Основы статистики

Одна из задач статистики сделать выводы о генеральной совокупности, основываясь на выборочных данных.

**Генеральная совокупность -** все представители

**Репрезентативная выборка –** случайначасть генеральной, отражающая все её свойства

# Типы переменных :

1. Количественные (дискретные.т е исчисляемые (кол-во детей в семье) и вещественные (рост людей) )
2. Качественные (название городов)

# Меры центральной тенденции (Мю и X):

Нужно для того, чтобы лучше охарактеризовать степень выраженности некоторого кол-во признака

Нужно 3 значения что бы, в зависимости от ситуации выбрать лучший, если наши значения асимметричны, то лучше использовать моду или медиану, чем среднее значение

**Мода –** значение самого встречаемого признака

**Медиана** – значение, которое делит наши упорядоченные признаки пополам

**Среднее значение (арифметическое) –** сумма всех деленное на кол-во всех

**Мю** - в генеральная совокупности

**X -** выборке

# Мера изменчивости:

Нужно для описания изменчивости или разброса данных

**Размах** – разность max и min значения, плохо описывает данные, если есть выбросы

**Дисперсия** – средний квадрат отклонения индивидуальных значений признака от их средней величины деленное на n

Если генеральная совокупность, то n – кол-во элементов

Если выборка, то n – кол-во элементов минус 1

**Среднеквадратическое отклонение** («сигма» или sd) – корень из дисперсии, но в отличии от дисперсии он показывает *реальное* среднее значение наших отклонений от среднего значения по выборке(X)

Если генеральная совокупность, то обозначается **Сигмой**, если по выборке, то **sd** (**стандартное отклонение**)

# Квартили распределения:

**Квантиль** — это значение, ниже которого лежит определённое число наблюдений

**Квантиль** — это значение признака, которые делят упорядоченные данные на некоторое число равных частей (медиана)

**Квартили** – это такие три значения, которые делят наши данные на 4 равные части

**График boxplot** – пример квартили, толста линия — это медиана, остальные 2 линии — это еще 2 медианы на оставшихся кусочках, расстояние между 2 линиями называется – межквартельный размах, линии вдоль это 1.5 межквартильного размаха.

Часто используется, когда надо сравнить 2 группы между собой по одинаковому признаку

# Нормальное распределение:

z- стандартизация – это такое преобразование данных, где среднее значение = 0, стандартное отклонение = 1. Используется для упрощения работы с данными.

Вычитаем из числа среднее и делим на sd, получаем число и посмотрев по таблице, узнаем какую часть из общей части они составляет

Таблица для вероятностей по уровню значимости - <https://gallery.shinyapps.io/dist_calc/>

# Центральная предельная теорема:

n – число наблюдений

Если исследуемый нами признак имеет нормальное распределение в генеральной совокупности, с некой Мю и сигмой и мы многократно извлекаем независимые выборки равные n по объему и в каждой выборке рассчитываем X, после этого строим распределение этих выборочных средних, так вот, такое распределение будет являться нормальным, с X приближенным с Мю, со **стандартным отклонением**(**ошибкой среднего SE**) = сигма деленая на корень n.

Если n больше 30 и выборки репрезентативны, то в ошибке среднего можно вместо сигмы использовать sd на корень n.

Это говорит о том, что если в нашей выборке 100 элементов (n = 100) и sd = 100, основывайся на данных только одной средней, мы можем предположить, как бы вели все выборочные среднее, если бы мы многократно повторяли это исследования

SE = 5 / корень(100) = 0.5

# Доверительные интервалы:

Предложим мы провели исследование, в котором n = 64 (люди от 18 до 34 лет), среднее = 100, sd = 4

Чему равно Мю? Ведь мы не можем позвать всех людей от 18 до 34 лет

Основываясь о среднем значении нашей выборки, мы не можем точно сказать Мю, но можем дать интервал, в котором она будет находиться при помощи ЦПТ

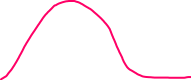
Если бы мы многократно повторяли наш эксперимент, то 95 % выборочных средних расположились бы в интервале Мю +- 2 сигма и с SE = sd на корень(число наблюдений)

Как это нам может помочь? Ведь мы и не знаем Мю…

Но если мы для каждого среднего значения будет рассчитывать 95% интервал   
(среднее +- 2 sd), то Мю попадет в него



95% средних значений



включили бы в свой

интервал Мю



В нашем исследовании среднее значение = 100, sd = 4,n = 64, дальше нашли Se = 0.5

, тогда Мю будет лежать в интервале среднее значение +- 2 se

среднее значение - 2 \* se < Мю < среднее значение + 2 \* se

Значит Мю с 95 % вероятностью лежит тут - 99 < Мю < 101

Для 99% уверенности можно взять интервал +- 2.6 se

# Идея статистического вывода:

Мы разрабатываем лекарство и предположим, что на выздоровление нам требуется 20 дней (Мю = 20), однако мы разработали новый препарат и хотим можно ли нам уменьшить этот срок. Мы набрали выборку из 64 человек и опробовали на них новый метод лечения, оказалось, что средней срок выздоровления X = 18,5 и sd = 4.

Нам надо понять, получилось это случайно или препарат работает?

Между собой будут конкурировать 2 гипотезы

1. Среднее значение тех пациентов, кто использует наш новый препарат не отличается от 20, а те воздействия не оказывает
2. Среднее значение тех пациентов, кто использует наш новый препарат отличается от 20, а те воздействия оказывает

Предположим, что 1 гипотеза верна, тогда согласно с ЦПТ, если бы мы многократно повторяли наш эксперимент, то выборочные средние распределились бы нормальным образом вокруг Мю = 20 и SE = sd / корень(n) = 0.5. На сколько далеко наше X отклонилось от Мю в единицах стандартного отклонения?

Тогда Z-преобразование, Z = (X – Мю) / se = (18.5 – 20) / 0.5 = -3

Это значит, что если Мю = 20, то наше X отклонилось в левую сторону на -3 \* сигма

Вероятность отклонения от Мю на 3 сигмы составляет p = 0.0027

# Уровень Значимости:

Если **уровень значимости** p < 0.05 можно смело принимать альтернативную гипотезу

(учитываются оба конца вероятности)

Надо отметить, что 0.0027 это не вероятность первой гипотезы! Если p принимает более меньшее значение, то это не говорит о том, что препарат еще сильнее бы влиял на скорость выздоровления!

Смысл уровня значимости в том, что если верна первая гипотеза, то получить такие значения можно с вероятность = p. P ничего не говорит о силе нашего эффекта или величине различий.

Если p > 0.05, то правильный вывод это, то что у нас недостаточно оснований что бы отклонить первую гипотезу. Если, например, p = 0.3, то это значит, что наши данные неплохо с ней согласуются, не значит, что она верна с вероятностью 30 %.

Сравнение средних

# t-распределение:

Если в нашей выборки число наблюдений не велико и **сигма не известна**, то используем распределение Стьюдента – **t-распределение** (более большие хвосты, чем у нормального).

Форма распределения Стьюдента зависит от **степеней** **свободы (df = n – 1),** чем больше df, тем больше оно похоже на нормальное

**степеней** **свободы –** кол-во элементов которые могут варьироваться при расчете статистического показателя

# Сравнение двух средних:

У выборки с большим n должна **быть больше** sd, только тогда все работает!

If (n1 > n2) => (sd1 > sd2) and (n1 < n2) => (sd1 < sd2)

(проблема Беренца-Фишера)

Если надо сравнить два выборочных средних, то используют **t-критерий** Стьюдента

Между собой будут конкурировать 2 гипотезы

1. Мю1 = Мю2
2. Мю1 != Мю2

Предположим, что 1 гипотеза верна, тогда если мы многократно будем извлекать выборки и рассчитывать X1 – X2, то X1 – X2 распределение приняло бы нормальный вид с Мю = 0 и Se = корень( sd1 \* sd1 / n1 + sd2 \* sd2/ n2)

Такое распределение будет иметь вид t-распределение с df = n1 + n2 - 2

Теперь узнаем на сколько наша разность X1 – X2 отклонилась от новой Мю = 0, т.е найдем **t-критерий** и найдем уровень значимости

**T** = (X1 – X2) / Se (сайт все тот же и уровень значимости тот же)

Изображение выглядит как текст, доска

Автоматически созданное описание

# Проверка распределения на нормальность:

Можно построить гистограмму и поверх наложить нормальное распределение или QQplot

# Ошибки:

Ошибка первого рода это, когда мы отклонили гипотезу, а она верная

Ошибка второго рода это, когда мы приняли гипотезу, а она ложная

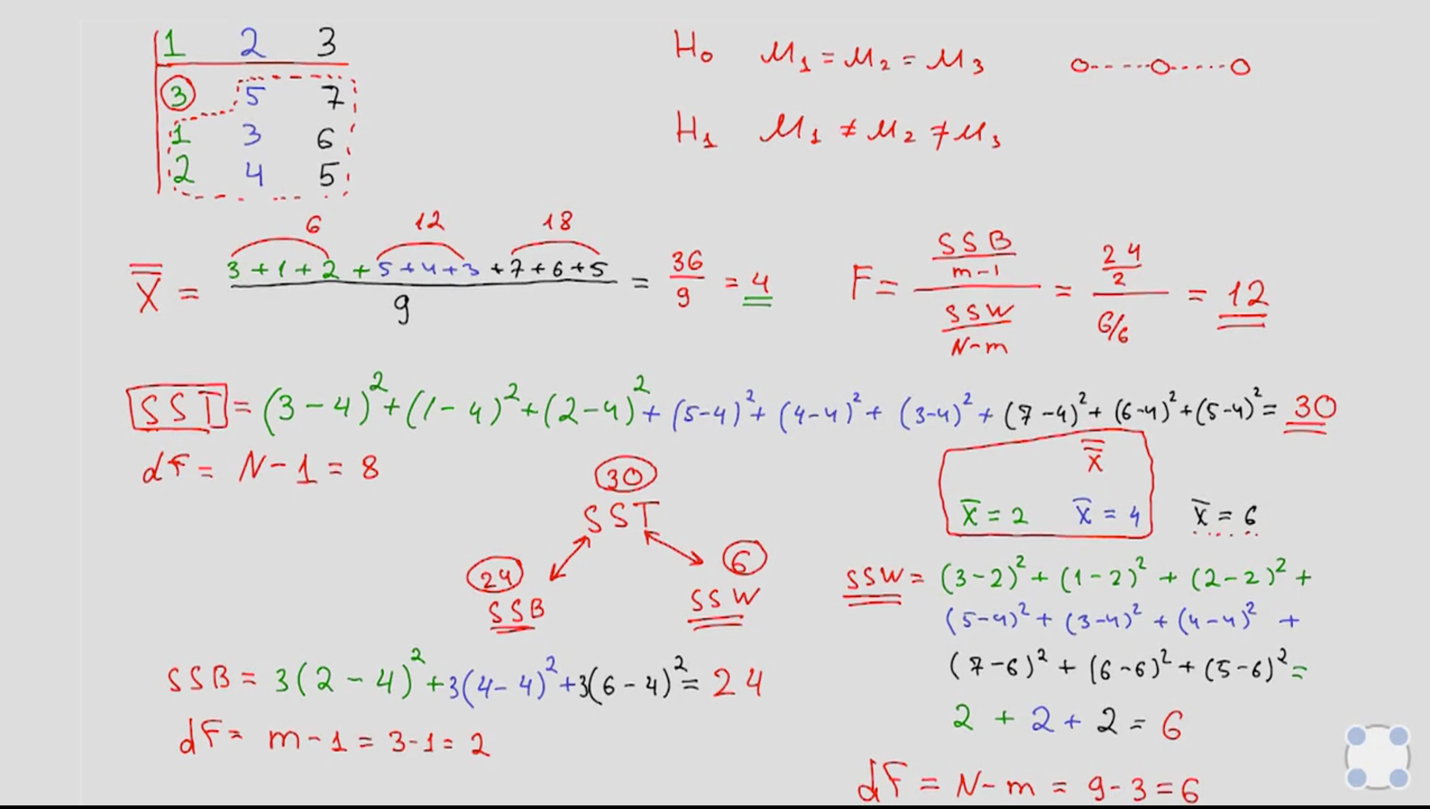
# Однофакторный дисперсионный анализ:

Учитываем взаимосвязь зависимой переменной от одной независимой переменной.

Если есть необходимость сравнить несколько групп между собой, таком случае мы можем применять однофакторный дисперсионный анализ.

**Независимая переменная** - переменная, которая будет разделять наших испытуемых или наблюдения на группы (номинативная переменная с нескольким градациями).(признаки)

**Зависимая переменная -** количественная переменная, по степени выраженности которой мы сравниваем группы.(целевая переменная – Мю)





У нас 3 выборки, надо узнать, схожи они между собой или нет, следовательно стоим 2 гипотезы. – среднее по всей всем выборкам, если бы они были как одна



**Общая сумма квадратов SST –** подобие дисперсии, без учета разделения их на группы

df = 8. Независимых чисел8 ,**ибо зная их** и SST мы **всегда можем найти** **оставшейся** 9-ое.

**SSB** – дисперсия (различия) между группами

**SSW** – сумма дисперсий внутри группы

Если SST в большей части состоит из SSB, то значит, группы значительно различаются между собой

Если SST в большей части состоит из SSW, то значит, группы более схожи и изменчивость внутри групп

**F-значение** – основной показатель дисперсионного анализа, как сильно отклонились

Если бы многократно повторяли бы наш эксперимент, при условии верности нулевой гипотезы, а т.е из одной генеральной совокупности набирали бы 3 выборки, то мы бы в большинстве случаев получали бы Sd схожими между собой числитель дроби F-значения стремился бы к 0 и результат был бы небольшим. Теперь смотрим уровень значимости такого события на том же сайте

Такое распределение называется **F-распределение (Фишера),** имеет явно выраженную асимметрию.

# Множественные сравнения в ANOVA:

Для множественного сравнения выборок надо использовать Бамферони, но лучше критерий Тьюки

Бамферони меняет наш порог альфа = альфа / кол-во парных сравнений

кол-во парных сравнений (n) = (n \* (n – 1)) / 2

критерий Тьюки составляет доверительный интервал из распределения средней разности между двумя выборками и если этот интервал не включает в себя ноль, тогда они различны

# Двухфакторный дисперсионный анализ:

Можем учитывать взаимосвязь нашей зависимой переменной от 2 независимых переменных, т.е влияние нескольких факторов на нашу переменную

Теперь SST = SSW + SSB(a) + SSB(b) + SSB(a) \* SSB(b)

Для дисперсионного анализа нужно

1. Нормальность распределения нашей зависимой переменной в каждой из групп
2. **Гомогенность** дисперсии (что бы она была одинаковая в каждой из групп)

Применение дисперсионного анализа еще не говорит о явной зависимости между ними

Корреляция и Регрессия

# Корреляция:

**Коэффициент корреляции (r)** – на сколько сильна линейная зависимость между двумя признаками, если она полож-ая, то зависимость возрастает, если нет, то убывает.

Желательно что бы распределение было нормальным и без выбросов.

Высокий r еще не говорит о причинах следственной зависимости, но может это делать, если проводить нормальный эксперимент. Может быть так же явление третей переменной, которая и обуславливает возникшую корреляцию. Размер ноги школьника и знания математики, чем выше размер ноги, тем он просто старше

**Коэффициент детерминации** - квадраткоэффициента корреляции**,** показывает, в какой степени дисперсия одной переменной обусловлена “влиянием” другой

Высокая корреляция не обязательно означает статистически значимую взаимосвязь, для это надо смотреть p уровень значимости

# Регрессия с одной независимой переменной:

Применяется для исследования взаимосвязи двух переменных (зависимой и независимой преременной)

Изображение выглядит как текст, карта

Автоматически созданное описание

Для проверки наличия регрессии( взаимосвязи между признаками) можно использовать T-критерий



**Коэффициент детерминации R -** доля дисперсии зависимой переменной(Y), объясняемая регрессионной моделью, насколько хорошо наша модель описывает поведение зависимой переменной, чем ближе к 1, тем лучше



b1 коэф показывает – с каждым единичным изменением только одной независимой переменной, целевая переменная менялась бы на b1

# Условия применения линейной регрессии

1. Линейная зависимость
2. Нормальное распределение остатков (они есть как снизу, так и сверху)
3. Гомоскедастичность – постоянная изменчивость остатков на всех уровнях независимой переменной (ведут ли они себя одинаково на всем регрессионном анализе)
4. Отсутствие Мультиколлинеарность

**Мультиколлинеарность** – очень сильная взаимосвязь (корреляцию) между нашими независимыми переменами или идентичность. Рост в метрах и в дюймах

Независимая переменная, которая хорошо коррелирует между остальными нп, может ухудшать модель

**Сэмплирование** - репрезентативная и меньшая по размеры выборка

**Эмпирическое распределение** - естественное приближение теоретической функции распределения построенной по выборке

На самом деле условий не 4, а 6

1. Линейная зависимость
2. Случайность выборки
3. Случайность ошибок
4. Нормальность ошибок (Нормальное распределение остатков)
5. Гомоскедастичность
6. Отсутствие мультиколлинеарности

# Проверка условий:

1 – Линейная зависимость

в точности не выполняется никогда, будем проверять, нет ли каких-то огромных отклонений от линейности y по x и проверять распределение и нормальность остатков

2 - Случайность выборки

Мы хотим, что выборка наша была независимой и одинаково распределенной.

Это предположение может нарушаться, если выборка отобрана из генеральной совокупности не случайно, а каким-то образом отфильтрована. Фильтровать генеральную совокупность по какому-то признаку z можно только в случае, если z не добавляет никакой новой информации об y.

3 - Случайность ошибок

На очереди предположение о случайности ошибки. Условное математические ожидание ε по x должно быть равно 0. Гипотезу о том, что оно равно 0, можно очень легко проверить по данным. Строите регрессию y на x, считаете остатки, проверяете гипотезу о том, что среднее значение остатков равно 0. Это можно сделать, например, критерием Стьюдента.

4 - Нормальность ошибок

Есть способ визуальный: вы строите ку-ку график и смотрите, лежат ли на этом графике точки более-менее на одной прямой. И есть способ формальный: вы можете использовать статистические критерии для проверки нормальности, среди всего разнообразия критериев рекомендуется использовать критерий Шапиро-Уилка.

5 - Гомоскедастичность

Это предположение можно проверять двумя способами. Первый, нестрогий, — это визуальный анализ.

Вы строите графики зависимости остатков от всех признаков xj и смотрите, выглядят ли точки на этом графике как горизонтальная полоса. Если вместо горизонтальной полосы вы видите что-то расширяющееся или сужающееся, наоборот, значит, гомоскедастичности не выполняется.

Формально это предположение можно проверять с помощью критерия Бройша-Пагана.

6 - Отсутствие мультиколлинеарности

Посмотреть на сильную корреляцию признаков или аналитически