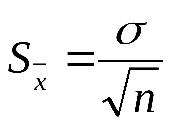
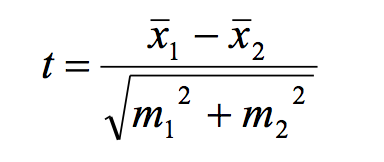
**Основные формулы**

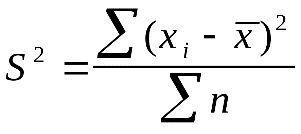
Формула расчета СТАНДРАТНОГО ОТКЛОНЕНИЯ в выборке (обозначается еще как sigma **σ)**

****

Формула расчета СТАНДРАТНОЙ ОШИБКИ в выборке(стандратное отклоение ГС\ корень кол-ва наблюдений выборки)

****

Формула расчета t-критерия Стьюдента



Формула расчета ДИСПЕРСИИ выборки (также для дисперсии)



Формула расчета доверительного интервала (где t-критерий Стьюдента, а mr стандратная ошибка(иначе обозначается,как Se(standard error))

**Конспект по основам статистики**

**Основные понятия:**

1. **Мода** – самый часто встречающийся показатель в Генеральной совокупности
2. **Медиана** – середина упорядоченной Генеральной совокупности (если в генеральной совокупности присутствует несколько одинаковых значений среднего, то найти медиану можно по формуле: сумма всех одинаковых значений среднего \ кол-во этих значений)
3. **Среднее значение (х̅)** - результат сложения всех показателей Генеральной совокупности, деленный на их кол-во.
4. **Выброс** – значение в Генеральной совокупности, значительно отличающееся как в меньшую, так и в большую стороны от большинства значений Генеральной совокупности.

Выше были рассмотрены основные методы нахождения типичного размера. Называются они **меры центральной тенденции**.

**Меры центральной тенденции:**

1. Мода
2. Медиана
3. Среднее значение **(х̅)**

**Меры изменчивости:**

1. **Размах** - является разностью между самым большим и самым маленьким котиком.
2. **Дисперсия** – среднее от квадратов отклонений (сумма квадратов всех отклонений, деленная на кол-во наблюдений).   
   Для удобства использования Дисперсию берут под корень, получается по итогу среднеквадратическое отклонение или стандартное отклонение.
3. **Стандартное отклонение (σ - “sigma”)** — это положительный квадратный корень дисперсии.

В связи с тем, что на практике мы не можем замерить всю Генеральную совокупность, как правило работа с данными происходит в небольшой выборке(ах). Очень важно, чтобы выборка была максимально похожа на Генеральную совокупность. Была максимально **Репрезентативна.**

**— Репрезентативность -** соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом.

Существует всего 2 формулы **Дисперсии**. Одна формула для генеральной совокупности и одна формула для выборки.

Формула для дисперсии генеральной совокупности:

***D = Ʃ(x- х̅)/n***

Формула для дисперсии в выборке:

***D = Ʃ(x- х̅)/n-1***

Корень из Дисперсии по генеральной совокупности называется:

**среднеквадратическим отклонением**

Корень из Дисперсии по выборке называется:

**стандартным отклонением**

**Средства визуализации Данных**

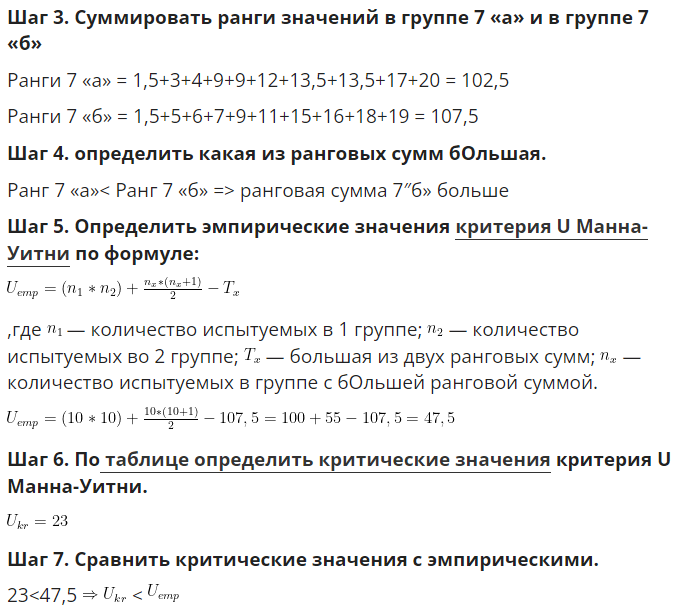
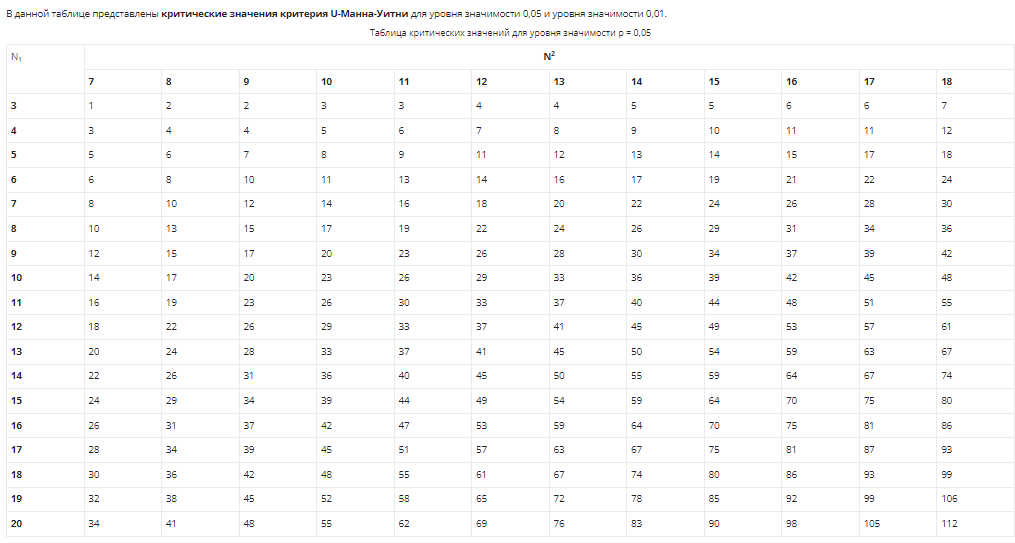
1. столбиковая диаграмма
2. полигон распределения
3. круговая диаграмма
4. пузырьковая диаграмма
5. **боксплот** (или «ящик с усами»)

**боксплот (Boxplot)-** позволяет компактно отобразить медиану, общий и межквартильный размах, а также прикинуть, насколько распределение ваших данных близко к нормальному и есть ли у вас выбросы.

**Меры различий для несвязанных выборок**

1. **t-критерий Стьюдента для несвязанных выборок** - оценивает, насколько различаются их средние значения. Чтобы рассчитать этот критерий, необходимо из среднего размера первой выборки вычесть средний размер второй выборки и поделить их на стандартную ошибку этой разности. Последняя вычисляется на основе стандартных отклонений первой и второй выборки деленной на корень из кол-ва наблюдений и нужна для приведения t-критерия к нужной размерности.(sigma\корень N)  
   Чем больше t-критерий, тем с большей вероятностью мы можем утверждать, что первая выборка в среднем ее значении отличается от второй выборки в её среднем значении
2. **u-критерий** (непараметрическим U-критерием Манна-Уитни). Этот критерий используется, когда в выборках присутствуют выбросы или не точные данные, например, когда не известна длинна в см, если мы рассматриваем рост, как переменную.

Чтобы рассчитать критерий Манна-Уитни, необходимо **проранжировать** все наблюдения по убыванию, от самого крупного к самому мелкому и назначить им ранги. Самому большому зверьку достанется первый ранг, а самому маленькому – последний.  
После этого мы снова делим их на две группы и считаем суммы рангов отдельно для песиков и для котиков.

будут различаться эти суммы, тем больше различаются песики и котики.****

Гипотеза H0 о незначительности различий между выборками принимается, если Uкритическая < uэмпирическая. В противном случае H0 отвергается и различие определяется как существенное.  
где Ukp - критическая точка, которую находят по таблице Манна-Уитни.

1. **F-критерий равенства дисперсий Фишера**f-критерий равенства дисперсий Фишера применяется в том случае, если нас могут интересовать различия по разнообразию в некоторой выборке. Он укажет нам, насколько различаются между собой показатели.  
   формула расчета критерия:  
   ****  
   Для вычисления $F_{{э}{м}{п}}$ нужно найти отношение дисперсий двух выборок, причем так, чтобы большая по величине дисперсия находилась бы в числителе, а меньшая знаменателе

**критерий Хи-квадрат Пирсона**подробное описание с примером

[**https://lit-review.ru/biostatistika/kriterijj-khi-kvadrat-pirsona/**](https://lit-review.ru/biostatistika/kriterijj-khi-kvadrat-pirsona/)

**Число степеней свобод**

Число степеней свобод обозначается **df**, говоря простыми словами число степеней свобод обозначает какое число измерений нам необходимо сделать в одной выборке или нескольких,чтобы понять размер каждого наблюдения,например высота животного.   
**Пример:**  
Предположим вы знаете, что сумма размеров всех ваших котиков равна 75 см, но не знаете величину каждого конкретного котика. Эти величины будут неизвестны ровно до тех пор, пока вы не начнете их измерять. Представим, что вы узнали размер первого котика и он оказался равен 20 см. После несложных вычислений можно убедиться, что сумма размеров оставшихся котиков будет 55 см. При этом их конкретные размеры до сих пор неизвестны. Измерим второго котика. Он оказался равен 25 см. Что мы можем сказать о размере третьего? То,что его размер равен разности 75 и 45,следовательно число степеней свобод это то кол-во котиков,которое мы должны измерить,чтобы однозначно узнать размер всех животных. Если выборка только одна,то число степеней свобод равно все животные минус 1,если выборки 2 и более,то общее кол-во животных минус кол-во выборок. Такой расчет нам и даст число степеней свобод.   
**Формула:**

Для 1 выборки : df = n -1   
Для n-ого кол-ва выборок: df = ***Ʃn – кол-во выборок***

**P-уровень значимости**

Что же такое p-уровень значимости? Во-первых, при выяснении уровня значимости делается предположение,а именно вводится **НУЛЕВАЯ ГИПОТЕЗА**, например: в рамках нулевой гипотезы ученые могут утверждать,что два разных вида животных идентичны друг другу. Следующим шагом они вычисляют вероятность того,что две случайные группы животных дадут значение критерия большее или равное тому, которое мы получили (чаще всего без учета его знака). Эта вероятность и называется p-уровень значимости.

Если р-уровень значимости меньше 5%(обычно это записывается как 0,05), то нулевая гипотеза отвергается и принимается гипотеза о том, что группы выбранных животных все таки отличаются. Такая гипотеза называется **АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА**.

**Доверительный интервал**

Доверительный интервал – диапазон в Генеральной Совокупности,в котором находится истинно среднее значение наблюдения. Рядом с доверительным интервалом всегда указывается вероятность. Например, 95%-й доверительный интервал означает, что мы с точностью 95% можем утверждать,что истинно средний размер наблюдений находится в этом диапазоне. Формула для расчета доверительного интервала:

(-+ делта = t-критерий Стьюдента \* se(standard error)).

**Основы однофакторного дисперсионного анализа** **(ANOVA)**

Дисперсионный анализ метод,который применяется в том случае, если мы не можем воспользоваться стандартными методами попарного сравнения,будь то t-критерия Стьюдента или U-критерий Манна-Уитни, так как выборок становится больше двух. Таким образом, дисперсионный анализ хорош тем,что кол-во групп (выборок) может быть не ограничено,будь то две, три или четыре группы. Нулевая гипотеза дисперсионного анализа состоит в том,что наблюдения из рассматриваемого кол-ва групп схожи между собой. Альтернативная же гипотеза утверждает,что хотя бы одно наблюдение из рассматриваемых групп отличается от остальных.

## Условия применения дисперсионного анализа ANOVA

Перед тем как приступить к применению дисперсионного анализа, который предназначен для минимизации риска неправильной оценки ошибки 1 рода в случае множественных сравнений необходимо убедиться в соблюдении ряда условий:

1. Количественный непрерывный тип данных, дискретные данные менее желательны.
2. Независимые между собой выборки.
3. Нормальное распределение признака в статистических совокупностях, из которых извлечены выборки.
4. Равенство (гомогенность) дисперсий изучаемого признака в статистических совокупностях из которых извлечены выборки, проверяется с помощью критерия Levene.
5. Независимые наблюдения в каждой из выборок.

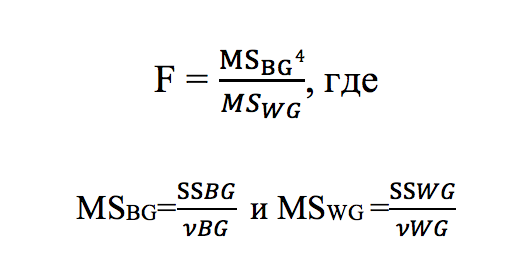
## Статистическая информация для применения однофакторного дисперсионного анализа

Ho в случае однофакторного дисперсионного анализа (ANOVA) подразумевает, что средние генеральных совокупностей из которых были извлечены выборки равны, другими словами все они относятся к одной генеральной совокупности и различия носят случайный характер. Для проверки теорий в случае дисперсионного анализа используется F-распределение. F-статистика принимает только положительные или нулевые значения.

Процедура дисперсионного анализа состоит в определении соотношения систематической (межгрупповой) дисперсии к случайной (внутригрупповой) дисперсии в измеряемых данных. В качестве показателя изменчивости используется сумма квадратов отклонения значений параметра от среднего: SS (от [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) Sum of Squares). Общая сумма квадратов SSTotal раскладывается на межгрупповую сумму квадратов SSBG[[1]](https://lit-review.ru/biostatistika/dispersionnyjj-analiz-anova/" \l "_ftn1) и внутригрупповую сумму квадратов SSWG***[[2]](https://lit-review.ru/biostatistika/dispersionnyjj-analiz-anova/" \l "_ftn2)***:

SSTotal = SSBG + SSWG

В случае если верна Ho, то как внутригрупповая, так и межгрупповая дисперсии служат оценками одной и той же дисперсии и должны быть приблизительно равны.

[](http://lit-review.ru/wp-content/uploads/2016/06/--------------------------2016-06-09----1.21.23.png)

Исходя из этого значение F должно быть близко к 1 в случае, если статистически значимых различий все-таки нет. Критическое значение F определяется уровнем значимости (обычно 0,05 или 0,01) и внутригрупповым и межгрупповым числом степеней свободы (ν). Оно достаточно сложно для вычисления, поэтому чаще используются табличные значения с указанием α, νBG,νWG.

**Межгрупповое число степеней свободы:**

νBG = m – 1.

m – число групп

**Внутригрупповое число степеней свободы:**

νWG = n – m

n – количество наблюдений в каждой из групп

## Апостериорные значения

Однако, при обнаружении статистически значимых отличий мы не сможем сказать лишь об их наличии, но какие именно группы отличаются друг от друга мы определить не сможем, для этого производят так называемые процедуры **апостериорных сравнений**. Апостериорные сравнения представляют собой попарные сравнения изучаемых групп для обнаружения различий между ними.

Апостериорные сравнения могут быть проведены с помощью критерия Стьюдента для независимых выборок, что может показаться странным, учитывая сказанное ранее о проблеме множественных сравнений. Однако в отличие от простых попарных сравнений при проведении апостериорных сравнений рассчитываются новые критические уровни значимости для удержания ошибки 1 типа в пределах 5 %.

Наиболее простым и наиболее популярным способом коррекции ошибки 1 типа является **поправка Бонферрони (Bonferroni)**, при которой уровень ошибки 1 типа делится на количество сравнений для получения нового критического уровня значимости. Так, если имеется 3 сравнения, то новый критический уровень должен быть 0,05 / 3 = 0,017. Поправка Бонферрони хорошо контролирует ошибку 1 типа, но является очень консервативной и приводит к повышению вероятности ошибки 2 типа (вероятности принятия решения об отсутствии различий там, где они на самом деле есть). Либеральные критерии, (например **критерий Тьюки**) в свою очередь, завышают вероятность ошибки 1 типа, то есть вероятность принятия решения о наличии различий там, где их нет.

Таким образом, при выборе статистического критерия для апостериорных сравнений необходимо принимать во внимание, как критерии контролируют ошибки 1 и 2 типов и как они работают при несоблюдении необходимых условий применения дисперсионного анализа.

Если данные не подчиняются нормальному распределению, то при анализе можно использовать два способа: применением различных арифметических преобразований до достижения нормальности распределения и дальше уже применять дисперсионный анализ, или использовать **критерий Краскела-Уоллиса (Kruskal-Wallis H-test)**, иногда его также называют непараметрическим дисперсионным анализом. Как и в большинстве непараметрических методов, работающих с количественными данными, исходный набор данных преобразуется в ранги и обрабатывается уже он. При обнаружении статистически значимых различий между группами стоит дальше проводить апостериорные сравнения с использованием критерия Манна-Уитни.

## Пример

В условиях крупной городской клинической больницы было решено провести исследование по оценке влияния возраста на длительность госпитализации после лапароскопической холецистектомии. 9 пациентов были разделены на 3 группы в зависимости от возраста

**Длительность госпитализации**

**после лапароскопической холецистектомии в зависимости от возраста, дни**

|  |  |  |
| --- | --- | --- |
| Группа №1Младше 45 лет | Группа №245-55 лет | Группа №3Старше 55 лет |
| 3 | 5 | 7 |
| 1 | 3 | 6 |
| 2 | 4 | 5 |
| x̄=2 | x̄=4 | x̄=6 |

Сделайте выводы о влиянии возраста на длительности госпитализации после лапароскопической холецистектомии.

1. Постановка нулевой гипотезы

H0указывает на отсутствие различий между группами, иными словами все группы по возрасту относятся к одной генеральной совокупности и соответственно средние равны друг другу

µ1= µ2= µ3

Альтернативная гипотеза выдвигает предположение, что длительно госпитализации зависит от возраста и средние в этих группах на самом деле не равны

µ1≠ µ2≠ µ3

1. Найдем общую сумму квадратов

Для этого нам нужно знать общую среднюю по всем выборкам, найдем ее:

x̄= (2+3+6)=4

SST =2= (3-4)2+(1-4)2+(2-4)2+(5-4)2+(4-4)2+(3-4)2+(7-4)2+

+(6-4)2+(5-4)2=30

1. Найдем сумму квадратов внутри групп последовательно вычитая из каждого значения в группе групповую среднюю:

SSWG = (3-2)2+ (1-2)2+ (2-2)2+ (5-4)2+ (3-4)2+ (4-4)2+ (7-6)2+ (6-6)2+ (5-6)2=2+2+2=6

1. Найдем внутригрупповую сумму квадратов.

Для этого нам необходимо найти квадрат отклонения каждой из выборочных средних относительно общей средней:

SSBG =3(2-4)2+3(4-4)2+3(6-4)2=24

1. Найдем значение критерия Фишера, исходя из средних квадратов отклонений внутри групп и между ними и соответствующих степеней свободы:

νBG = m – 1 = 3-1 = 2

νWG = n – m = 9 – 3 = 6

F= 12, Fкрит.= 5,143 при α = 0,05

F > Fкрит

1. Делаем вывод о наличии статистически значимых отличий между группами:

так как наше значение F больше критического значения при заданном количестве наблюдений и количестве групп, иными словами наша дисперсия между группами вносит больший вклад в любую сумму дисперсий, чем таковая внутри самих групп.

Возраст влияет на длительность госпитализации после холецистектомии.

[[1]](https://lit-review.ru/biostatistika/dispersionnyjj-analiz-anova/" \l "_ftnref1) Sum of squares between groups

[[2]](https://lit-review.ru/biostatistika/dispersionnyjj-analiz-anova/" \l "_ftnref2) Sum of squares within groups

[[3]](https://lit-review.ru/biostatistika/dispersionnyjj-analiz-anova/" \l "_ftnref3)MSBG — Средний квадрат отклонения между группами и MSWG — Средний квадрат отклонения внутри групп