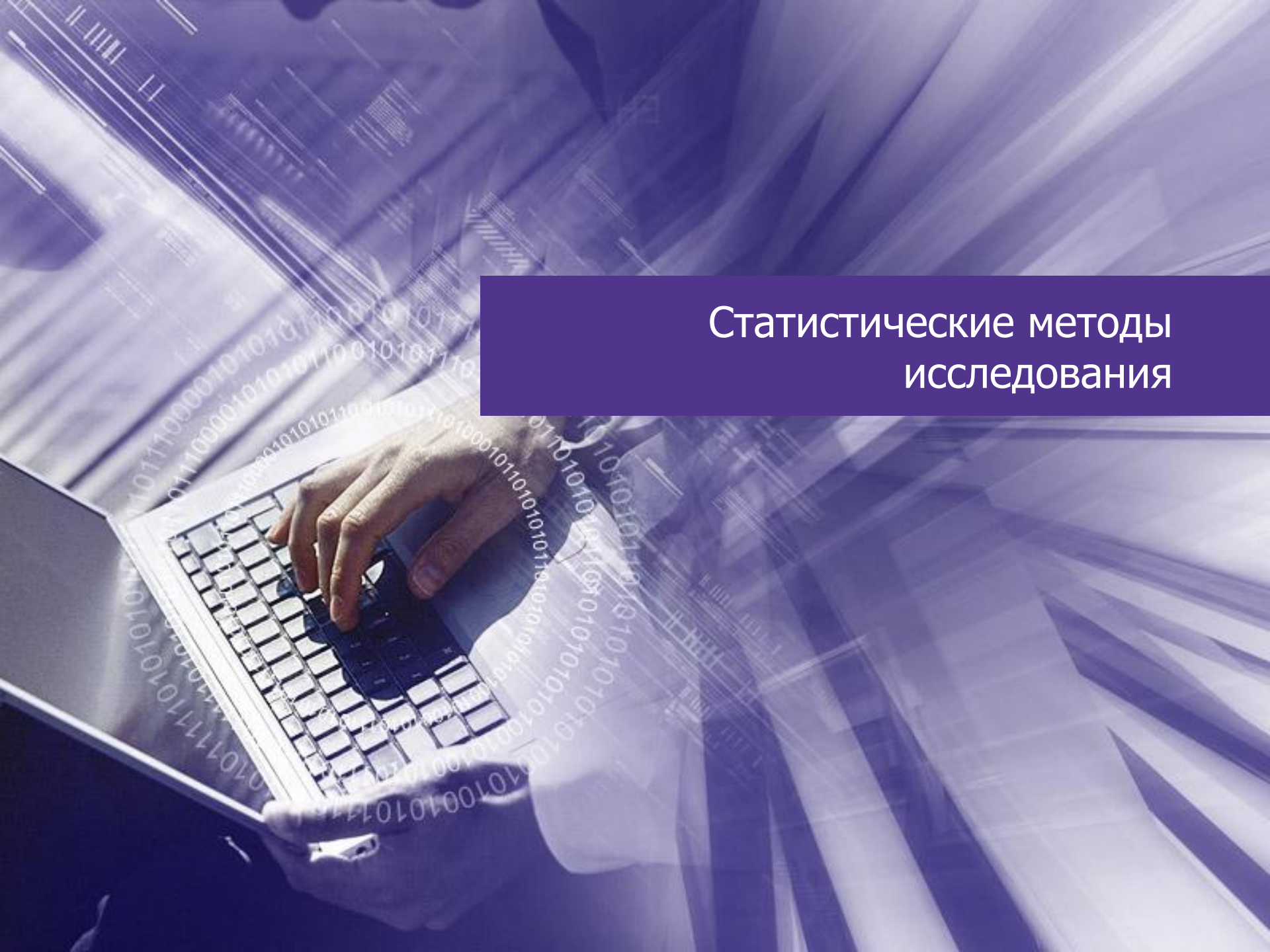


Статистические методы исследования

Я подготовил тезисы своего доклада, а вы подберите немного статистики, чтобы их обосновать.

Математическая статистика

Математическая статистика - область науки, изучающая случайные явления, разрабатывающая математические методы систематизации, обработки и использования статистических данных для научных и практических выводов.

Составными частями математической статистики являются:

(1) описание данных,

(2) статистическое оценивание

(3) проверка статистических гипотез.

Замечания

- Статистические методы *основаны на логике.*
- Следует опасаться применения статистических методов без их глубокого понимания и без контекста, который может оказаться крайне важным.
- Только после постижения внутренней логики каждого из методов можно с уверенностью говорить о способности исследователя без труда применять статистику для изучения явлений.

Статистические данные

- Числовые → Числовая статистика
- Числовые статистические данные – это числа, вектора, функции. Их можно складывать, умножать на коэффициенты. Поэтому в числовой статистике большое значение имеют разнообразные суммы.
- Математический аппарат анализа сумм случайных элементов выборки – это (классические) законы больших чисел и центральные предельные теоремы

Статистические данные

- Нечисловые → Нечисловая статистика
- Нечисловые статистические данные – это категоризованные данные, вектора разнотипных признаков, бинарные отношения, множества, нечеткие множества и др. Их нельзя складывать и умножать на коэффициенты. Поэтому не имеет смысла говорить о суммах нечисловых статистических данных. Они являются элементами нечисловых математических пространств (множеств).
- Математический аппарат анализа нечисловых статистических данных основан на использовании расстояний между элементами (а также мер близости, показателей различия) в таких пространствах.

Переменные

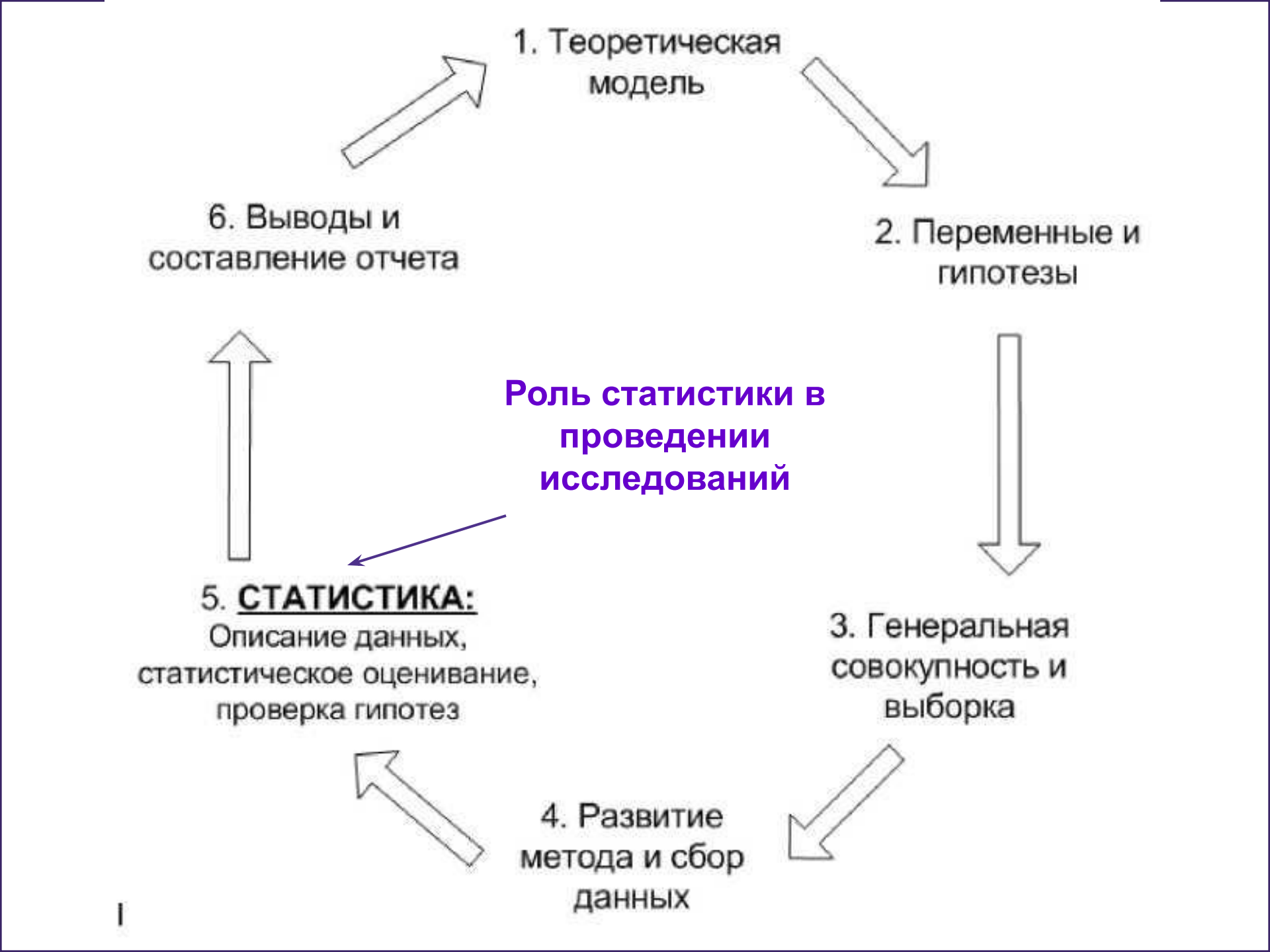
- **Данные (data)** представляют собой результаты наблюдений, испытаний, накапливаемые с целью последующего изучения и анализа.
- **Переменная, признак (variable)** - это некоторая общая для всех изучаемых объектов характеристика или свойство, конкретные проявления которого могут меняться от объекта к объекту.
- Проявления признака называют **значениями, показателями, альтернативами, градациями.**
- **Распределение переменной (distribution of the variable)** - совокупность различных значений, которые переменная принимает для различных изучаемых объектов.

Определения

- **Генеральная совокупность (population)** - вся интересующая исследователя совокупность изучаемых объектов.
- **Выборка, выборочная совокупность (sample)** - некоторая, обычно небольшая, часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о свойствах генеральной совокупности.

- **Параметры (parameters)** - числовые характеристики генеральной совокупности.
- **Статистики (statistics)** - числовые характеристики выборки.

- **Гипотеза (hypothesis)** - предположение относительно параметров генеральной совокупности, которое подлежит проверке на основе анализа выборки.


1. Теоретическая модель
2. Переменные и гипотезы
3. Генеральная совокупность и выборка
4. Развитие метода и сбор данных
5. СТАТИСТИКА:
Описание данных, статистическое оценивание, проверка гипотез
6. Выводы и составление отчета
Роль статистики в проведении исследований

Измерение явлений

- **Измерение (measurement)** означает присвоение чисел характеристикам изучаемых объектов, явлений согласно некоторому правилу.
- **Шкала (scale)** есть правило или алгоритм, в соответствии с которым изучаемым объектам, явлениям присваиваются числа.

Типы данных

- **Дискретные данные (discrete data)** представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счетным, т.е. может быть подсчитано натуральными числами от одного до бесконечности.
- **Непрерывные данные (continuous data)** могут принимать любое значение в некотором интервале.

Критерии измерений

- **Надежность измерения (reliability)** означает возможность получить согласующиеся результаты при повторных
- **Достоверность измерения (validity)** означает соответствие между результатами измерения и его целями, между выбранной шкалой и исследуемыми переменными. измерениях характеристик объекта.
- **Завершенность измерения (exhaustive)** означает, что в результате измерения мы должны получить какой-либо результат.
- **Единственность измерения (mutually exclusive)** означает, что в результате измерения мы получим только одно значение переменной.

Измерительные шкалы (С. Стивенс)

- номинативная, или номинальная, или шкала наименований (в том числе дихотомическая)
- порядковая, или ранговая, или ординальная шкала
- интервальная, или шкала равных интервалов
- шкала равных отношений или реляционная шкала

- Стивенсовская типология измерительных шкал получила повсеместное распространение, однако, по мнению Суходольского Г.В. к числу измерительных шкал относятся только интервальные и реляционные шкалы.

- Применение статистического метода определяется, прежде всего, шкалой в которой измерена переменная.

Шкала	Особенности	Пример
Номинальная	Содержит только категории, данные не могут упорядочиваться	Хобби студента. Только название.
Дихотомическая	Содержит две категории	Пол студента. Третьего не дано, если не рассматривать исключения.
Порядковая	Категории могут упорядочиваться, но разности не имеют смысла	Место на соревнованиях. Лучше результат - выше место.
Интервальная	Разности между значениями могут быть вычислены, но нет отношений	Температура студента. У больного выше на 1-2°C
Относительная	Имеется точка отсчета, возможны отношения между значениями	Рост студента. Один в 1,2 раза выше другого

Представление данных

- Группировка
- Табулирование
- Ранжирование
- Распределение частот
- Интервальное распределения частот
- Статистические ряды
- Графическое представление данных

Типическое значение	Номинальные данные	Порядковые данные	Интервальные данные
Мода	●	●	●
Медиана		●	●
Среднее			●

Меры центральной тенденции

Mo

Md

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

$$gm = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots x_n}$$

$$hm = \frac{1}{[(1/x_i) + (1/x_i) + \ldots (1/x_i)]/n} = \frac{n}{\sum_{i=2}^{n} 1/x_i}$$

- Мода
- Медиана
- Среднее арифметическое значение
- Среднее геометрическое
- Среднее гармоническое

Меры изменчивости (вариативности)

$$R = x_{\max} - x_{\min}$$

$$IQR = Q_3 - Q_1$$

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

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

$$CV = \frac{\sigma}{\bar{x}} \cdot 100\%$$

$$A = \frac{\sum(x_i - \bar{x})^3}{n \cdot \sigma^3}$$

$$E = \frac{\sum(x_i - \bar{x})^4}{n \cdot \sigma^4} - 3$$

- Размах
- Квартильный размах
- Дисперсия
- Стандартное отклонение
- Коэффициент вариации
- Асимметрия
- Эксцесс

Наименование характеристики	Для генеральной совокупности	Для выборки
Количество элементов	N	n
Частота	M	m
Частость (доля)	$p = \frac{M}{N}$	$w = \frac{m}{n}$
Среднее	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Дисперсия	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Стандартное отклонение	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Множество X

Множество Z

Множество Y

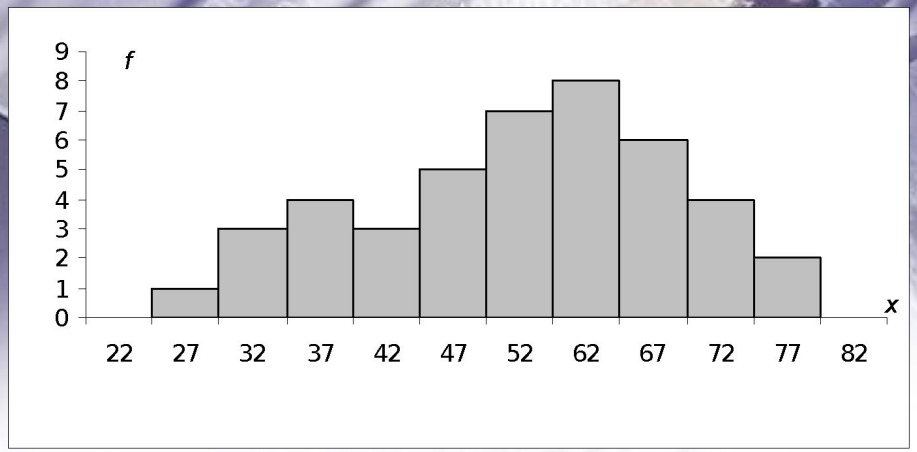
$$z_i = \frac{x_i - \bar{x}}{s_x}$$

$$y_i = cz_i + d$$

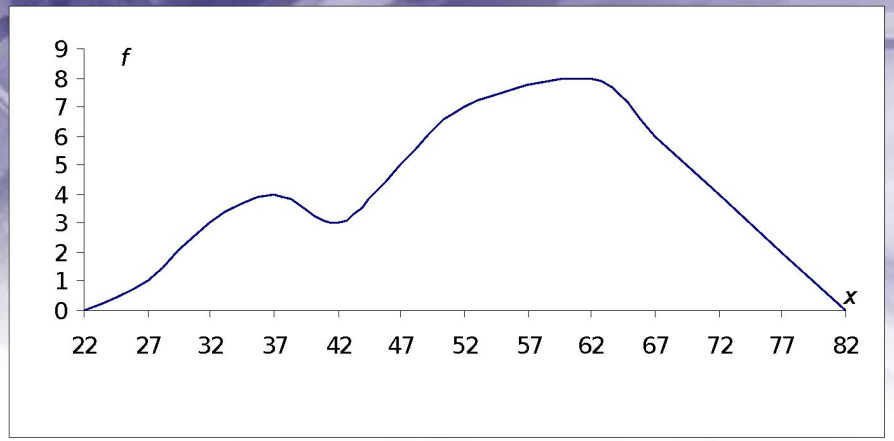
Стандартизация шкал

- Любое множество *n* данных со средним $\bar{x}$ и стандартным отклонением *Sx* можно преобразовать в другое множество со средним 0 и стандартным отклонением 1 таким образом, что преобразованные значения будут непосредственно выражаться в отклонениях исходных значений от среднего, измеренных в единицах стандартного отклонения. Новые значения называют значениями *z.*

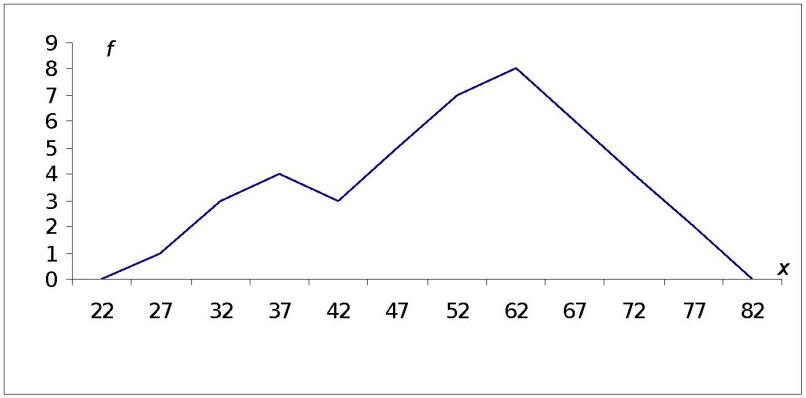
- Множество данных можно расположить на любой шкале, то есть им можно приписать желаемые среднее (d) и стандартное отклонение (c), пользуясь выражением



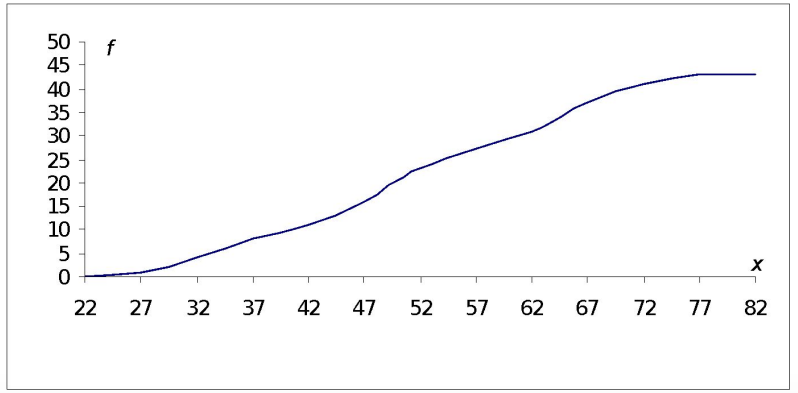
Гистограмма результатов тестирования 43 абитуриентов



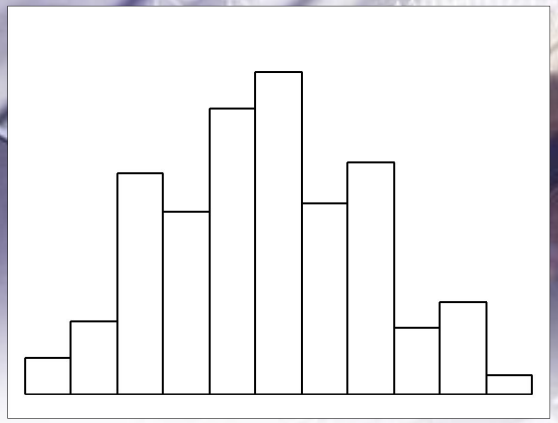
Кривая распределения

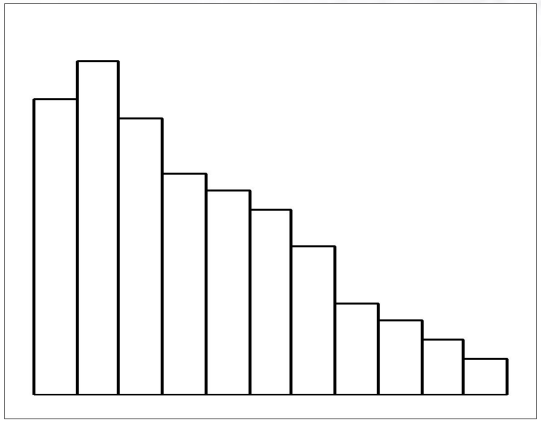


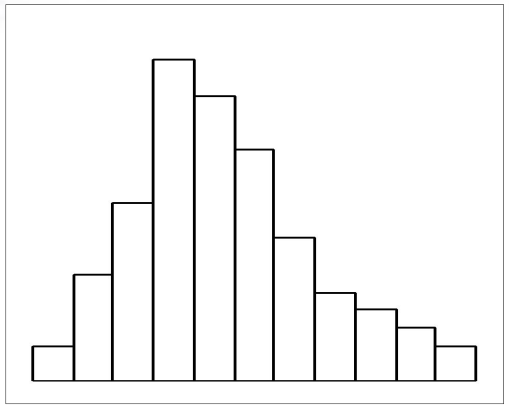
Полигон распределения

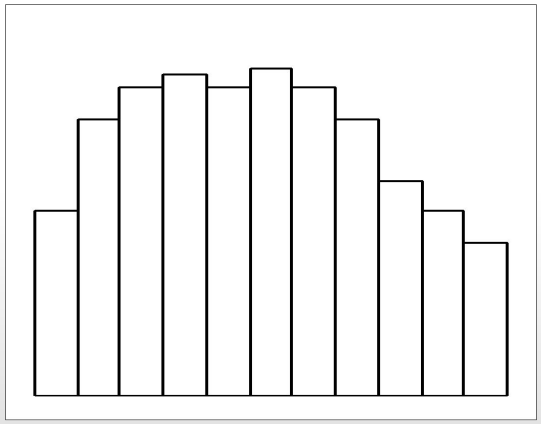


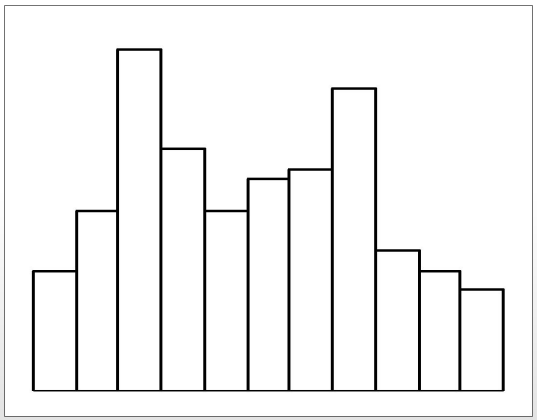
Кумулятивный полигон.

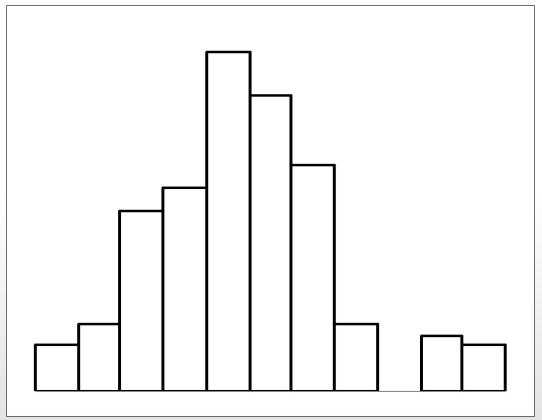












Исследовательский анализ данных

- **Исследовательский анализ данных (Exploratory Data Analysis - EDA)** представляет собой применение статистических методов для представления, упорядочения данных и понимания их важнейших характеристик.
- Это комплексный анализ характеристик распределения

1. **Измерение центральной тенденции**
2. **Измерение вариации.**
3. **Нахождение и анализ выбросов.** Выделение границ для выбросов, анализ экстремальных и умеренных выбросов.
4. **Анализ формы распределения.** Вычисление и анализ коэффициентов асимметрии и куртозиса.

Вероятность (классическое определение)

$$P(A) = \frac{m}{n}$$

- **Вероятностью события А** назовем отношение числа благоприятных исходов к общему числу элементарных исходов (классическое определение вероятности).

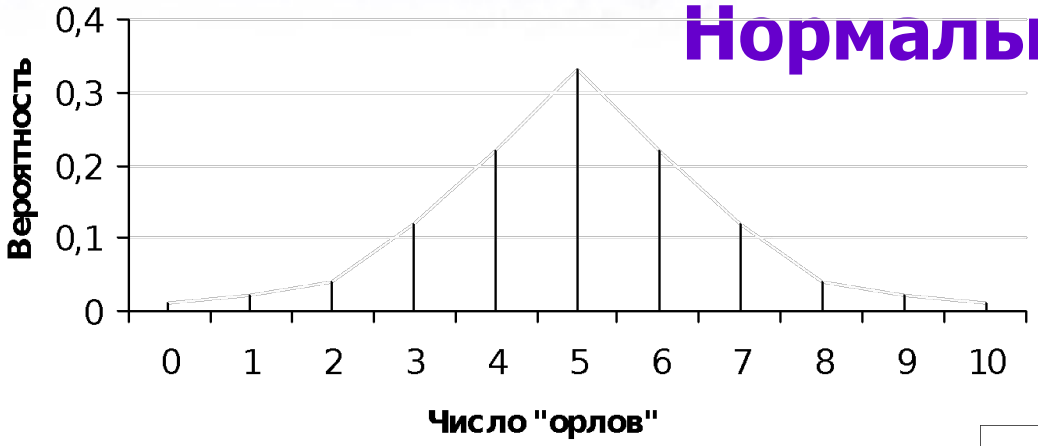
- Вероятность достоверного события равна единице
- Вероятность невозможного события равна нулю.
- Вероятность любого события не может быть меньше нуля и больше единицы: $0 < p(A) < 1$.

Вероятность (статистической определение)

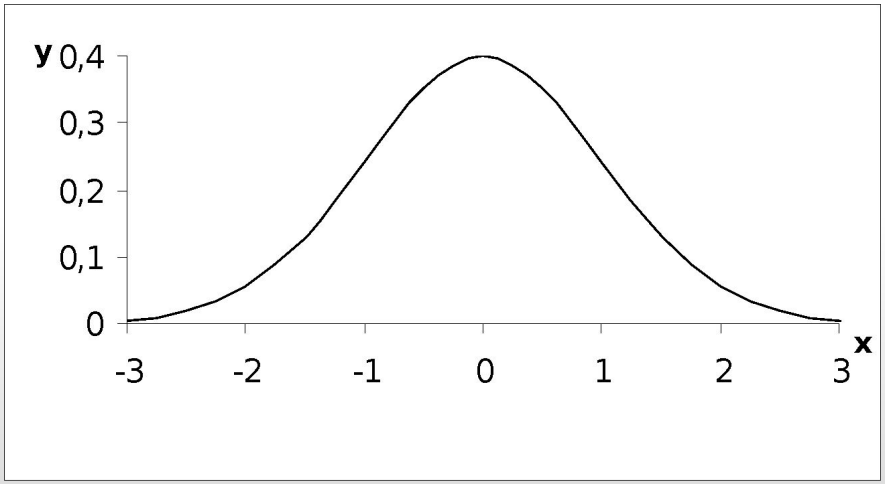
$$P(A) = \lim_{n \to \infty} \frac{s}{n},$$

- **Вероятность события А** - предельная относительная частота появления события А при проведении серии испытаний, при неограниченном увеличении их числа.

Нормальное распределение



$$u = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-(X-\mu)^2/2\sigma^2}$$

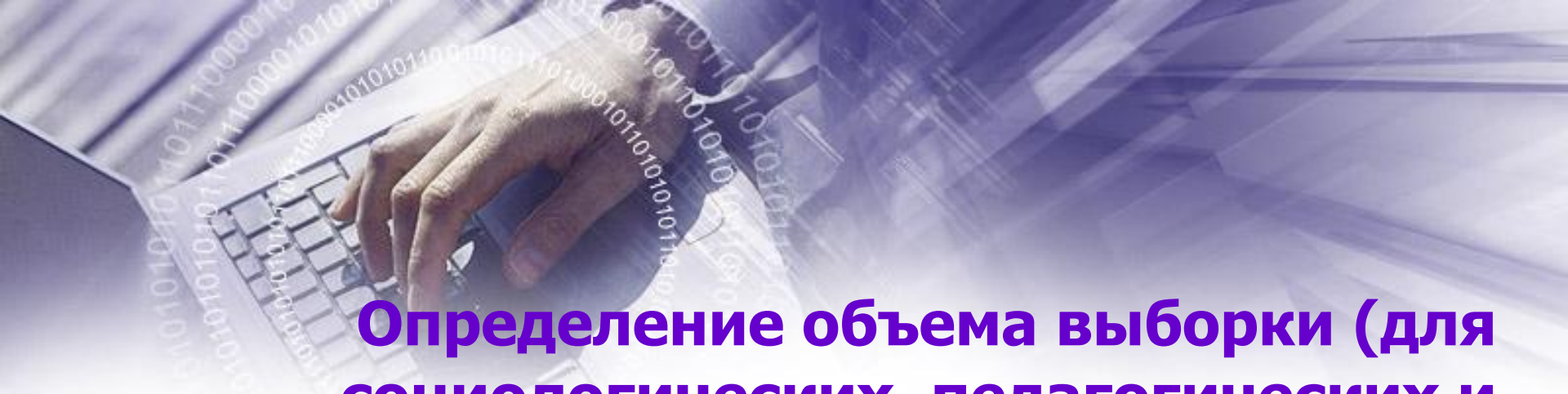


Выборки

- Зависимые (связные)
- Независимые (несвязные)

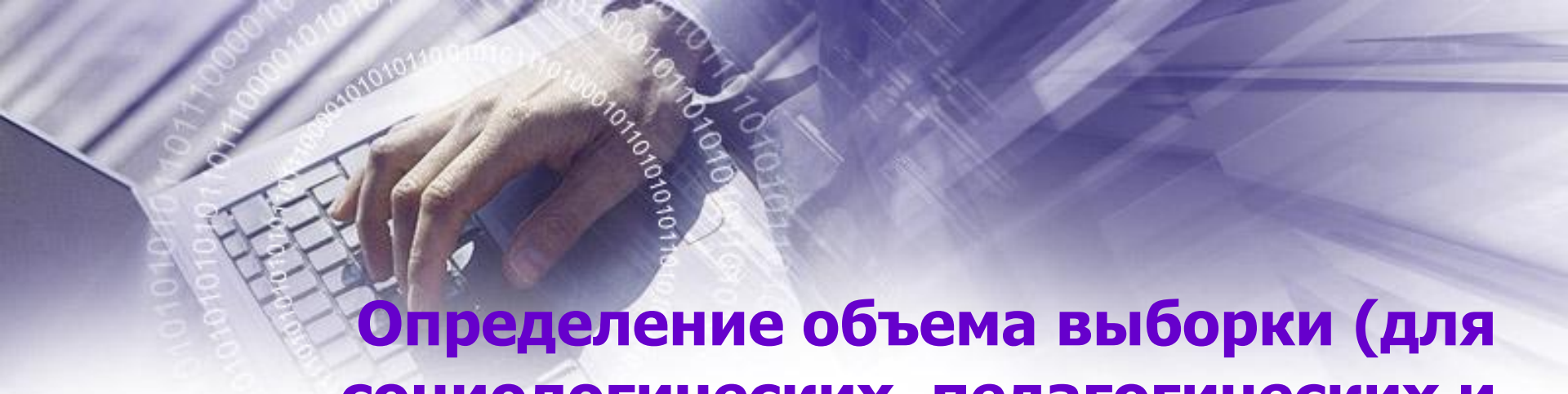
Требования к формированию выборок:

- Однородность
- Репрезентативность
- Повторность или безповторность



Определение объема выборки (для социологических, педагогических и психологических исследований)

- принято считать, что при $n \geq 60$ выборка большая или репрезентативная, но такое деление тоже весьма условно;
- наибольший объем выборки необходим при разработке диагностической методики – от 200 до 1000-2500 человек;
- если необходимо сравнивать две выборки, их общая численность должна быть не менее 50 человек; численность сравниваемых выборок должна быть приблизительно одинаковой;



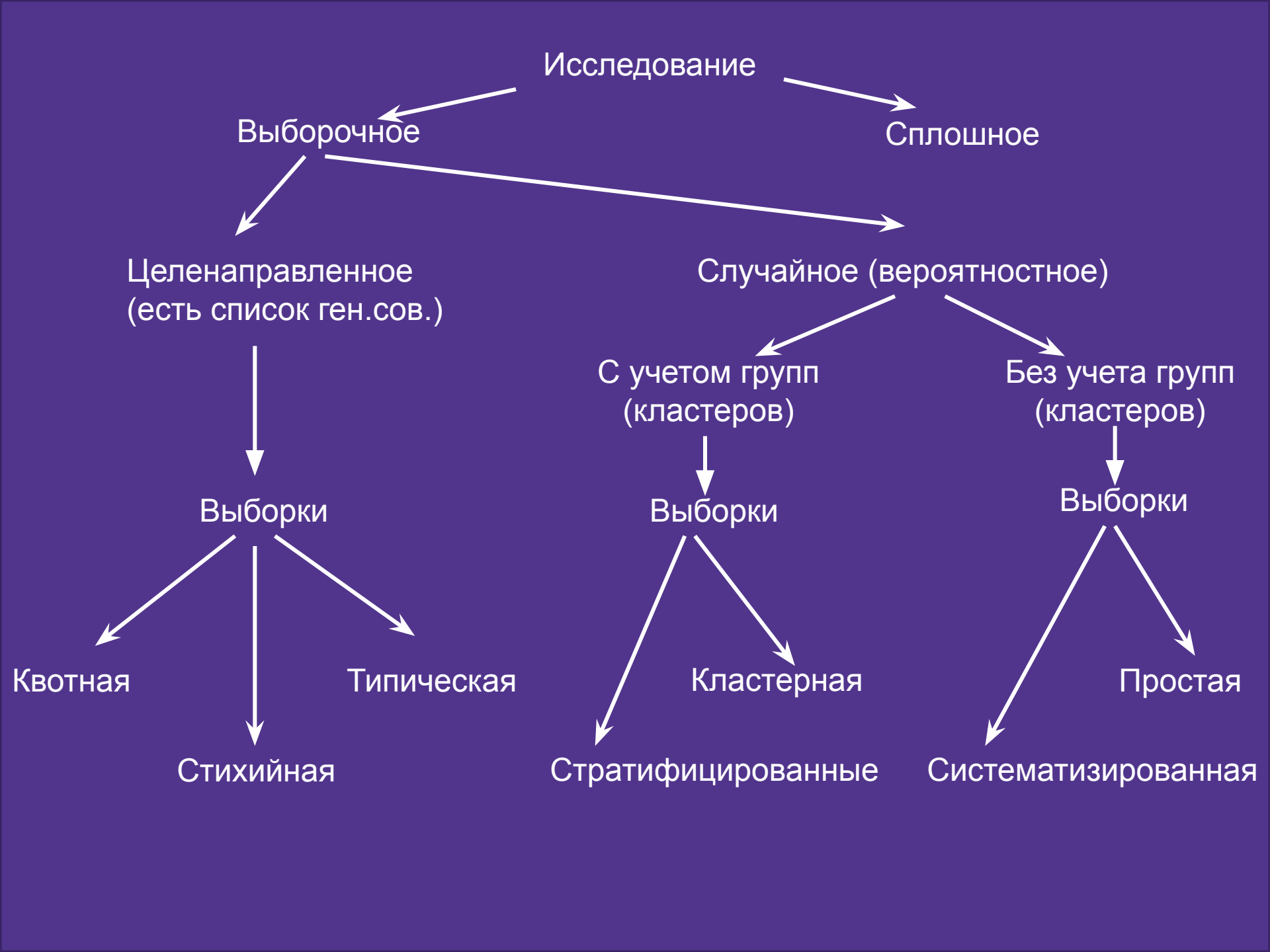
Определение объема выборки (для социологических, педагогических и психологических исследований)

- если изучается взаимосвязь между какими-либо свойствами, то объем выборки должен быть не меньше 30-35 человек;
- чем больше изменчивость изучаемого свойства, тем больше должен быть объем выборки. Поэтому изменчивость можно уменьшать, увеличивая однородность выборки, например по полу, возрасту и т.д.. при этом, естественно, уменьшаются возможности генерализации выводов.

Статистический метод определения объема безповторной выборки

$$n = \frac{t^2 \cdot \sigma^2 \cdot N}{N \cdot \alpha^2 + t^2 \cdot \sigma^2},$$

- где n – объем выборки,
- σ – стандартное отклонение,
- N – объем генеральной совокупности,
- α – предельная ошибка репрезентативности, задается обычно в пределах от 0,01 до 0,10 с наиболее частым употреблением 0,05 (5%);
- t – табулированная константа, табличные значения этой величины следующие: t=1,96, при α=0,05; t=2,58, при α=0,01.


Исследование
Выборочное
Сплошное
Целенаправленное
(есть список ген.сов.)
Случайное (вероятностное)
Выборки
С учетом групп
(кластеров)
Без учета групп
(кластеров)
Выборки
Выборки
Квотная
Стихийная
Типическая
Кластерная
Стратифицированные
Простая
Систематизированная

Алгоритм решения

- Определить, какая модель кажется наиболее подходящей для доказательства научных предположений
- Ознакомиться с описанием метода, примерами и задачами
- Рассмотреть ограничения критерия и возможность сбора необходимых данных.
- Определить объем выборки
- Обеспечить доступ к выборке
- Провести исследование, обработать полученные данные по заранее выбранному алгоритму
- Если ограничения выполнить не удалось, обратиться к предыдущим шагам, когда данные уже получены.

Схема применения статистических методов

- Формулируются ***статистические гипотезы:***
  Но: гипотеза об отсутствии различий (так называемая ***нулевая гипотеза*)**
  H1: гипотеза о значимости различий (так называемая ***альтернативная гипотеза***)
- Для принятия решений о том, какую из гипотез следует принять, используют решающие правила – ***статистические критерии***
- То есть, на основании информации о результатах наблюдений вычисляется число, называемое ***эмпирическим значением*** критерия.
- Это число сравнивается с известным (например, заданным таблично) эталонным числом, называемым ***критическим значением*** критерия.

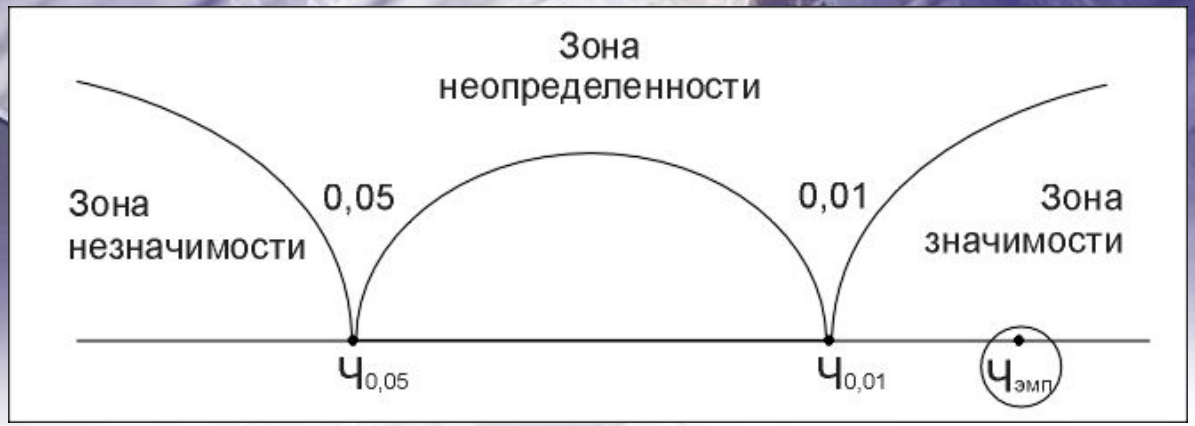
Критические значения

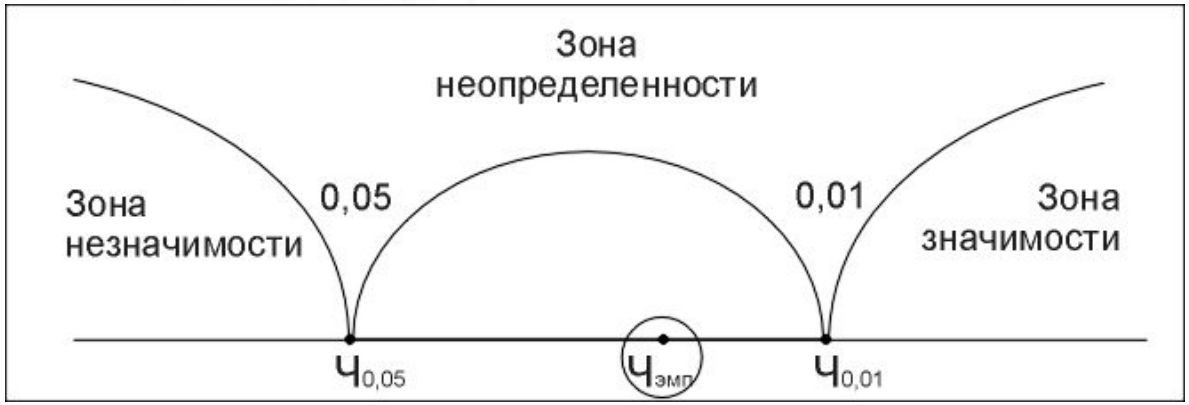
- Находится по специальным таблицам – для каждого метода свои таблица
- Зависят или от объема выборки, или от количества интервалов, или количества выборок
- Зависят от уровня значимости
  - *Уровни значимости - вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, когда она верна, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны.*
  - *Обычно различают (p) 0,05, 0,01 и 0,001.*

Правила принятия гипотез

- Если полученное исследователем эмпирическое значение критерия оказывается меньше или равно критическому, то принимается ***нулевая гипотеза***.
- В противном случае, если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается ***альтернативная гипотеза***.
- В разных науках принято считать низшим разный уровень статистической значимости, например
  - **в психологии – это 0,05**
  - **в экономике, физике – это 0,01**

Графическая интерпретация


Зона
неопределенности
Зона
незначимости
0,05
0,01
Зона
значимости
Ч0,05
Ч0,01
Чэмп


Зона
неопределенности
Зона
незначимости
0,05
0,01
Зона
значимости
Ч0,05
Чэмп
Ч0,01


Зона
неопределенности
Зона
незначимости
0,05
0,01
Зона
значимости
Чэмп
Ч0,05
Ч0,01

Интерпретация ответов

- r_{xy} = 0,669. Гипотеза *H0* отвергается и принимается гипотеза *H1* (при α≤0,01).

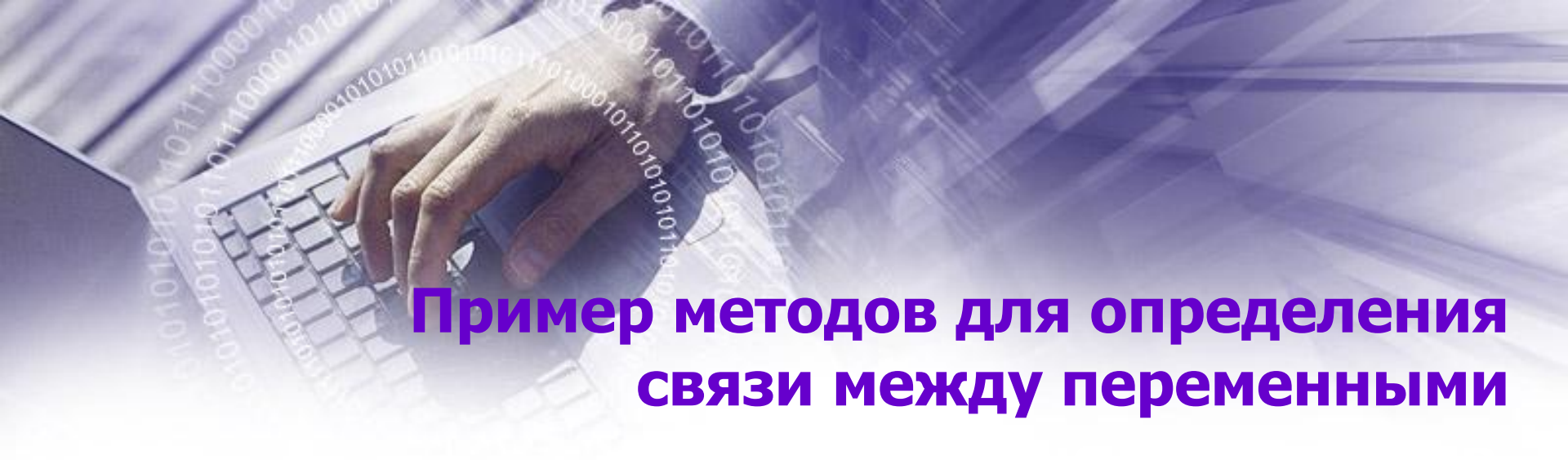
≈ *Достоверность составляет 99%.*

- r_{xy} = 0,669. Гипотеза *H0* отвергается и принимается гипотеза *H1* (при α≤0,05).

≈ *Достоверность составляет 95%.*

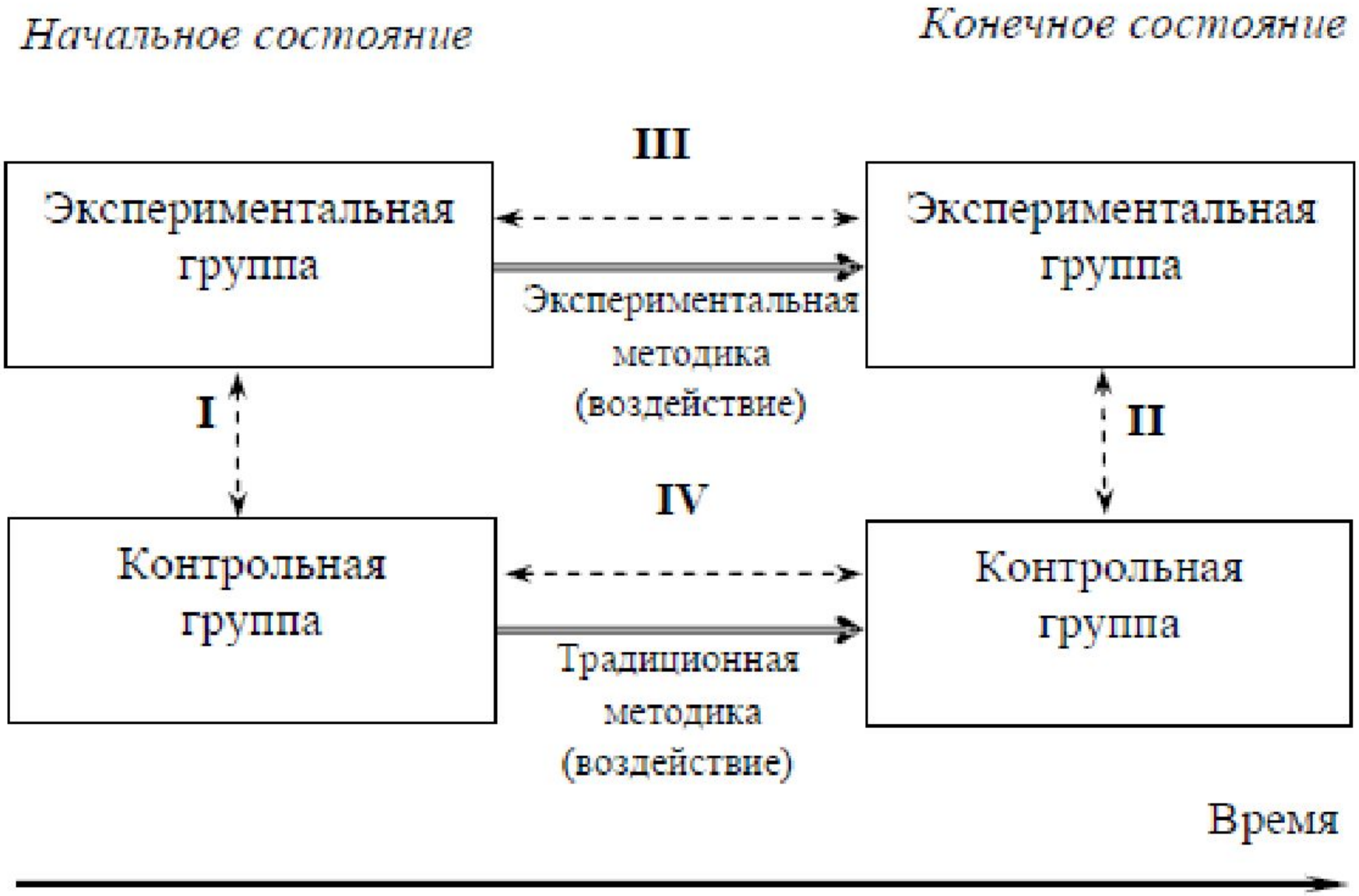
Классификация задач

1. Выявление различий в уровне исследуемого признака
2. Оценка сдвига значений исследуемого признака
3. Выявление различий в распределении признака.
4. Выявление степени согласованности изменений
5. Анализ изменений признака под влиянием контролируемых условий
6. Методы многомерного анализа

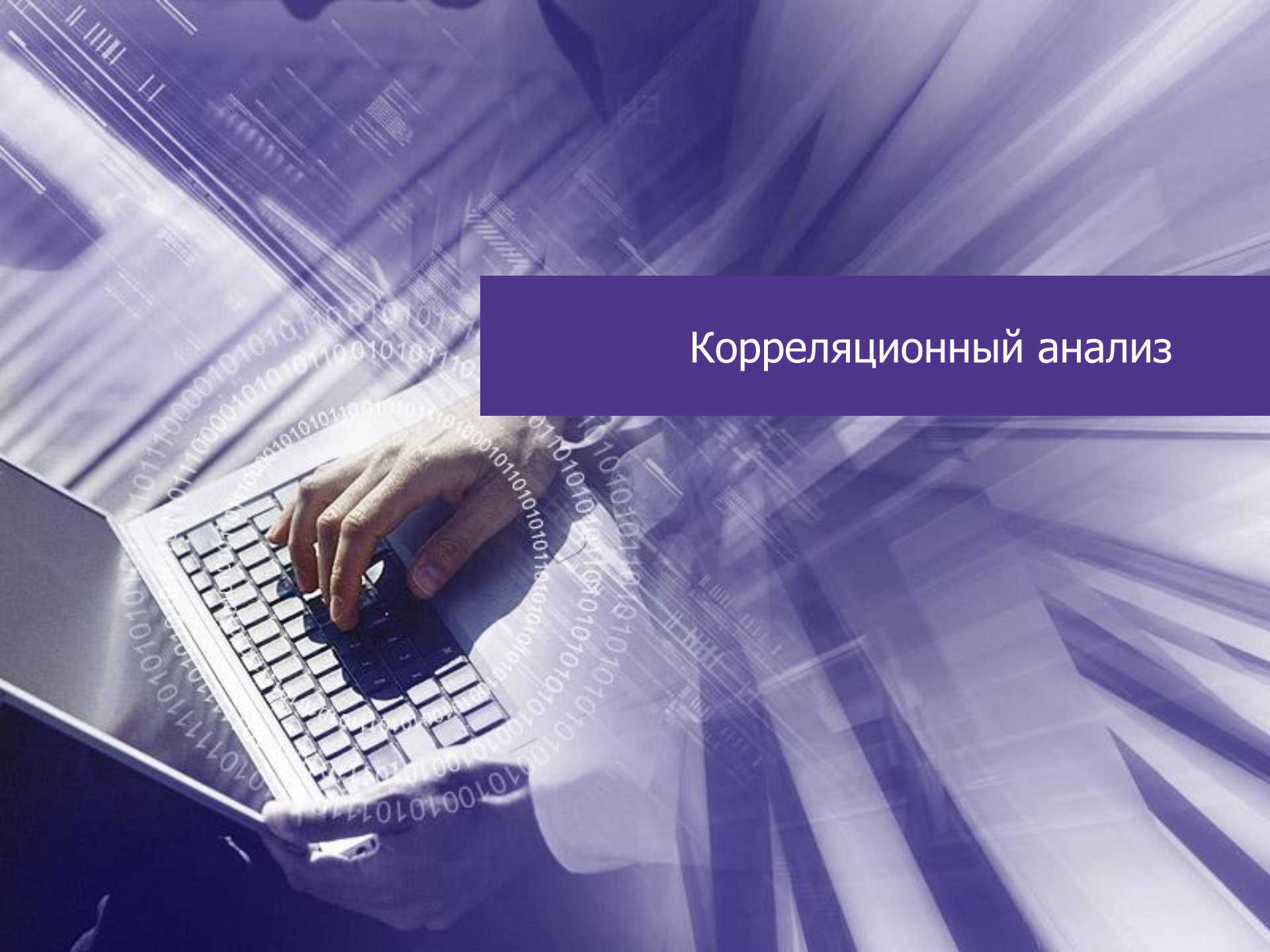


Пример методов для определения связи между переменными

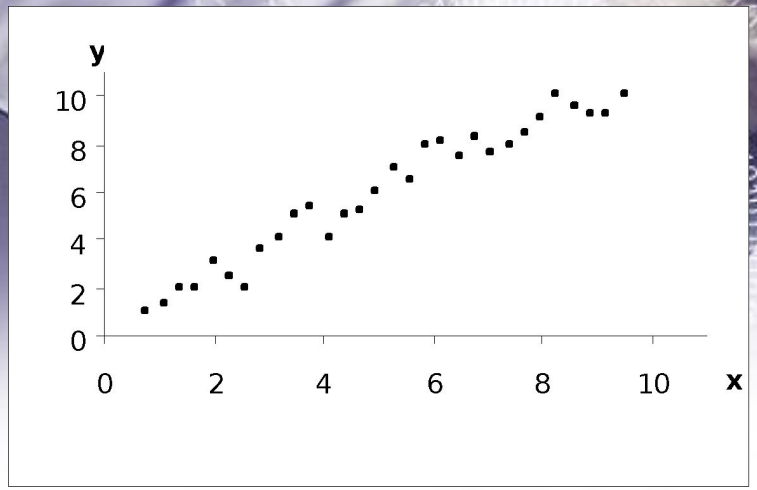
- ϕ коэффициент корреляции Пирсона
- τ - коэффициент корреляции Кендалла
- R – бисериальный коэффициент корреляции
- η - корреляционное отношение Пирсона
- r_S - коэффициент ранговой корреляции Спирмена
- r_{xy} - коэффициент линейной корреляции Пирсона
- Множественная и частная корреляция
- Линейная, криволинейная и множественная регрессия
- Факторный и кластерный анализы



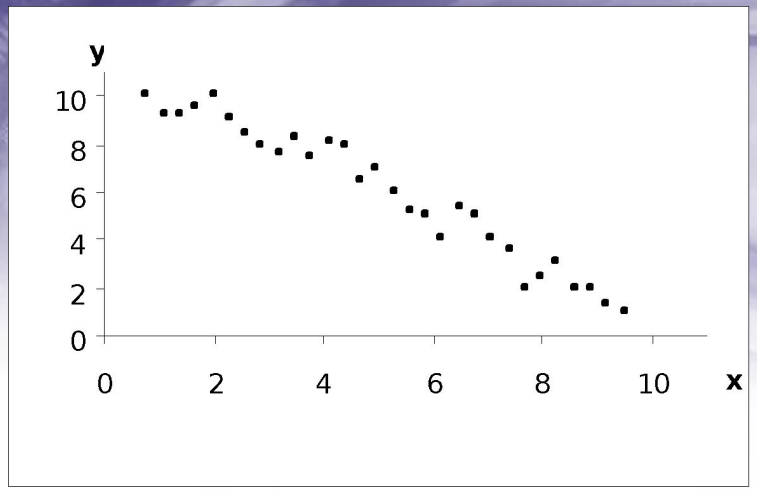
Структура эксперимента



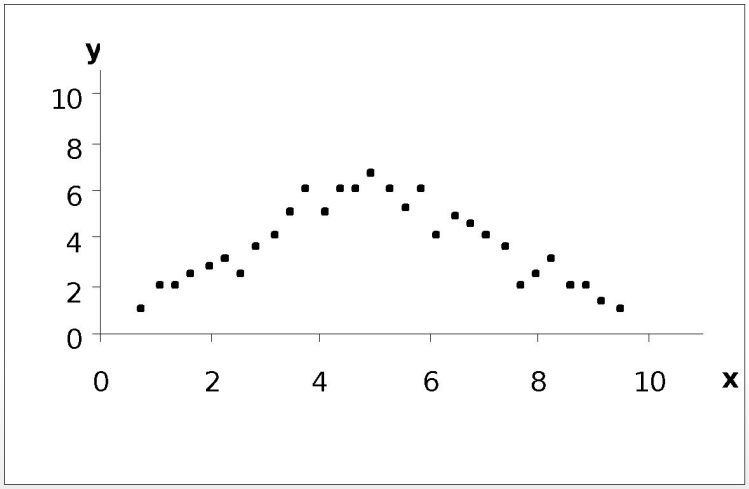
Корреляционный анализ



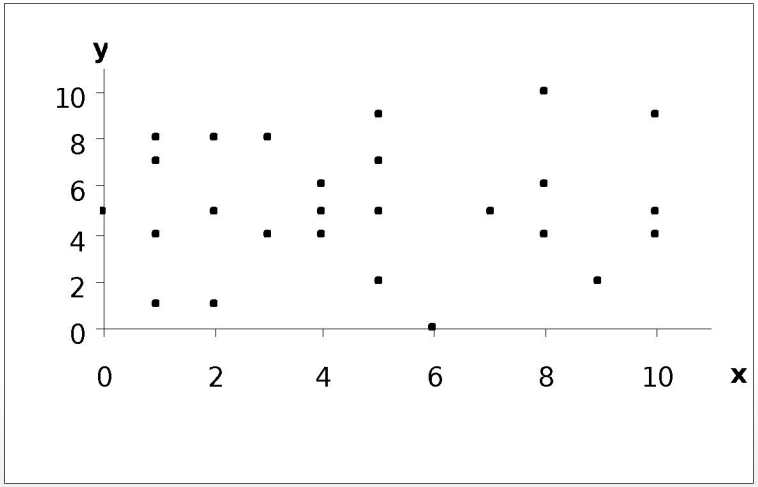
Линейная положительная связь



Линейная отрицательная связь



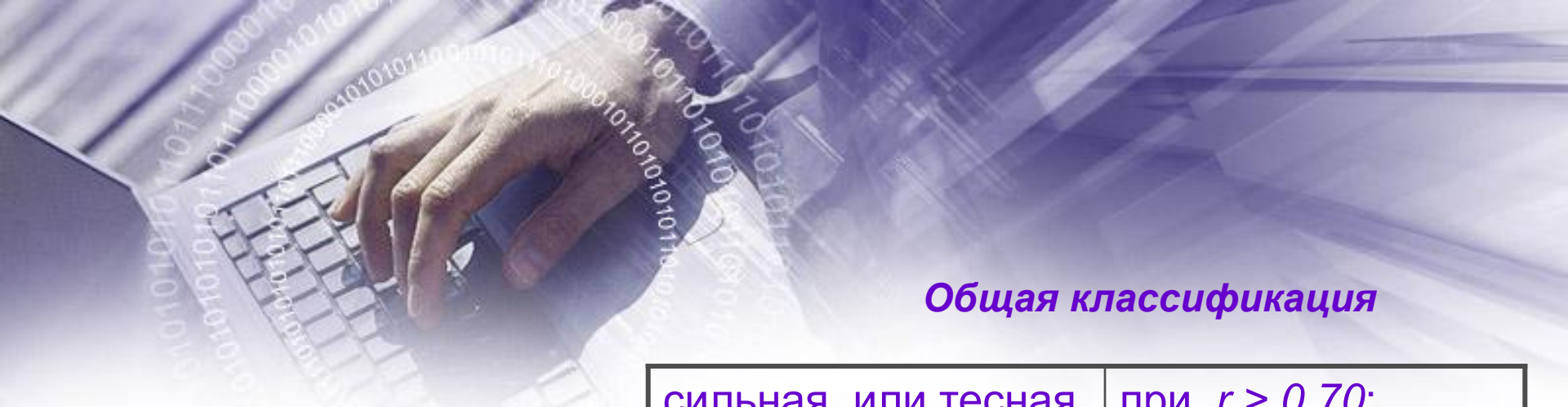
Криволинейная связь



Случайная связь

- ***Степень*** (сила или теснота) корреляционной связи определяется по величине коэффициента корреляции, обозначающегося часто как ***r*.**
- $-1 \leq r \leq +1$.
- ***Сила*** связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции ***/r/*.**
- Если коэффициент корреляции по модулю оказывается близким к 1, то это соответствует высокому уровню связи между переменными.

Тип шкалы		Мера связи
Переменная А	Переменная В	
Интервальная или отношений	Интервальная или отношений	r_{xy} – коэффициент линейной корреляции Пирсона
Порядковая, интервальная или отношений	Порядковая, интервальная или отношений	r_S - коэффициент ранговой корреляции Спирмена
Порядковая	Порядковая	τ - коэффициент корреляции Кендалла
Дихотомическая	Дихотомическая	ϕ коэффициент корреляции Пирсона
Дихотомическая	Порядковая	R_{rb} – рангово-бисериальный коэффициент корреляции
Дихотомическая	Интервальная или отношений	$R_{бис}$ – бисериальный коэффициент корреляции
Интервальная	Порядковая	Не разработан



Пример

- r= 0,67
- r= 0,12
- r= 0,98
- r= -0,67
- r= -0,13
- r= 0,79

?

Общая классификация

сильная, или тесная	при $r > 0,70$;
средняя	при $0,50 < r < 0,69$;
умеренная	при $0,30 < r < 0,49$;
слабая	при $0,20 < r < 0,29$;
очень слабая	при $r < 0,19$.

Частная классификация

- $r_{0,05}$= 0,69
- $r_{0,01}$= 0,89
- $r_{0,01}$= 0,78