЕРЫВНЫЕ ПРИЗНАКИ

Интервальные – признаки, измеряющиеся в абсолютных величинах, имеющих физический смысл.

Относительные – признаки, отражающие долю измерения (увеличение или уменьшение), значения признака по отношению к исходному значению этого признака.

ВИД РАСПРЕДЕЛЕНИЯ

- соответствие, устанавливаемое между всеми возможными числовыми значениями случайной величины и вероятностями их появления в совокупности.

Может быть представлен:

1. аналитической зависимостью в виде формулы;
2. в виде графического изображения;
3. в виде таблицы.

Виды распределения

Дискретные

- Биноминальные
- Распределение Пуассона
- Распределение Бернулли

Непрерывные

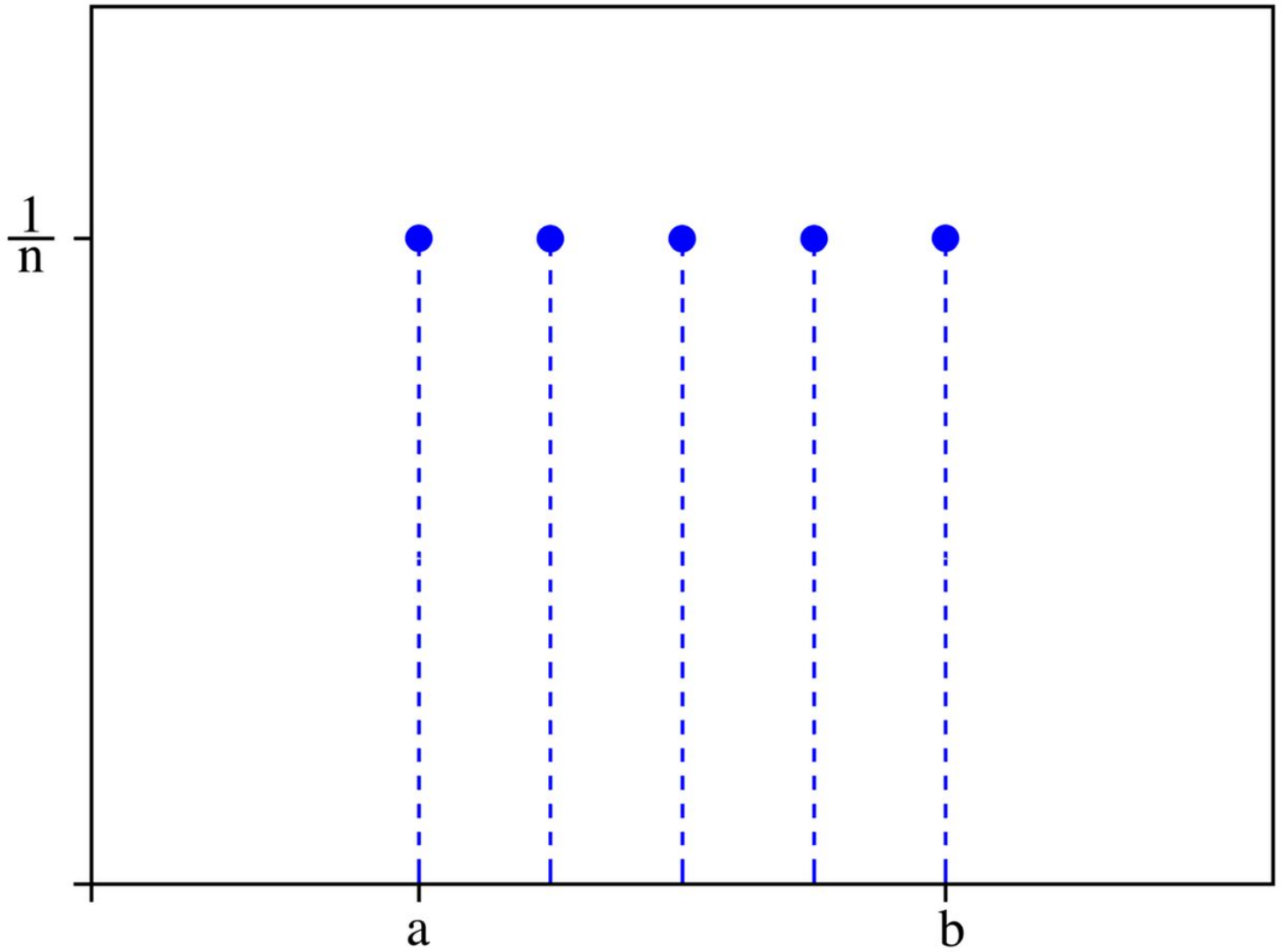
- Нормальное
- Логнормальное
- Постоянное
- Экспоненциальное
- Хи-квадрат и другие

О.Ю. Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера, 2002. – 312 с.

ДИСКРЕТНОЕ РАСПРЕДЕЛЕНИЕ

Используются для описания событий с недифференцируемыми характеристиками, определёнными в изолированных точках.

ДИСКРЕТНОЕ РАСПРЕДЕЛЕНИЕ



БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Описывает распределение частоты события, обладающего постоянной вероятностью появления при многократных испытаниях.

То есть это распределение количества «успехов» в последовательности из некоторого числа независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна.

РАСПРЕДЕЛЕНИЕ ПУАССОНА

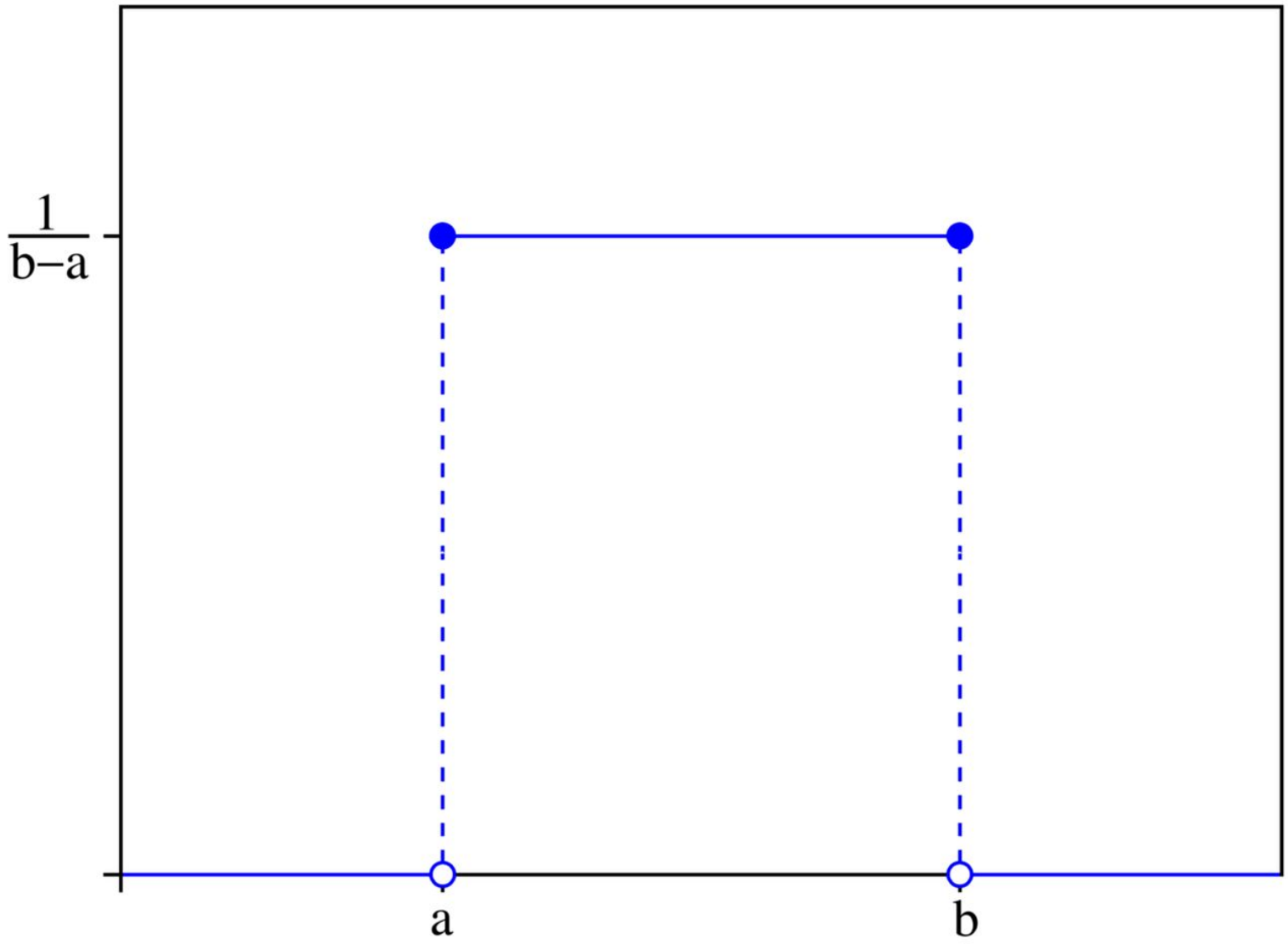
Описывает события, при которых с возрастанием значения случайной величины, вероятность появления ее в совокупности резко уменьшается. Характерно для редких событий.

НЕПРЕРЫВНОЕ РАСПРЕДЕЛЕНИЕ

- это распределение случайной вещественной величины, принимающей значения, принадлежащие некоторому промежутку конечной длины, характеризующееся тем, что плотность вероятности на этом промежутке почти всюду постоянна.

По другому, непрерывной называется случайная величина, которая может принимать любые значения внутри некоторого интервала (масса, температура, рост)

НЕПРЕРЫВНОЕ РАСПРЕДЕЛЕНИЕ

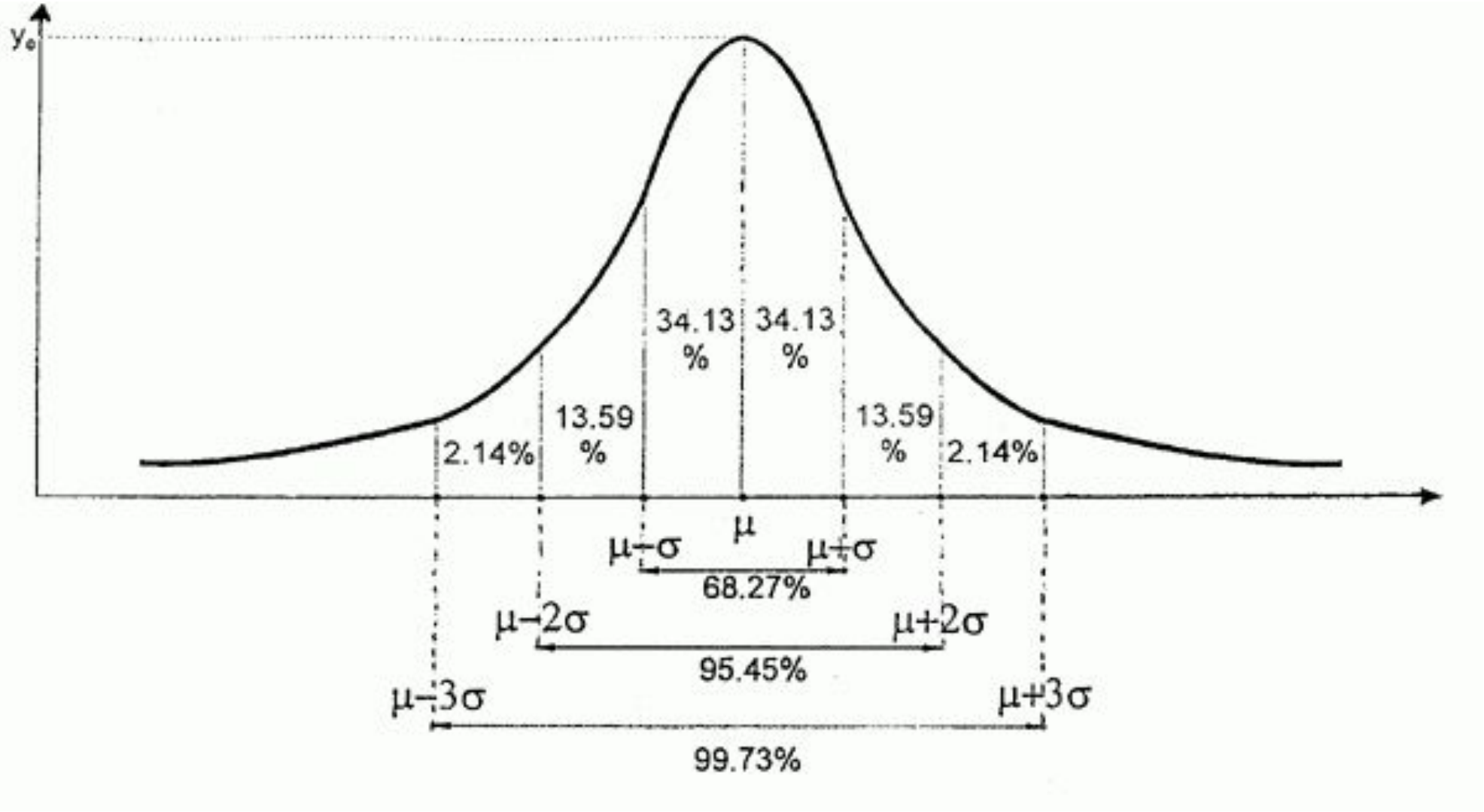


НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ (ГАУССОВО, СИММЕТРИЧНОЕ, КОЛОКОЛООБРАЗНОЕ)

Описывает совместное воздействие на изучаемое явление небольшого числа случайно сочетающихся факторов (по сравнению с общей суммой факторов), число которых неограниченно велико.

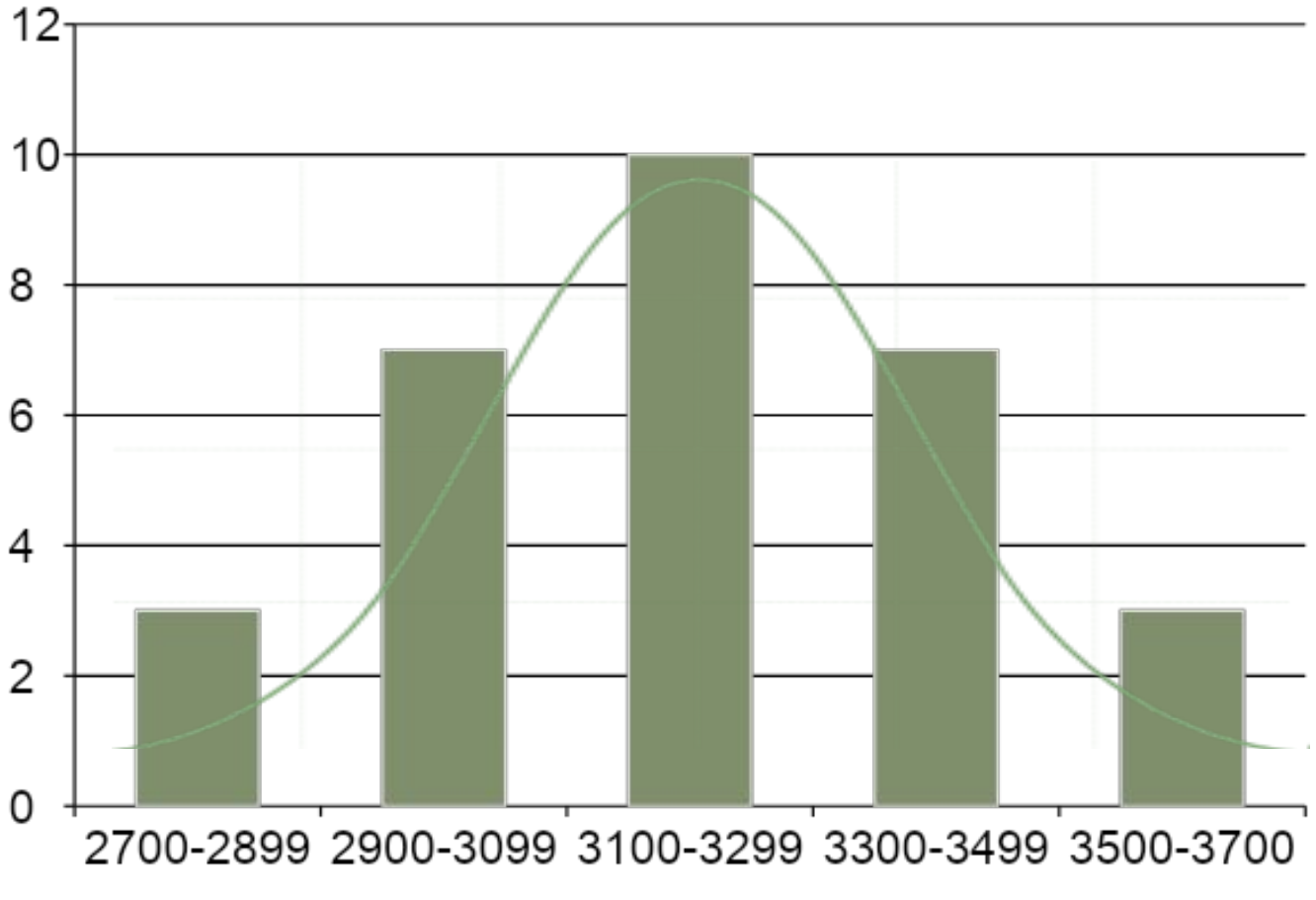
Встречается в природе наиболее часто, поэтому называется «нормальным»

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ



	1	2	3
1	1	2780	
2	2	3560	
3	3	3100	
4	4	3056	
5	5	3267	
6	6	2902	
7	7	3198	
8	8	3905	
9	9	3303	
10	10	3605	
11	11	3402	
12	12	3140	
13	13	3350	
14	14	3270	
15	15	3360	
16	16	3201	
17	17	3987	
18	18	3400	
19	19	3250	
20	20	3380	
21	21	2960	
22	22	3166	
23	23	2802	
24	24	3090	
25	25	3678	
26	26	3170	
27	27	3043	
28	28	3120	
29	29	2859	
30	30	3370	
31			

Вес	Количество детей
2700-2899	3
2900-3099	7
3100-3299	10
3300-3499	7
3500-3700	3



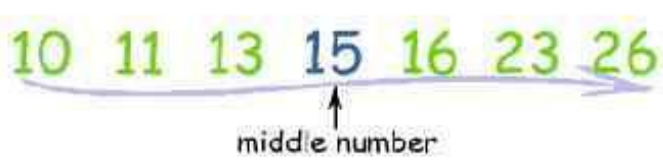
ВСЕ СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ ДЕЛЯТСЯ НА 3 БОЛЬШИЕ ГРУППЫ:

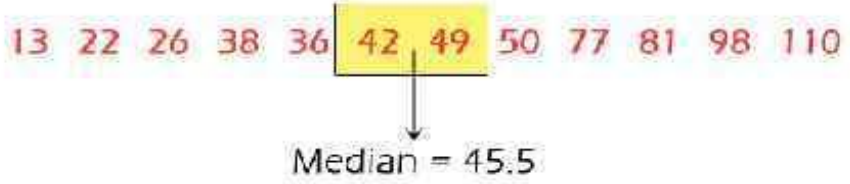
- ***Меры центральной тенденции*** - показывают расположение среднего, типичного значения признака, вокруг которого сгруппированы остальные наблюдения
- ***Меры рассеяния*** (меры изменчивости, показатели вариации) - характеризуют значения между отдельными показателями выборки. Позволяют судить о степени однородности полученного множества, и о надежности полученных результатов
- ***Меры связи*** (меры корреляции) - позволяют изучить взаимосвязь между двумя признаками/переменными

МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ (МЕРЫ ПОЛОЖЕНИЯ, МЕРЫ ЛОКАЛИЗАЦИИ)

Показывают наиболее типичное значение для данной выборки

- Среднее значение (M) - среднее арифметическое
- Медиана (Me) - средняя точка распределения
  - Если кол-во значений нечетное, то Me - среднее значение в ранжированном списке
  - Если кол-во значений четное, то Me - среднее арифметическое между двумя центральными значениями





- Мода (Mo) - наиболее часто встречающееся значение признака в выборке

1 2 2 3 3 3 3 4 5 6 7 8 9

МЕРЫ РАССЕЯНИЯ
(МЕРЫ ИЗМЕНЧИВОСТИ, ПОКАЗАТЕЛИ ВАРИАЦИИ)

Показывают разброс значений признака в выборке

- Дисперсия - характеризует, насколько частные значения отклоняются от средней величины в данной выборке *(чем больше дисперсия, тем больше "разброс данных").*
- Среднее квадратическое (стандартное) отклонение (СКО, s, SD) - позволяет оценить, насколько бОльшая часть результатов данного исследования отклоняется от среднего значения.
- Стандартная ошибка (SE-standard error) - оценка возможного отличия между значением среднего в анализируемой выборке и истинным средним, характерным для всей популяции. С увеличением выборки уменьшается данная ошибка, так как чем больше наблюдений, тем больше вероятность, что полученные данные близки к истинным.

МЕРЫ РАССЕЯНИЯ (МЕРЫ ИЗМЕНЧИВОСТИ, ПОКАЗАТЕЛИ ВАРИАЦИИ)

Показывают разброс значений признака в выборке

- Размах - разность максимального и минимального значения *(Недостаток: не характеризует распределение целиком, а только крайние значения)*
- Интерпроцентильный размах/интервал - значения каких-либо процентилей распределения, например, 10-го и 90-го
- Интерквартильный размах/интервал - значения 25-го и 75-го процентилей (такой интервал независимо от вида распределения включает 50% значений признака в выборке)

ПОНЯТИЕ О КВАНТИЛЯХ

Квантили (ед.ч. - Квантиль) - величины, разделяющие ранжированный ряд на равные части.

Разновидности квантилей:

- 1. **Медиана** - делит на 2 равные части (пополам)
- 2. **Квартили** - делит на 4 равные части
- 3. **Децили** - делит на 10 равных частей
- 4. **Перцентили** - делит на 100 равных частей

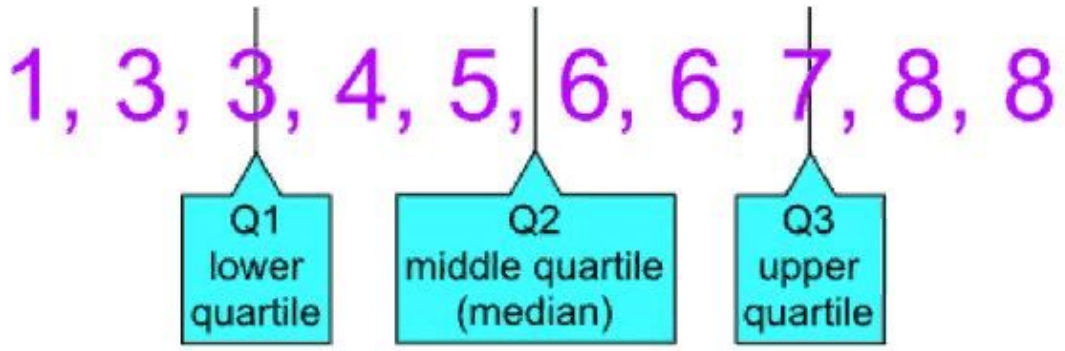
ПОДРОБНЕЕ О КВАРТИЛЯХ

Квартили делят ранжированный ряд на 4 равные части

- ***Нижний (первый) квартиль Q1*** - это медиана левой половины упорядоченного ряда. 25% значений меньше Q1

- ***Верхний (третий) квартиль Q3*** - медиана правой половины упорядоченного ряда. 25% значений больше Q3

- ***Второй квартиль Q2*** - медиана



АНАЛИЗ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Первый этап - анализ вида распределения

От вида распределения зависят:

- *Выбор способа описания центральной тенденции*
- *Выбор способа описания изменчивости значений признака*
- *Выбор методов дальнейшего анализа данных*

КАК ОПРЕДЕЛИТЬ ВИД РАСПРЕДЕЛЕНИЯ?

4 способа с помощью программы STATISTICA:

Качественные:

1. Построение гистограммы
(Graphs => Histograms=> "выбираем необходимые признаки" => OK)

2. График функции распределения в специальных координатах
(Graphs => 2D Graphs => Probability-Probability plots =>
=> Distribution – normal => "выбираем необходимые признаки" => OK)

Количественные:

3. Оценка симметричности распределения признаков

$$СКО<(М/2)$$

(Среднее квадратическое отклонение должно быть меньше половины *среднего арифметического)*

4. Проверка статистических гипотез *(используется крайне редко)*:

- Нулевая гипотеза (H0) - утверждает, что распределение исследуемого признака в генеральной совокупности соответствует закону нормального распределения
- Альтернативная гипотеза (H1) - утверждает, что распределение исследуемого признака в генеральной совокупности **не** соответствует закону нормального распределения

3 критерия:

1. Колмогорова - Смирнова (λ-критерий): применяется, если среднее значение и среднее квадратическое отклонение известны априори
2. Лиллиефорса: применяется, когда среднее значение и среднее квадратическое отклонение **не известны** априори, а вычисляются по выборке
3. Шапиро-Уилка: применяется так же, если известны среднее значение и среднее квадратическое отклонение априори, однако *данный критерий предпочтителен, так как является самым "мощным", точным и универсальным*

ОПРЕДЕЛЕНИЕ КРИТЕРИЕВ В ПРОГРАММЕ STATISTICA

Statistics => Basic Statistics/Tables =>
=>Descriptive statistics => Normality ***(здесь же, но во вкладке Advanced можно высчитать моду, медиану и среднее значение)*** => "выбираем критерии" => => Histograms

Далее оцениваем гистограмму и значение p

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

После использования программы STATISTICA будут получены результаты анализа распределения каждого признака - **p**.

- Если p < 0,05 => принимается альтернативная гипотеза -> распределение отличается от нормального -> далее будут использованы *непараметрические методы анализа данных*
- Если p ⩾ 0,05 => принимается нулевая гипотеза -> нормальное распределение -> далее будут использованы *параметрические методы анализа данных*

<u>P никак не отражает величину различий между группами, поэтому часто рассчитывают ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ (ДИ)</u>

Доверительный интервал - диапазон значений вокруг истинного значения.

ДИ с определённой вероятностью включает в себя истинные значения в генеральной совокупности.

КАКИЕ ДАННЫЕ НЕОБХОДИМО УКАЗЫВАТЬ ПРИ ОПИСАНИИ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ?

Для описания нормального распределения:

- Число наблюдений (объектов исследования)
- Среднее значение
- Среднее квадратическое отклонение (СКО)

Для описания распределения, отличающегося от нормального:

- Число наблюдений (объектов исследования)
- Медиану
- Верхний и нижний квартили

ПРИ ОПИСАНИИИ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ СЛЕДУЕТ ОБЯЗАТЕЛЬНО УКАЗЫВАТЬ ЧИСЛО НАБЛЮДЕНИЙ (ОБЪЕКТОВ ИССЛЕДОВАНИЯ) - N

Пример:

Исследуют группу из 1600 человек по 2-ум признакам: вес и анализ крови.

По каким-то причинам в ходе исследования не была получена информация о весе 10-ти объектов исследования и не были получены результаты анализа крови у 16-ти объектов.

Следовательно, мы должны указать, что:

Для признака ВЕС n=1590

Для признака АНАЛИЗ КРОВИ n=1584

В данном случае разница допустима (это нормально)

Второй этап анализа - выбор статистического метода

Статистические методы делят на:

- *Параметрические* (основываются на оценке параметров: среднее значение или стандартное отклонение; применяются для количественных признаков, если наверняка известно, что вид распределения - **нормальный**)
- *Непараметрические (*не связаны напрямую с оценкой параметров; могут применяться для количественных признаков при **любом** виде распределения + для качественных признаков)

Так как непараметрические методы можно использовать при любом виде распределения, то их используют гораздо чаще

СРАВНЕНИЕ ПАРАМЕТРИЧЕСКИХ И НЕПАРАМЕТРИЧЕСКИХ МЕТОДОВ

К ***преимуществам непараметрических методов*** можно отнести следующие:

- могут быть использованы, когда характеристики популяции, из которой делается выборка, частично неизвестны;
- бóльшая мощность;
- относительная несложность вычислений (в большинстве случаев);
- менее жесткие начальные допущения

Недостатками непараметрических методов являются:

- меньшая эффективность, чем у параметрических методов;
- меньшая специфичность;
- потенциальная трудоемкость при применении к большим массивам данных.

ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

- 1. ***Непарный t-тест (тест Стьюдента)*** - с его помощью проводят проверку нулевой гипотезы ("***Но")*** об отсутствии различий средних значений переменной в двух независимых выборках *(историческое значение)*
- 2. Если данные зависимые (повторные наблюдения за одним и тем же человеком или исследование людей по парам), то рекомендуется применять ***парный t-тест***
- 3. ***Т-тест Уэлча (t-критерий неравных дисперсий) -*** используется для проверки гипотезы о том, что две популяциии меют равные средние значения.
- 4. ***Дисперсионный анализ -*** направлен на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

- I. Для непрерывных переменных (данные, полученные на непрерывной шкале: АД, масса, рост)
  - U-тест Манна-Уитни (Mann-Whitney U) или тест Манна-Уитни-Вилкоксона (MWW)
  - Тест Крускала-Уоллиса (Kruskal-Wallis)
  - Тест знаковых рангов Вилкоксона (Wilcoxon signedrank)
- II. Для дискретных переменных (данные в виде целых чисел: кол-во людей)
  - точный тест Фишера (англ. Fisher’s exact test)
  - хиквадрат ($\chi 2$) тест (англ. chi-square test); или «хи-квадрат Пирсона» (с англ. - Pearson’s chisquare)

U-ТЕСТ МАННА-УИТНИ (MANN-WHITNEY U) ИЛИ ТЕСТ МАННА-УИТНИ-ВИЛКОКСОНА (MWW)

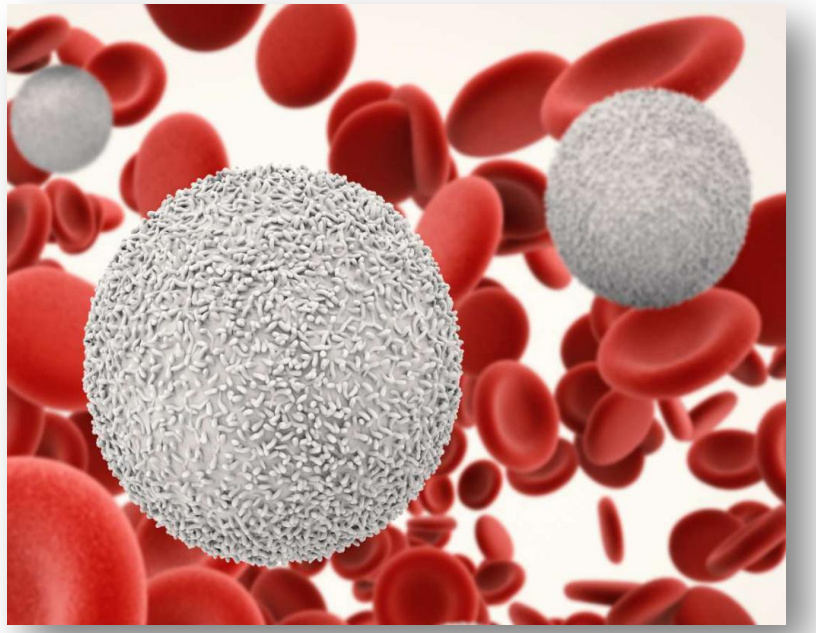
- U-критерий Манна-Уитни - используется для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно.
- Метод основан на определении того, достаточно ли мала зона перекрещивающихся значений между двумя ранжированными рядами.
- Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны.

Statistics => Nonparametrics => Comparing to independent samples => Variables (в первом окне выбираем зависимую переменную - возраст; во втором - группирующую переменную - пол) => M-W U test => оцениваем p

(P должен быть больше 0,05)

ДВЕ ПЕРЕМЕННЫЕ

?

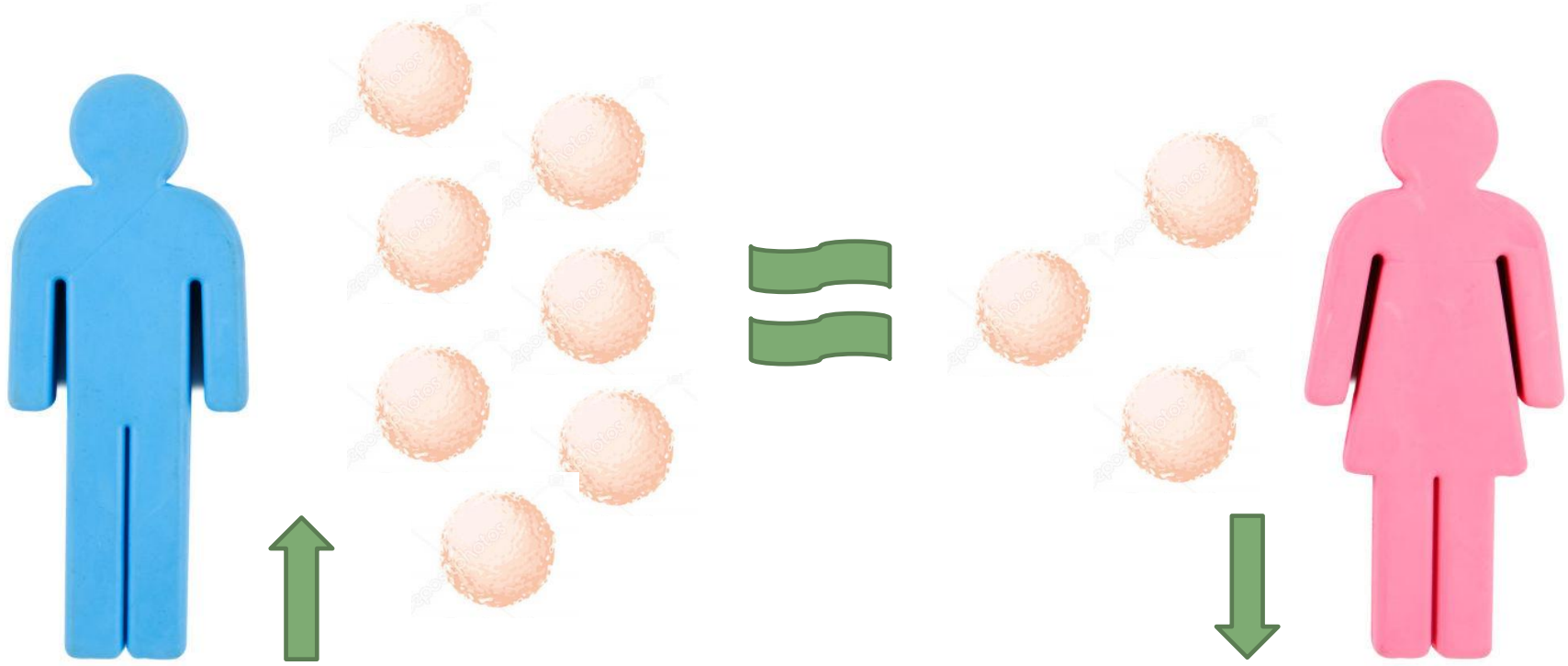


Качественная переменная

Количественная переменная

КАК УЗНАТЬ, БУДУТ ЛИ ЗАВИСИМЫ ДРУГ ОТ ДРУГА ДВЕ ПЕРЕМЕННЫЕ?

Две разные переменные зависимы в том случае, если они согласованы.



STATISTICA 64 - [Data: Spreadsheet14* (2v by 100c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help

Add to Workbook Add to Report Add to MS Word

Arial 10 B I U

	1 Номер	2 Мужчин
1	1	3,1
2	2	4,8
3	3	8,9
4	4	8,4
5	5	6,8
6	6	7,1
7	7	3,9
8	8	5,7
9	9	6
10	10	7,3
11	11	7,6
12	12	6,9
13	13	5,8
14	14	5,7
15	15	4,8
16	16	5,5
17	17	7
18	18	8,1
19	19	5,7
20	20	7,4
21	21	8
22	22	7,5
23	23	7,2
24	24	6,2

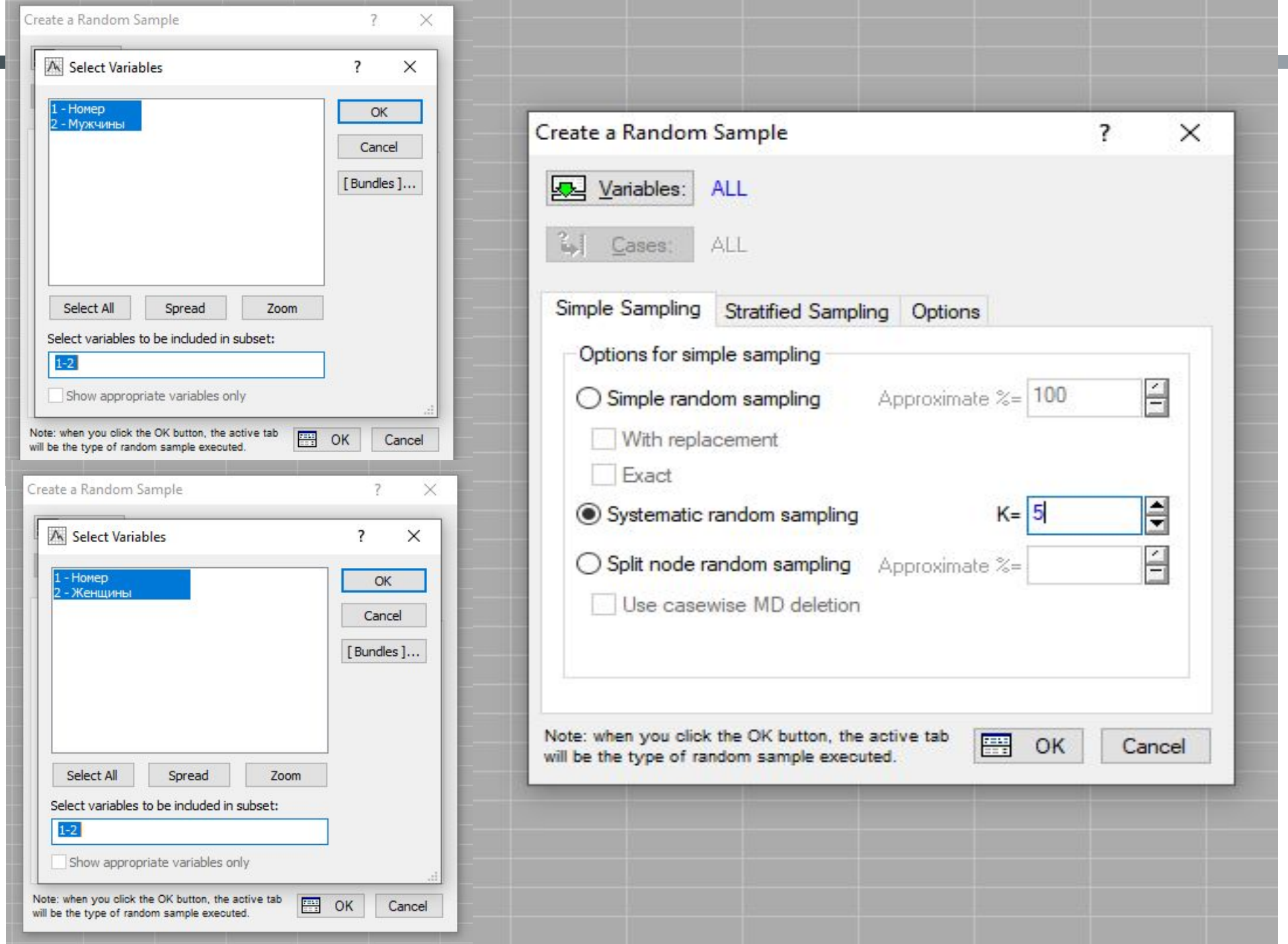
STATISTICA 64 - [Data: Spreadsheet16* (2v by 100c)]

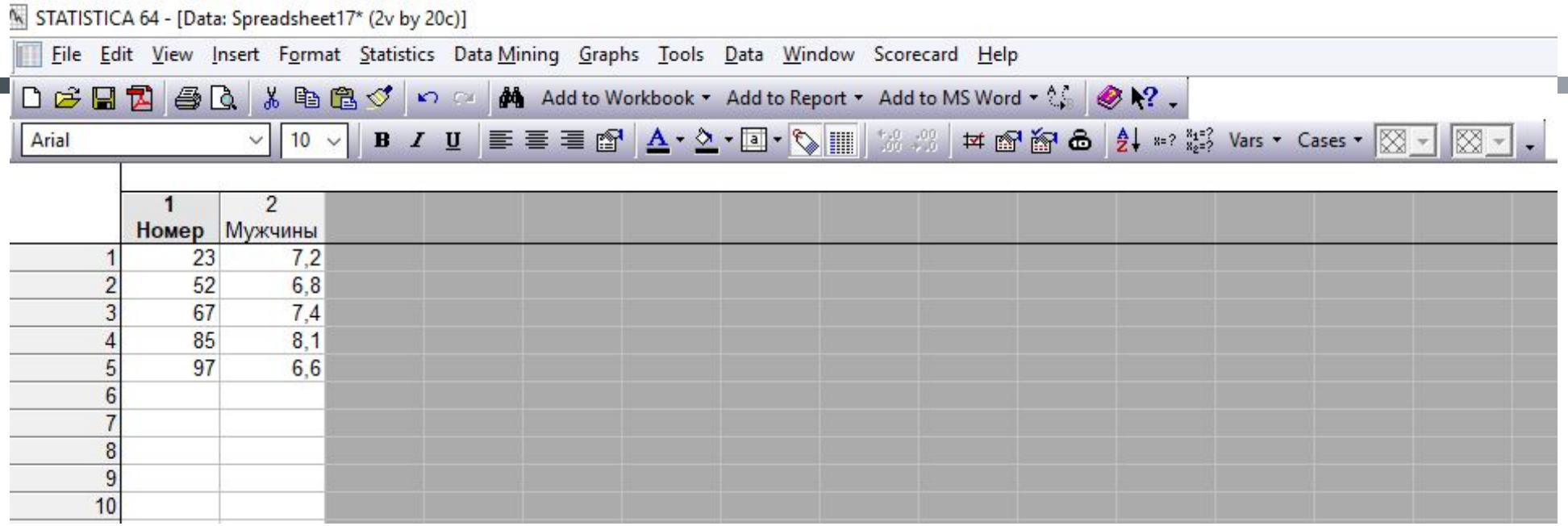
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help

Add to Workbook Add to Report Add to MS Word

Arial 10 B I U

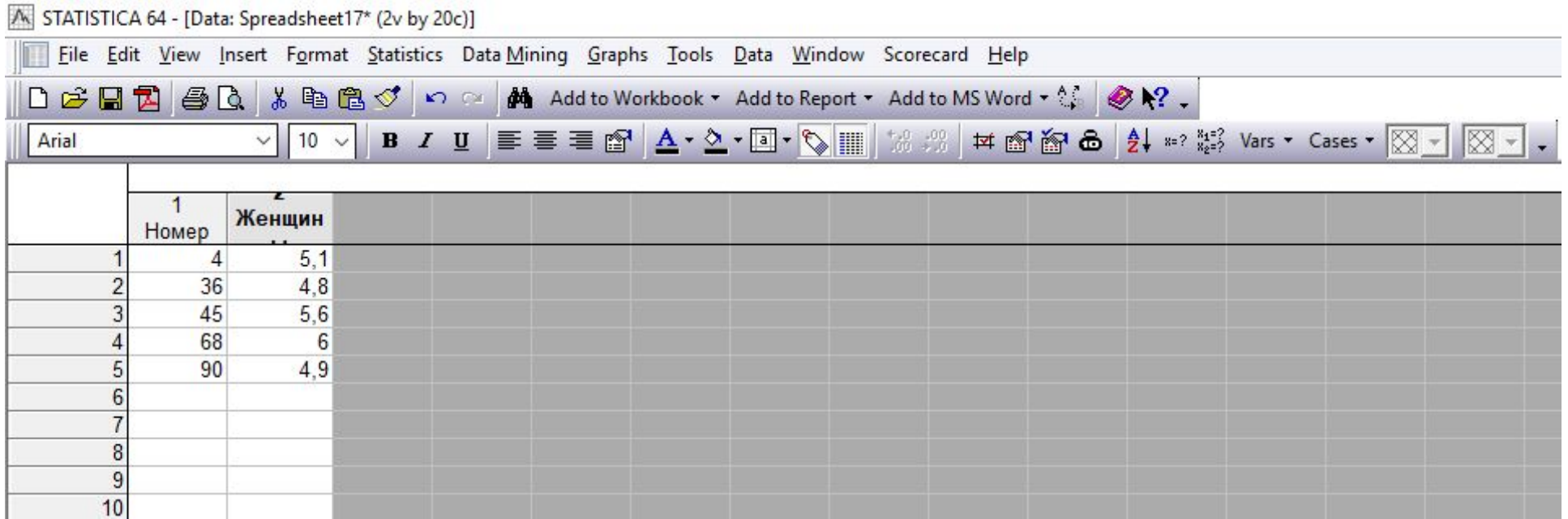
	1 Номер	2 Женщин
1	1	1,9
2	2	4,1
3	3	6,1
4	4	5,1
5	5	2,8
6	6	7,4
7	7	2,8
8	8	6,4
9	9	7,1
10	10	5,4
11	11	6,2
12	12	4,1
13	13	3,5
14	14	5,2
15	15	4,6
16	16	5,3
17	17	8
18	18	5,3
19	19	4,6
20	20	7,1
21	21	4
22	22	5,6
23	23	4,7
24	24	4,2


Create a Random Sample
Select Variables
1 - Номер
2 - Мужчины
OK
Cancel
[Bundles]...
Select All
Spread
Zoom
Select variables to be included in subset:
1-2
Show appropriate variables only
Note: when you click the OK button, the active tab will be the type of random sample executed.
OK
Cancel
Create a Random Sample
Select Variables
1 - Номер
2 - Женщины
OK
Cancel
[Bundles]...
Select All
Spread
Zoom
Select variables to be included in subset:
1-2
Show appropriate variables only
Note: when you click the OK button, the active tab will be the type of random sample executed.
OK
Cancel
Create a Random Sample
Variables: ALL
Cases: ALL
Simple Sampling
Stratified Sampling
Options
Options for simple sampling
Simple random sampling
Approximate %= 100
With replacement
Exact
Systematic random sampling
K= 5
Split node random sampling
Approximate %=
Use casewise MD deletion
Note: when you click the OK button, the active tab will be the type of random sample executed.
OK
Cancel



	1 Номер	2 Мужчины
1	23	7,2
2	52	6,8
3	67	7,4
4	85	8,1
5	97	6,6
6		
7		
8		
9		
10		

1 выборка случайных переменных



	1 Номер	Женщин
1	4	5,1
2	36	4,8
3	45	5,6
4	68	6
5	90	4,9
6		
7		
8		
9		
10		

ВЕЛИЧИНА

Может предсказать зависимость двух переменных при случайно выборке

100

Из случайной выборки у каждого мужчины лейкоцитов больше, чем у случайно выбранных женщин

23	$7{,}2*10^9$
52	$6{,}8*10^9$
67	$7{,}4*10^9$
85	$8{,}1*10^9$
97	$6{,}6*10^9$

Из случайной выборки у каждой женщины лейкоцитов меньше, чем у случайно выбранных мужчин

4	$5{,}1*10^9$
36	$4{,}8*10^9$
45	$5{,}6*10^9$
68	$6{,}0*10^9$
90	$4{,}9*10^9$

100

STATISTICA 64 - [Data: Spreadsheet17* (2v by 20c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help

Add to Workbook Add to Report Add to MS Word

Arial 10 B I U Vars Cases

	1 Номер	2 Мужчины
1	24	6,2
2	55	5,8
3	60	6,1
4	78	7,2
5	99	8
6		
7		
8		
9		
10		

2 выборка случайных переменных

STATISTICA 64 - [Data: Spreadsheet17* (2v by 20c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help

Add to Workbook Add to Report Add to MS Word

Arial 10 B I U Vars Cases

	1 Номер	2 Женщины
1	2	4,1
2	34	5,7
3	54	8,1
4	59	3,2
5	91	4,8
6		
7		
8		
9		
10		

НАДЕЖНОСТЬ (ИСТИННОСТЬ)

Показывает, распространяется ли данная зависимость на все случайные выборки

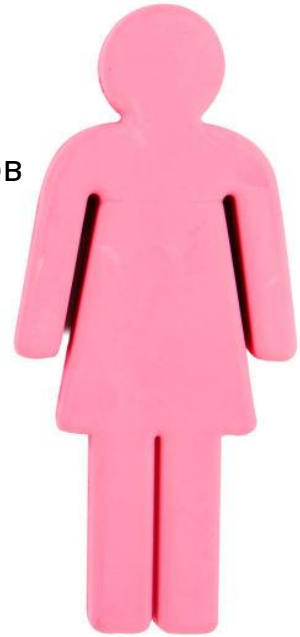
100

Из случайной выборки у мужчин лейкоцитов больше, чем у случайно выбранных женщин, кроме одной

24	$6{,}2*10^9$
55	$5{,}8*10^9$
60	$6{,}1*10^9$
78	$7{,}2*10^9$
99	$8{,}0*10^9$

Из случайной выборки у одной женщины лейкоцитов больше, чем у случайно выбранных мужчин

2	$4{,}1*10^9$
34	$5{,}7*10^9$
54	$8{,}1*10^9$
59	$3{,}2*10^9$
91	$4{,}8*10^9$



100

ЧТО ТАКОЕ P-УРОВЕНЬ (ЗНАЧИМОСТЬ)

Значимость – оценённая мера уверенности в его «истинности». P-уровень находится в обратной зависимости от надежности результата. Более высокий p-уровень соответствует более низкому уровню доверия к найденной в выборке зависимости между переменными.

P-уровень  **Надежность**

ЗНАЧИМОСТЬ

Данная зависимость встретилась лишь 5 раз из 100 выборок.
P-уровень = 0,05. Связь является значимой лишь в этих 5 случайных выборках.

23	$7,2*10^9$	4	$5,1*10^9$
52	$6,8*10^9$	36	$4,8*10^9$
67	$7,4*10^9$	45	$5,6*10^9$
85	$8,1*10^9$	68	$6,0*10^9$
97	$6,6*10^9$	90	$4,9*10^9$

				1	3,1*109	3	6,1*109
				32	$4,8*10^9$	23	$4,7*10^9$
				42	$7,1*10^9$	43	$6,1*10^9$
				32	$6,2*10^9$	56	$2,2*10^9$
				34	$8,0*10^9$	76	$3,8*10^9$
100	24	6,2*109	2	4,1*109			
	55	$5,8*10^9$	34	$5,7*10^9$			
	60	$6,1*10^9$	54	$8,1*10^9$			
	78	$7,2*10^9$	59	$3,2*10^9$			
	99	$8,0*10^9$	91	$4,8*10^9$			

100

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ - МЕРА УВЕРЕННОСТИ В "ИСТИННОСТИ" РЕЗУЛЬТАТА

- Статистическая значимость определяется значением p-уровня *(p-value)*
- Чем выше p-уровень, тем ниже уровень доверия к полученным результатам *(обратная зависимость)*

↑ p-уровень ⇒ ↓ уровень доверия

- P > 0,05 результатам нельзя доверять
- p ⩽ 0,05 статистически значимые результаты
- P < 0,01 статистически высокозначимые результаты

<u>Пример:</u> p-уровень - 5% (0,05) показывает, что сделанный при анализе вывод является случайной особенностью с вероятностью 5%. Другими словами, с вероятностью 95% вывод можно распространить на все объекты.