Анализ данных

Лекция 11

Звоновский, к.с.н.

Статистические характеристики

Среднее арифметическое (мат. ожидание) – отношение суммы значений всех элементов совокупности к числу этих элементов

Далеко не всегда среднее арифметическое хорошо описывает переменную. Часто оно скорее затуманивает оценку, чем уточняет ее. В этих случаях используют другие средние статистические характеристики.

Мода – наиболее часто встречающееся значение переменной. Востребовано при анализе нечисловых переменных.

Например, русские составляют в России большинство, или Клинское – наиболее распространенная марка пива.

Статистические характеристики

Медиана - возможное значение признака, которое делит ранжированную совокупность на две равные части: 50 % «нижних» единиц ряда данных будут иметь значение признака не больше, чем медиана, а «верхние» 50 % — значения признака не меньше, чем медиана.

Например, 20 респондентов отмечают уровень своего дохода в рублях. Первые 19 по 10 т.р., а 20-ый – 100 млн р. Среднее будет 5 000 т.р. и оно не указывает на характеристики выборочной совокупности. А медиана, т.е. **значение**, при котором половина ранжированных элементов расположена слева по линии возрастания значений, а половина – справа, будет характеризовать совокупность.

Статистические характеристики

Уровень измерения	Допустимые меры средних
Номинальный	Мода
Порядковый	Мода, медиана
Метрический	Мода, медиана, среднее арифметическое

В случае использования среднего значения для оценки генерального среднего указывается величина дисперсии, стандартного отклонения (СКО) или доверительного интервала.

Доверительный интервал чаще используется, потому что его значения легче интерпретировать и они интуитивно ясны и очевидны.

Стандартизация показателей

Чаще всего значения доверительного интервала имеют различную размерность и для адекватного сравнения со средним требуется знание величины отклонения, выраженного не в единицах измерения переменной, а в степени этого отклонения.

Это позволяет делать сравнения любых величин между собой. Например, степени отклонения дохода мужчин, вступающих в брак, от среднего и степени отклонения роста вступающих в брак женщин от среднего. Или – сравнить степень рассеяния значений доходов американцев и россиян.

Чаще всего используется z-стандартизация, предполагающая нормальный закон распределения выборочных значений.

Интервальное оценивание

Поскольку данные, полученные в результате измерения, носят статистический характер необходимо оценить интервал, в которые с тем или иным значением вероятности показатели выборочной совокупности совпадают с аналогичными в генеральной.

Если полученное значение равно, например, 8,9% и измерение проведено на объеме 1000 единиц с дисперсией генеральной 3,4%, то доверительный интервал будет равен:

$$H=z*\sigma/\sqrt{n}$$

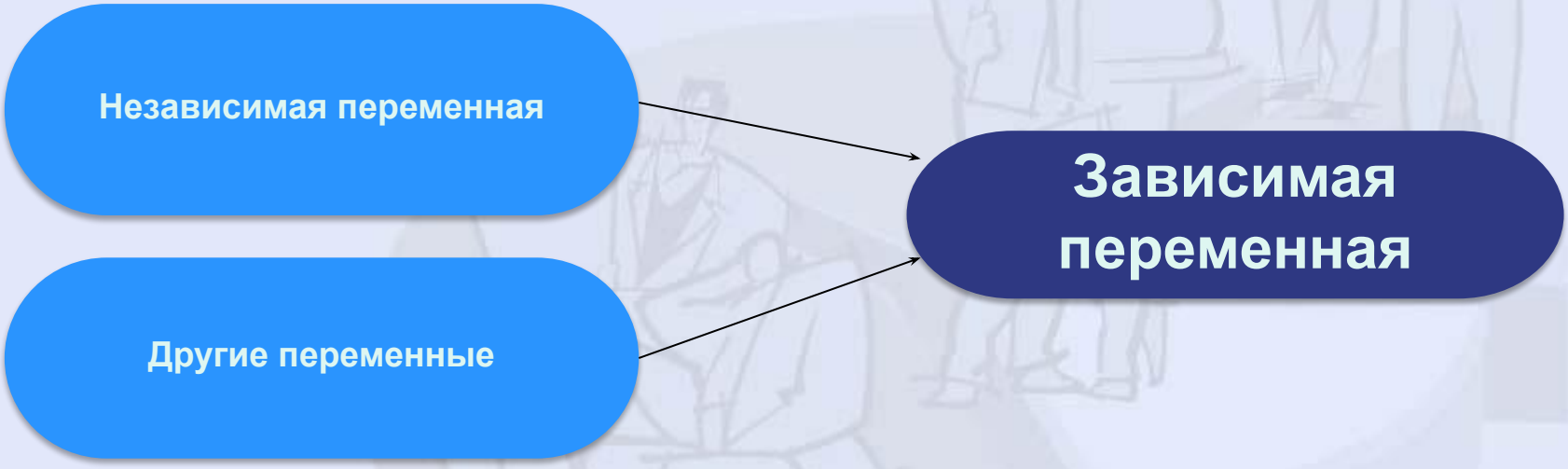
$$2*3,4/\sqrt{1000}=2,1\%$$

Т.е. значение генеральной с вероятностью 95% будет находиться в промежутке от 6,8% до 11,0%

Взаимосвязь переменных

В задачи исследования часто входит не только измерение величин, но выявление и измерение взаимосвязи переменных.

Пример: социальные группы, различающиеся по образованию, предпочитают разные формы досуга.



Независимые – объясняющие переменные
Зависимые – объясняемые переменные

Анализ взаимосвязи переменных

Лекция 12

Звоновский, к.с.н.

Взаимосвязь переменных

Основная задача анализа данных, собранных в результате количественного социологического или маркетингового исследования состоит в **поиске различий** в формах социального поведения и массового сознания между отдельными социальными группами

Отличия в математике идентифицируются как зависимость одной переменной (зависимой) от другой (независимой) и обнаруживаются в теории статистики двумя основными способами.

Анализ таблиц сопряженности и визуальное обнаружение различий между группами

Расчет коэффициентов зависимости (корреляции) между переменными.

Анализ сопряженности

Пол * Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?
Crosstabulation

Count		Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?				Total
		хорошее	удовлетворительное	плохое	затрудняюсь ответить	
Пол	Мужчины	15	242	84	19	360
	Женщины	9	314	78	39	440
Total		24	556	162	58	800

Пол * Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?
Crosstabulation

Count		Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?				Total
		хорошее	удовлетворительное	плохое	затрудняюсь ответить	
Пол	Мужчины	4	67	23	5	100
	Женщины	2	71	18	9	100
Total		3	70	20	7	100

Коэффициенты связи для номинальных переменных

Зависимость – это отсутствие независимости.

Два события считаются независимыми, если вероятность того, что они произойдут одновременно, равна произведению вероятностей того, что произойдет каждое из них.

Вероятность того, что на двух монетах выпадут два орла, равна произведению выпадения орла на каждой из них. 0,5*0,5=0,25.

Если реальная частота выпадения двух орлов будет отличаться от прогнозного, значит, два события зависимы друг от друга.

Коэффициент X^2

Пол * Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?

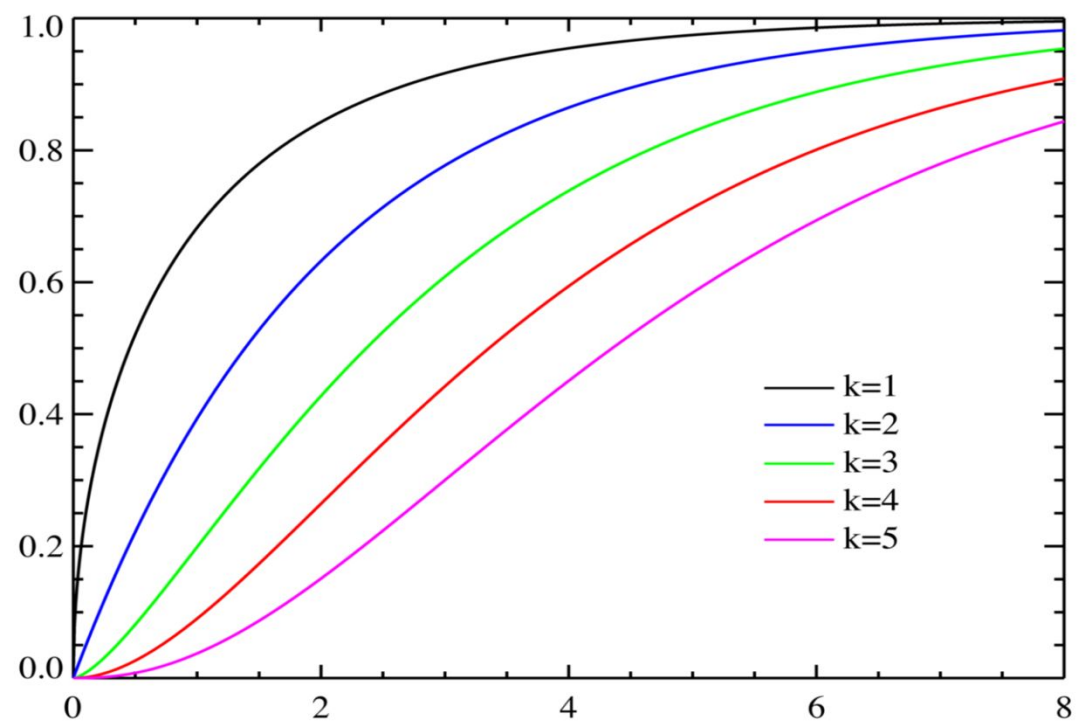
Crosstabulation

Count		Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?				Total
		хорошее	удовлетворительное	плохое	затрудняюсь ответить	
Пол	Мужчины	15	242	84	19	360
		10,8	250,2	72,9	26,1	360
	Женщины	9	314	78	39	440
		13,2	305,8	89,1	31,9	440
	Total	24	556	162	58	800
		24	556	162	58	800

$$\chi^2 = \sum (O - E)^2 / E$$

X^2 - распределение

Число степеней свободы
N=(r-1)*(c-1)



Если значение X равно 5, то вероятность, что наблюдаемые и ожидаемые частоты значимо различаются, равна (при 5 степенях свободы) 0,08.

Ограничения X^2

Коэффициент X^2 будет иметь распределение X^2 лишь в случае, если ожидаемые частоты в таблице имеют значения не меньше 5.

Если таких ячеек в таблице более 20% или в одной ячейке ожидаемая частота меньше 1, расчет не дает надежных результатов.

В любом случае, не следует использовать данную статистику при малых объемах выборки.

Использование двух параметров (собственно значение коэффициента и число степеней свободы) делает невозможным сравнение коэффициентов и затрудняет использование

Коэффициент X^2 по Пирсону

Пирсон уточнил коэффициент X^2

$$c = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Достоинства коэффициент Пирсона:

- растет вместе с X^2
- меняется от 0 до 1

Недостатки коэффициент Пирсона:

- зависит от N
- нельзя сравнивать различные пары переменных между собой

Коэффициент Крамера

Более удобным является коэффициент Крамера

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

K – наименьшее из чисел (r,c), где r - число рядов, c- число столбцов

Достоинства коэффициент Крамера:

- меняется от 0 до 1
- равен 1 лишь в случае полной детеминированности одной переменной другой.

Недостатки коэффициент Крамера:

- зависит от N
- нельзя сравнивать различные пары переменных между собой

Коэффициент, основанные на прогнозе

Эта группа коэффициентов основана на идее Гутмана:

Насколько улучшится наш прогноз ответа случайно взятого респондента на вопрос 1, если мы будем знать его ответы на вопрос 2.

$$\frac{\text{ошибка при первом прогнозе} - \text{ошибка при втором прогнозе}}{\text{ошибка при первом прогнозе}}$$

Первый прогноз – это модальное значение (А) предсказываемой переменной. Вероятность этого прогноза – Pr A. Вероятность ошибки прогноза 1 – Pr A. Ошибка во втором прогнозе будет средней по каждой из строк таблицы сопряженности

$$P2 = \sum (1 - Pr\ A_i) \ / r$$

Прогноз для модального значения предложил в 1941 году Гутман.

Коэффициент, основанные на прогнозе

Пол * Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?

Count		Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?				Total
		хорошее	удовлетвори тельное	плохое	затрудняюсь ответить	
Пол	Мужчины	4	67	23	5	100
	Женщины	2	71	18	9	100
	Total	3	70	20	7	100

$$\frac{\text{ошибка при первом прогнозе} - \text{ошибка при втором прогнозе}}{\text{ошибка при первом прогнозе}}$$

$$\frac{0{,}70 - 242 / 360}{0{,}70}$$

$$\lambda = \frac{\sum (\max n\, ij - max\ n\ *j)}{\max n\ *j}$$

Достоинство λ – очевидный физический смысл – степень улучшения вероятности правильного прогнозирования

Недостаток - равенство коэфициента нулю не указывает на независимость переменных

Коэффициент, основанные на прогнозе

Коэффициент тау (т) Гудмена-Краскала рассчитывает улучшение прогноза не только по модальным значениям, а по всем ячейкам таблицы сопряженности

	Как бы Вы сегодня оценили состояние экономики в нашей области - как хорошее, удовлетворительное или плохое?			
	хорошее	удовлетвори тельное	плохое	затрудняюс ь ответить
Мужчины	4	67	23	5
	15	242	84	19

$$\frac{\text{ошибка при первом прогнозе} - \text{ошибка при втором прогнозе}}{\text{ошибка при первом прогнозе}}$$

$$\tau = \frac{0{,}70 - (360 - (4 * 15\% + 242 * 67\% + 84 * 23\% + 19 * 5\%)) / 360}{0{,}70}$$

Как и в случае λ мы получаем три коэффициента – для случаев зависимой и независимой переменных и среднюю.

Коэффициенты для порядковых шкал

Критерии наличия связи между порядковыми шкалами основаны на количестве нарушений порядка (инверсий).

Коэффициент Гудмена-Краскала представляет собой отношение разности сумм ячеек без нарушения порядка и с нарушением порядка к сумме этих сумм.

$$\gamma = (S - D) / (S + D)$$

Коэффициент может меняться от -1 до 1 и равен нулю, если число инверсий равно числу проверсий, т.е. когда зависимости между двумя переменными нет.

Случай, когда $\gamma = -1$ отражает разнонаправленность двух однозначно связанных переменных.

γ однозначно интерпретируется лишь для монотонных зависимостей.

Коэффициенты для порядковых шкал

В случае, если между переменной A и B, коэффициент Гудмена-Краскала больше, чем между A и C, это значит, что C более чувствительно к изменению A, чем B. При этом, очевидно, что B и C должны быть измерены по одинаковой шкале.

Кроме коэффициент Гудмена-Краскала существуют критерии:

ρ Спирмена

τ Кендэлла

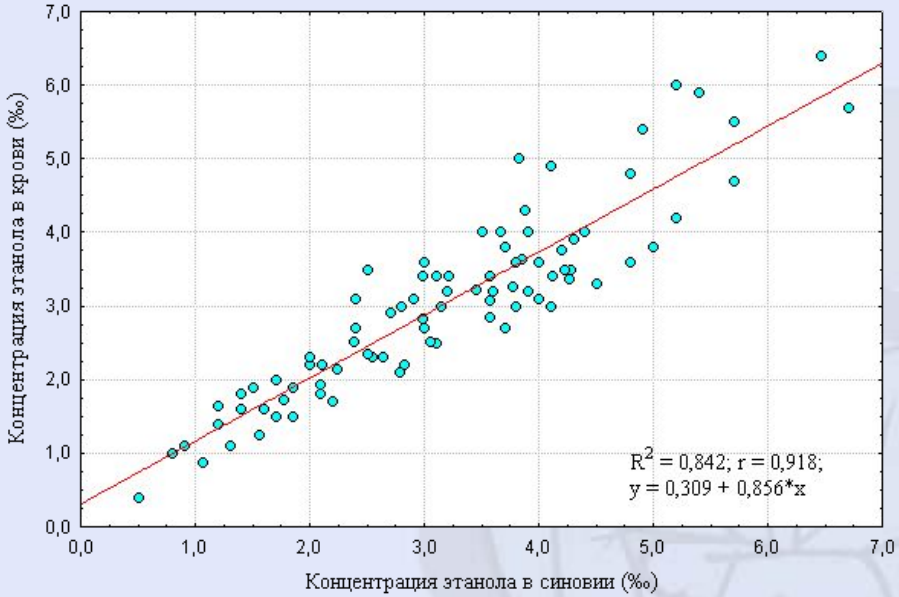
d Соммера

Все они основаны на совпадении/несовпадении изменения порядка переменных.

Коэффициенты для метрических шкал

В случае, когда обе переменные измерены по метрическим шкалам, появляется возможность использовать коэффициент Пирсона.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)s_x s_y}$$



Коэффициент Пирсона меняется от -1 до 1.
Значение 1 говорит о полной и положительной детерминации одной переменной другой
Значение -1 указывает на полную и отрицательную детерминацию одной переменной другой.
Значение 0 говорит об отсутствии линейной зависимости

Коэффициенты Пирсона для метрических переменных можно сравнивать друг с другом.
Коэффициент r имеет доверительный интервал и уровень значимости.

Сравнение средних

Лекция 13

Звоновский, к.с.н.

Анализ средних

Задача сравнения средних значений (means) возникает в случаях, когда необходимо убедиться в том, что различия между значениями какого-либо метрической переменной в двух или более группах статистически значимы.

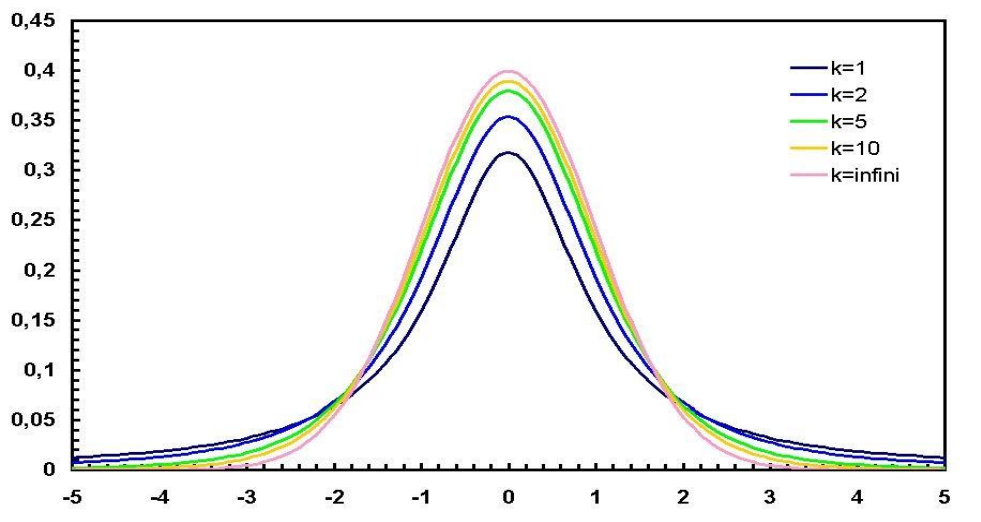
Например, мы хотим проверить гипотезу о том, что доходы избирателей партии Единая Россия ниже доходов избирателей ЛДПР. Для этого мы сравниваем среднее значение переменной «доход» в двух группах, сформированных по номинальной шкале «партийные предпочтения».

Анализ средних

Для определения значимости различий средних значений в группах необходимо провести проверку t-тест, или тест Стьюдента.

Существуют три вида t-теста: для двух независимых выборок, для одной выборки, для парных (связанных) данных.

Все t-тесты основаны на t-распределении. Оно приближается к нормальному с ростом степеней свободы (объемом выборки) при больших (n > 30). И практически не отличается от нормального при n > 100.



Предполагается, что распределение генеральной нормальное или близкое к нему, а используемая шкала, – как минимум, интервальная.

Анализ средних

Случай двух независимых выборок. Одна выборка не зависит от другой в том случае, если попадание элемента в одну из них не влияет на вероятность попасть в другую

Примеры.

- Средний доход жителей Самары выше среднего дохода жителей Тольятти.
- Молодежь выше оценивает сервис сотового оператора, чем старшее поколение
- Число посещений кинотеатров выше у студентов гуманитарных вузов, чем у студентов технических.

Анализ средних

Случай одной выборки. Необходимо сравнить выборочное значение параметра с каким-то либо внешним параметром, чаще всего, произвольно выбранным.

Примеры.

- Проверить утверждение о том, что число поездок на границу для данной социальной группы - 2,0 .
- Оценка функций нового принтера не меньше 7,5 баллов по 10-балльной шкале.

Анализ средних

Случай двух зависимых выборок. Зависимые выборки – выборки, где выпадение элементов одной выборки влияет на выпадение элементов (формирует) другой.

Примеры.

- Необходимо сравнить оценки двух ресторанов одной и той же группой.
- При покупке автомашины люди цену считают более важной характеристикой марки, чем ее имидж.
- Дети больше тратят денег на мороженное, чем на пиццу.

Анализ средних

- Сформулировать H_0 и H,
- Выбрать подходящую статистику
- Выбрать уровень значимости
- Собрать данные и рассчитать проверочную статистику
- Определить вероятность выбранной статистики и сравнить с выбранным значением значимости
- Отклонить или принять H_0
- Сделать вывод и принять решение

Непараметрические тесты

Все рассмотренные выше случаи касались параметрических тестов.

Параметрические тесты – тесты, основанные на допущении, что выборочная совокупность подчиняется нормальному закону распределения.

Непараметрические тесты – тесты, не требующие какого-то конкретного закона распределения выборочной совокупности.

Параметрические тесты работают лишь с метрическими шкалами и чувствительны к выбросам.

Непараметрические тесты, поскольку обрабатывают не само значение, а его ранг, позволяют работать и с порядковыми переменными.

Непараметрические тесты

Виды теста	Тесты
Случай одной выборки	Тест Колмогорова-Смирнова
Случай двух независимых выборок	Тест Манна-Уитни Тест Вальда-Вулфовитца
Случай к независимых выборок	Тест Краскала-Уоллиса Медианный тест
Случай двух зависимых выборок	Тест Уилкоксона Тест на знак
Случай к зависимых выборок	Тест Кендалла Тест Кохрана

Непараметрические тесты

Для определения, какому закону распределения подчиняется данная переменная, используется **тест Колмогорова-Смирнова.**

Чаще всего, тест Колмогорова-Смирнова используется для доказательства нормального распределения.

Биноминальный тест – тест на значимость различия среднего значения в двух подвыборках, составляющих вместе общую выборку.

Например, насколько значимо отличается число побед у двух команд.

Поиск последовательности – тест на поиск закономерностей в последовательности.

Обнаруживает наличие закономерности в последовательности дихотомических значений по сравнению со случайной последовательностью.

Дисперсионный анализ

Лекция 14

Звоновский, к.с.н.

Дисперсионный анализ (ANOVA)

В случаях, когда необходимо сравнить не одну, а несколько средних, используют анализ вариаций, или, **дисперсионный анализ.**

Гипотезы, которые требуют использования дисперсионного анализа:

- Сегменты рынка отличаются по объему потребления товара (нулевая гипотеза – отличий нет)
- Оценки товара среди групп, просмотревших различную рекламу этого товара, отличается (нулевая гипотеза – различий нет)
- Число прочитанных статей на политическую тему среди сторонников различных кандидатов отличаются (нулевая гипотеза – не отличаются)

Одномерный дисперсионный анализ

В дисперсионном анализе важно различать **зависимые** и **независимые** переменные.

Независимая – переменная, которая оказывает влияние на значения зависимой. Иногда такие переменные называют **факторами**, ее значения - **уровнями.** Таких переменных может быть несколько. Если она одна, то такой анализ будет **одномерным (one-way)**.

Например, зависимость числа походов студентов в кинотеатр от вуза, где они обучаются.

Если несколько переменных будет признано независимыми, то такой анализ будет многомерным.

Все эти переменные должны быть или **номинальными** (чаще всего), или порядковыми.

Зависимая – переменная, значение которой измеряется в зависимости от значений независимых(ой). Она должна быть **метрической**.

Одномерный дисперсионный анализ

Основная идея анализа состоит в разделении дисперсии на дисперсию, вносимую независимыми переменными (межгрупповую) и дисперсию (внутригрупповую), вносимую из-за ошибки, учитывающую влияние других факторов, сведенных к случайному воздействию.

Пример. Студенты СГЭУ ходят в кино чаще, чем студенты СГАУ.

Два допущения. 1. Мы предполагаем, что различия существуют (не все μ равны между собой). 2. Мы предполагаем, что на число походов в кино влияет не только вуз, где проходит обучение, но и доход в семье, образование родителей и другие факторы.

Тогда можно утверждать, что дисперсия (степень различия) числа походов в кино будет составляться из дисперсии между группами студентов различных вузов (S1) и из дисперсии внутри каждой из этих групп (S2).

S^2=$S1^2$ (межгрупповая) + $S2^2$ (внутригрупповая)

Одномерный дисперсионный анализ

Дисперсионный анализ разделяет дисперсию на дисперсию, вносимую независимыми переменными (межгрупповую) и дисперсию (внутригрупповую), вносимую из-за ошибки, учитывающую влияние других факторов, сведенных к случайному воздействию.

Пример. На рынок выводится новый бренд шампуня.

Наша гипотеза: Оценка данного бренда по шкале Лайкерта отличается в группах постоянных (hard), периодических (medium), случайных (light) пользователей и непользователей. Нулевая гипотеза состоит в утверждении, что отличий в этих группах нет.

Одномерный дисперсионный анализ

x_{ij} - оценка *i-ого* респондента из j-ой группы

μ - средняя оценка по всему массиву данных

μ_j - средняя оценка по j-ой подгруппе респондентов

Тогда оценка *i-ого* респондента из j-ой группы –

$$x_{ij} = \mu + (x_{ij} - \mu_j) - (\mu_j - \mu)$$

межгрупповая **внутригрупповая**

Межгрупповая и внутригрупповая дисперсии вычисляются с помощью статистики Фишера, распределение которой зависит от числа степеней свободы каждой из переменных.

Одномерный дисперсионный анализ

В результате теста на статистически значимые различия зависимой переменной мы получаем доказательство лишь существования таких различий. Но не можем сказать между какими именно группами есть такие различия.

Для этого используются **тесты множественных сравнений**. Они указывают на отличия или сходства значений зависимой переменной в группах всех значений независимой переменной попарно.

Данные тесты бывают с предполагаемым равенством дисперсии в группах и непредполагаемым их равенством. В общем случае рекомендуется использовать критерии с непредполагаемым равенством дисперсий.

Тест Краскала-Уоллиса

Рассмотренные виды анализа имеют два существенных недостатка.

Во-первых, они предполагают, что зависимая переменная распределена нормально. Однако, часто мы не можем утверждать этого. В других случаях мы точно знаем, что она не имеет нормального распределения

Во-вторых, дисперсионный анализ применяется лишь в случае, когда зависимая переменная – метрическая. А часто хотелось бы понять, есть ли значимые различия между двумя порядковыми переменными или даже номинальными.

В этом случае используются **тест Краскала-Уоллиса.** Данный тест применяется при измерении (зависимой) переменной по порядковой шкале. Значения такой переменной упорядочиваются от минимального до максимального (от несогласия к согласию) и значениям приписываются ранги, которые принимаются за количественные значения, равномерно распределенные по шкале.