 **Statistica**

**ГОТОВИМСЯ к отчету НИР с программой Statistica**

Приближается время отчетов по НИР, набора и анализа данных, интерпретации полученных результатов экспериментов. В настоящее время хорошим инструментом для исследователя может стать программа Statistica. **Русская версия программы Statistica для Windows  –** продуманный комплекс, основным значением которого является статистический анализ информации. Специфичное и достаточно функциональное решение пользуется заслуженным спросом. В умелых руках и при правильном подходе этот инструмент даст точные результаты, поможет решить множество задач.

Программа Statistica создана компанией Statsoft. Statistica имеет пакет прикладных программ – набор модулей, каждый из которых содержит определенную группу инструментов. Программное решение работает с текстовой и численной информацией. Все данные, которые фиксируются в продукте, преобразованы в виде переменных и наблюдений. Пользователи могут объединять обрабатываемую информацию в одну из четырех групп: управление, анализ, получение или визуализация.

При помощи нескольких несложных действий пользователь сможет выбрать и применить аналитические инструменты, которые требуются для статистического анализа.

Statistica содержит качественные инструменты для построения графиков и схем; в этой программе можно сформировать отчеты и переносить их в Word; программа обеспечивает высокую точность расчетов; есть руководство по решению статистических операций; разнообразные методы редактуры графиков, схем, диаграмм; возможность запуска системы через Excel; доступность средств детализации и качественной графики.

С помощью этой программы возможно выполнение таких процедур, как: вычисление корреляций; множественной регрессия; выполнение графическогой анализа таблиц; вычисление экстремумов и т.д. Программа позволяет использовать разведочные технологии при помощи деревьев классификации, провести кластерный, дискриминантный и факторный анализы и другие позиции. **Statistica имеет электронное руководство с примерами выполнения анализа, что очень ценно при освоении программы.** Электронное руководство Statistica предоставляет подробную информацию обо всех процедурах, возможностях и параметрах системы Statistica. На панели инструментов программы нажмите на одну из ссылок для получения более подробной информации, или используйте *Содержание*, находящееся слева в *Электронном руководстве*.

**Перечислим основные термины и понятия в Statistica**

Основу [статистического исследования](http://www.grandars.ru/student/statistika/statisticheskoe-issledovanie.html) составляет множество данных, полученных в результате измерения одного или нескольких признаков. Реально наблюдаемая совокупность объектов, статистически представленная рядом наблюдений случайной величины , является **выборкой**, а гипотетически существующая (домысливаемая) — **генеральной совокупностью**. Генеральная совокупность может быть конечной (число наблюдений **N = const**) или бесконечной (**N = ∞**), а выборка из генеральной совокупности — это всегда результат ограниченного ряда наблюдений.



Пусть Х1, Х2 ... Xn – **выборка** независимых случайных величин.

Упорядочим эти величины по возрастанию, иными словами, построим вариационный ряд:

Х(1) < Х(2) < ... < X (n) ,   (\*)

где Х(1) = min ( Х1, Х2 ... Xn),

Х(n) = max ( Х1, Х2 ... Xn).

Элементы вариационного ряда (\*) называются порядковыми статистиками.

Величины d(i) = X(i+1) - X(i) называются спейсингами или расстояниями между порядковыми статистиками.

**Размахом** выборки называется величина

R = X(n) - X(1)

Иными словами, размах это расстояние между максимальным и минимальным членом вариационного ряда.

Выборочное среднее равно:  = (Х1 + Х2 + ... + Xn) / n

**Среднее арифметическое.**

Большинство из вас использовало такую важную описательную статистику, как среднее. Среднее – очень информативная мера "центрального положения" наблюдаемой переменной, особенно если сообщается ее доверительный интервал. Исследователю нужны такие статистики, которые позволяют сделать вывод относительно популяции в целом. Одной из таких статистик является среднее.

**Доверительный интервал** для среднего представляет интервал значений вокруг оценки, где с данным уровнем доверия, находится "истинное" (неизвестное) среднее популяции.

Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p = .95 равны 19 и 27 соответственно, то можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее популяции.

Если вы установите больший уровень доверия, то интервал станет шире, поэтому возрастает вероятность, с которой он "накрывает" неизвестное среднее популяции, и наоборот.

Хорошо известно, например, что чем "неопределенней" прогноз погоды (т.е. шире доверительный интервал), тем вероятнее он будет верным. Заметим, что ширина доверительного интервала зависит от объема или размера выборки, а также от разброса (изменчивости) данных. Увеличение размера выборки делает оценку среднего более надежной. Увеличение разброса наблюдаемых значений уменьшает надежность оценки.

Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Если это предположение не выполнено, то оценка может оказаться плохой, особенно для малых выборок.

При увеличении объема выборки, скажем, до 100 или более, качество оценки улучшается и без предположения нормальности выборки.

Довольно трудно «ощутить» числовые измерения, пока данные не будут содержательно обобщены. Диаграмма часто полезна в качестве отправной точки. Мы можем также сжать информацию, используя важные характеристики данных. В частности, если бы мы знали, из чего состоит представленная величина, или если бы мы знали, насколько широко рассеяны наблюдения, то мы бы смогли сформировать образ этих данных.

Среднее арифметическое, которое очень часто называют просто «среднее», получают путем сложения всех значений и деления этой суммы на число значений в наборе.

Это можно показать с помощью алгебраической формулы. Набор n наблюдений переменной X можно изобразить как X1, X2, X3, ..., Xn. Например, за X можно обозначить высоту растений (см), X1 обозначит высоту 1-го растения, а Xi — высоту i-го растения. Формула для определения среднего арифметического наблюдений  (произносится «икс с чертой»):

 = (Х1 + Х2 + ... + Xn) / n

Можно сократить это выражение:



где  (греческая буква «сигма») означает «суммирование», а индексы внизу и вверху этой буквы означают, что суммирование производится от i = 1 до i = n. Это выражение часто сокращают еще больше:

 или 

**Медиана**

Если упорядочить данные по величине, начиная с самой маленькой величины и заканчивая самой большой, то медиана также будет характеристикой усреднения в упорядоченном наборе данных.

Медиана делит ряд упорядоченных значений пополам с равным числом этих значений как выше, так и ниже ее (левее и правее медианы на числовой оси).

Вычислить медиану легко, если число наблюдений n нечетное. Это будет наблюдение номер (n + 1)/2 в нашем упорядоченном наборе данных.

Например, если n = 11, то медиана — это (11 + 1)/2, т. е. 6-е наблюдение в упорядоченном наборе данных.

Если n четное, то, строго говоря, медианы нет. Однако обычно мы вычисляем ее как среднее арифметическое двух соседних средних наблюдений в упорядоченном наборе данных (т. е. наблюдений номер (n/2) и (n/2 + 1)).

Так, например, если n = 20, то медиана — это среднее арифметическое наблюдений номер 20/2 = 10 и (20/2 + 1) = 11 в упорядоченном наборе данных.

**Мода**

Мода — это значение, которое встречается наиболее часто в наборе данных; если данные непрерывные, то мы обычно группируем их и вычисляем модальную группу.

Некоторые наборы данных не имеют моды, потому что каждое значение встречается только 1 раз. Иногда бывает более одной моды; это происходит тогда, когда 2 значения или больше встречаются одинаковое число раз и встречаемость каждого из этих значений больше, чем любого другого значения.

Как обобщающую характеристику моду используют редко.

**Среднее геометрическое**

При несимметричном распределении данных сред­нее арифметическое не будет обобщающим показа­телем распределения.

Если данные скошены вправо, то можно создать более симметричное распределе­ние, если взять логарифм (по основанию 10 или по основанию е) каждого значения переменной в наборе данных. Среднее арифметическое значений этих логарифмов — характеристика распределения для преобразованных данных.

Чтобы получить ме­ру с теми же единицами измерения, что и первона­чальные наблюдения, нужно осуществить обратное преобразование — потенцирование (т. е. взять анти­логарифм) средней логарифмированных данных; мы называем такую величину среднее геометрическое.

Если распределение данных логарифма приблизитель­но симметричное, то среднее геометрическое подобно медиане и меньше, чем среднее необработанных дан­ных.

**Взвешенное среднее**

Взвешенное среднее используют тогда, когда не­которые значения интересующей нас переменной x более важны, чем другие. Мы присоединяем вес wi к каждому из значений xi в нашей выборке для то­го, чтобы учесть эту важность.

Если значения x1, x2 ... xn имеют соответствующий вес w1, w2 ... wn, то взвешенное арифметическое среднее выглядит следующим образом:



Например, предположим, что мы заинтересованы в определении средней урожайности зерновых культур (пшеницы, ячменя и др.) в области и знаем среднюю урожайность в каждом хозяйстве района. Учитываем количество информации, в первом при­ближении принимая за вес каждого наблюдения число разных зерновых культур.

Взвешенное среднее и среднее арифметическое идентичны, если каждый вес равен единице.

**Размах** (интервал изменения)

Размах  — это разность между максимальным и минимальным значениями переменной в наборе данных; этими двумя величинами обозначают их разность. Обратите внимание, что размах вводит в заблуждение, если одно из значений есть выброс.
 **Размах, полученный из процентилей**

Предположим, что мы расположим наши данные упорядоченно от самой маленькой величины перемен­ной X и до самой большой величины. Величина X, до которой расположен 1% наблюдений (и выше которой расположены 99% наблюдений), называется первым процентилем.

Величина X, до которой находится 2% наблюдений, называется 2-м процентилем, и т. д.

Величины X, которые делят упорядоченный набор значений на 10 равных групп, т. е. 10-й, 20-й, 30-й,..., 90 и процентили, называются децилями. Величины X, которые делят упорядоченный набор значений на 4 равные группы, т.е. 25-й, 50-й и 75-й процентили, называются квартилями. 50-й процентиль — это ме­диана.

Применение процентилей

Мы можем добиться такой формы описания рас­сеяния, на которую не повлияет выброс (аномальное значение), исключая экстремальные величины и определяя размах остающихся наблюдений.

Межквартильный размах — это разница между 1-м и 3-м квартилями, т.е. между 25-м и 75-м процентилями. В него входят центральные 50% наблюдений в упорядоченном наборе, где 25% наблюдений находятся ниже центральной точки и 25% — выше.

Интердецильный размах содержит в себе центральные 80% наблюдений, т. е. те наблю­дения, которые располагаются между 10-м и 90-м процентилями.

Мы часто используем размах, который содержит 95% наблюдений, т.е. он исключает 2,5% наблюдений снизу и 2,5% сверху. Указание такого интервала актуально, например, для осуществления диагностики заболеваний растений. Такой интервал называется референтный интервал, референтный размах или нормальный размах.

**Степени свободы**

Количество степеней свободы — это количество значений в итоговом вычислении [статистики](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_%28%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B8%29), способных варьироваться. Иными словами, количество степеней свободы показывает [размерность](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%B0) [вектора](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80_%28%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0%29) из случайных величин, количество «свободных» величин, необходимых для того, чтобы полностью определить вектор.

Количество степеней свободы может быть не только [натуральным](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%82%D1%83%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%87%D0%B8%D1%81%D0%BB%D0%BE), но и любым [действительным](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D1%89%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D1%87%D0%B8%D1%81%D0%BB%D0%BE) числом, хотя стандартные таблицы рассчитывают *p-value*наиболее распространённых [распределений](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9) только для натурального числа степеней свободы.

**Дисперсия**

Один из способов измерения рассеяния данных за­ключается в том, чтобы определить степень отклоне­ния каждого наблюдения от средней арифметической. Очевидно, что чем больше отклонение, тем больше изменчивость, вариабельность наблюдений.

Однако мы не можем использовать среднее этих отклонений как меру рассеяния, потому что положительные от­клонения компенсируют отрицательные отклонения (их сумма равна нулю). Чтобы решить эту проблему, мы возводим в квадрат каждое отклонение и находим среднее возведенных в квадрат отклонений; эта величина называется вариацией, или дисперсией.

Возьмем n наблюдений x1, x2, х3, ..., xn, среднее которых равняется .

Вычисляем дисперсию:



В случае, если мы имеем дело не с генеральной совокупностью, а с выборкой, то вычисляется выборочная дисперсия:



Теоретически можно показать, что полу­чится более точная дисперсия по выборке, если разделить не на n, а на (n-1).

Единицы измерения (размерность) вариации — это квадрат единиц измерения первоначальных на­блюдений.

Например, если измерения производятся в килограммах, то единица измерения вариации будет килограмм в квадрате.

**Среднеквадратическое отклонение, стандартное отклонение выборки**

Среднеквадратическое отклоне­ние — это положительный квадратный корень из [дисперсии](http://statistica.ru/theory/opisatelnye-statistiki/#%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D1%8F).

Стандартное отклонение выборки - корень из выборочной дисперсии:



Мы можем представить себе стандартное отклоне­ние как своего рода среднее отклонение наблюдений от среднего. Оно вычисляется в тех же единицах (размерностях), что и исходные данные.

Если разделить стандартное отклонение на сред­нее арифметическое и выразить результат в процен­тах, получится коэффициент вариации.

Он являет­ся мерой рассеяния, не зависит от единиц измерения (безразмерный), но имеет некоторые теоретические не­удобства и поэтому не очень одобряется статистиками.

Вариация в пределах субъектов и между субъектами

Если провести повторные измерения непрерывной переменной у исследуемого объекта, то можно увидеть ее изме­нения (внутрисубъектные изменения). Это можно объяснить тем, что объект не всегда может дать точные и те же самые ответы, и/или ошибкой, погрешностью измерения. Однако при измерениях у одного объекта вариация обычно меньше, чем вариация единичного измерения в группе (межсубъектные изменения).

Например, масса тысячи семян растения от 360 до 387 г, когда измерения повторяются не менее 10 раз; если провести однократное измерение у 10 растений такой же популяции, то масса будет между 298 и 433 г. Эти концепции важны в плане исследования.

**t-критерий Стьюдента** – общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Наиболее частые случаи применения t-критерия связаны с проверкой равенства средних значений в двух выборках.

Применяется для проверки гипотезы об отличии среднего значения  от некоторого известного значения m



Количество степеней свободы рассчитывается как v=n-1. Граничные значения критерия можно расчитать по таблице



**Что такое статистическая значимость (*p*-уровень)?**

Статистическая значимость результата представляет собой оцененную меру уверенности в его "истинности". Выражаясь более технически, *p*-уровень – это показатель, находящийся в убывающей зависимости от надежности результата. Более высокий *p*-уровень соответствует более низкому уровню доверия к найденной в выборке зависимости между переменными.

Именно, *p*-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю популяцию. Например, *p*-уровень = .05 (т.е. 1/20) показывает, что имеется 5% вероятность, что найденная в выборке связь между переменными является лишь случайной особенностью данной выборки. Иными словами, если данная зависимость в популяции отсутствует, а вы многократно проводили бы подобные эксперименты, то примерно в одном из двадцати повторений эксперимента можно было бы ожидать такой же или более сильной зависимости между переменными.

Во многих исследованиях *p*-уровень .05 рассматривается как "приемлемая граница" уровня ошибки.

**Коэффициент корреляции**

Корреляционный анализ занимается степенью связи между двумя переменными, x и y.

Сначала предполагаем, что как x, так и y количественные, например высота и масса растения. Предположим, пара величин (x, у) измерена у каждого из n объектов в выборке.

Мы можем отметить точку, соответствующую паре величин каждого объекта, на двумерном графике рассеяния точек.

Обычно на графике переменную x располагают на горизонтальной оси, а у — на вертикальной. Размещая точки для всех n объектов, получают график рассеяния точек, который говорит о соотношении между этими двумя переменными.

Коэффициент корреляции Пирсона

Соотношение х и у линейное, если прямая линия, проведенная через центральную часть скопления точек, дает наиболее подходящую аппроксимацию наблюдаемого соотношения.

Можно измерить, как близко находятся наблюдения к прямой линии, которая лучше всего описывает их линейное соотношение путем вычисления коэффициента корреляции Пирсона, обычно называемого просто коэффициентом корреляции.

Его истинная величина в популяции (генеральный коэффициент корреляции) (греческая буква «r») оценивается в выборке как r (выборочный коэффициент корреляции), которую обычно получают в результатах компьютерного расчета.

Пусть (x1. y1), (x2, y2),…,(xn, yn) – выборка из n наблюдений пары переменных (X, Y).

Выборочный коэффициент корреляции r определяется как

,

где ,  - выборочные средние, определяющиеся следующим образом:



Свойства коэффициента корреляции r

r изменяется в интервале от —1 до +1.

Знак r означает, увеличивается ли одна переменная по мере того, как увеличивается другая (положительный r), или уменьшается ли одна переменная по мере того, как увеличивается другая (отрицательный r).

Величина r указывает, как близко расположены точки к прямой линии. В частности, если r = +1 или r= —1, то имеется абсолютная (функциональная) корреляция по всем точкам, лежащим на линии (практически это маловероятно); если , то линейной корреляции нет (хотя может быть нелинейное соотношение). Чем ближе r к крайним точкам (±1), тем больше степень линейной связи.

Коэффициент корреляции r безразмерен, т. е. не имеет единиц измерения.

Величина r обоснована только в диапазоне значений x и y в выборке. Нельзя заключить, что он будет иметь ту же величину при рассмотрении значений x или y, которые значительно больше, чем их значения в выборке.

x и y могут взаимозаменяться, не влияя на величину r ().

Корреляция между x и у не обязательно означает соотношение причины и следствия.

 представляет собой долю вариабельности у, которая обусловлена линейным соотношением с x.

Когда не следует рассчитывать r

Расчет r может ввести в заблуждение, если:

соотношение между двумя переменными нелинейное, например квадратичное;

данные включают более одного наблюдения по каждому случаю;

есть аномальные значения (выбросы);

данные содержат ярко выраженные подгруппы наблюдений.

 Для определения статистической достоверности [корреляционной связи r Пирсона](https://statpsy.ru/correlations/linear-pirson/)

используется таблица критических значений корреляции Пирсона



**Регрессия**— зависимость [математического ожидания](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BE%D0%B6%D0%B8%D0%B4%D0%B0%D0%BD%D0%B8%D0%B5) (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть . Регрессионным анализом называется поиск такой функции , которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.



где  — функция регрессионной зависимости, а  — аддитивная случайная величина с нулевым матожиданием. Предположение о характере распределения этой величины называется [гипотезой порождения данных](http://www.machinelearning.ru/wiki/index.php?title=%D0%93%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0_%D0%BF%D0%BE%D1%80%D0%BE%D0%B6%D0%B4%D0%B5%D0%BD%D0%B8%D1%8F_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85&action=edit). Обычно предполагается, что величина  имеет [гауссово распределение](http://www.machinelearning.ru/wiki/index.php?title=%D0%93%D0%B0%D1%83%D1%81%D1%81%D0%BE%D0%B2%D0%BE_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) с нулевым средним и дисперсией .

В целом программа Statistica позволяет решить общие вопросы сбора, измерения, мониторинга и анализа массовых статистических (количественных или качественных) данных; изучить количественную сторону различных явлений в числовой форме путем использования различных видов анализа; предоставляет сотни двумерных и трехмерных графиков, включая двумерные и трехмерные тернарные графики, специализированные четырехмерные графики, многомерные графики, матричные графики, пиктограммы, мозаики, спектральные двумерные и трехмерные графики, составные графики и много других специализированных процедур.

**Дополнительно StatSoft Russia рекомендует следующие учебные пособия на русском языке:**

*Боровиков В.П. STATISTICA, искусство анализа данных на компьютере, Питер 2003 (второе издание).* В книге изложена концепция и технология современного анализа данных на компьютере. На основе элементарных понятий описываются углубленные методы анализа данных, иллюстрированные примерами из экономики, маркетинга, рекламы, бизнеса, медицины, промышленности и других областей.

*Ивченко Г.И., Медведев Ю.И. Математическая статистика, Высшая школа, 1992.* В пособии на современном научном уровне изложены основные разделы статистической теории.

*Боровиков В.П. Популярное введение в программу STATISTICA, Компьютер Пресс 1998.* Книга, посвященная анализу данных, построению зависимостей, классификации и прогнозированию в системе STATISTICA.

*Боровиков В.П., Боровиков И.П. STATISTICA. Статистический анализ и обработка данных в среде Windows, Филинъ 1998.* Справочное и учебное пособие по системе *STATISTICA*.

*Боровиков В.П, Ивченко Г.И. Прогнозирование в системе STATISTICA в среде Windows, Финансы и статистика 1999.* Учебное пособие, содержащее описание практических методов и приемов прогнозирования и изложение теоретических основ.

*Розанов Ю.А. Теория вероятностей, случайные процессы и математическая статистика, Наука 1985.* Книга представляет собой единый учебный курс теории вероятностей, случайных процессов и математической статистики. Изложение материала таково, что книга во многих важных разделах доступна широкому кругу читателей.

Замечательным введением в элементарную статистику с разнообразными примерами из медицины и генетики является книга *Ю. Неймана Вводный курс теории вероятностей и математической статистики, Наука, 1968 (перевод с английского).*

Отличным справочным пособием для практиков является книга коллектива авторов *Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер Многомерный статистический анализ в экономике.* М., Юнити 1999. Достаточно полно представлены теоретические основы и важнейшие методы многомерной статистики, открывающей для исследователя широкие возможности моделирования сложных реальных процессов, явления и визуализации данных.