

# Вариационный ряд, таблицы сопряженности признаков и проверка гипотез

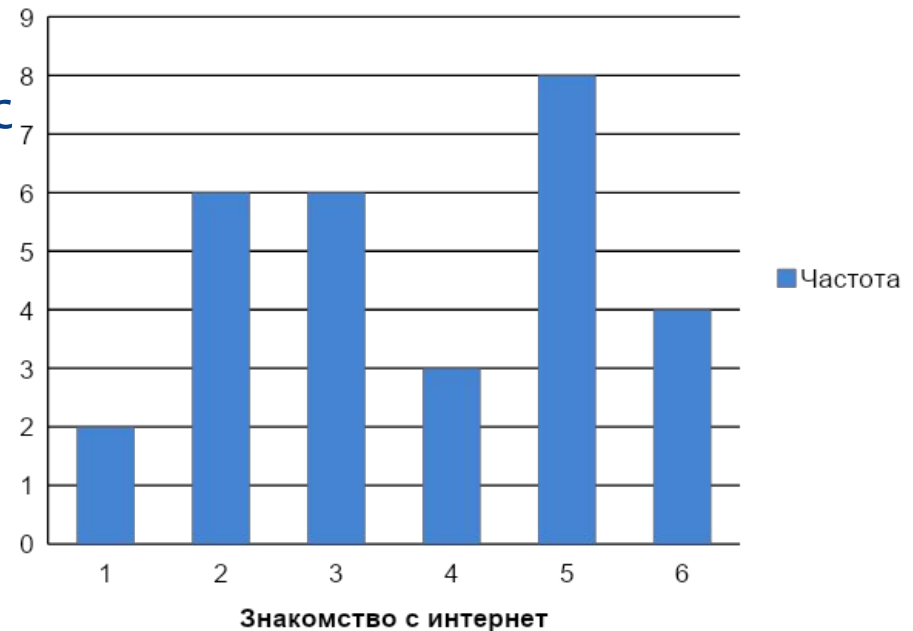
*Подготовлено А.В. Омеляненко*

# Вопросы к экзамену:

- \* Вариационный ряд. Статистики, связанные с распределением частот. Общая схема проверки гипотезы
- \* Построение таблиц сопряженности признаков. Статистики сопряженности признаков. Практика построения таблиц сопряженности признаков.
- \* Проверка гипотез: различия между значениями переменных. Параметрические критерии. Непараметрические методы проверки гипотез.

# ВАРИАЦИОННЫЙ РЯД

- \* Вариационный ряд, распределение частот значений переменной (frequency distribution)
- \* Математическое распределение, цель которого — подсчет ответов, связанных с различными значениями одной переменной (частот), и дальнейшее выражение их в процентном виде (частоты).



# ВАРИАЦИОННЫЙ РЯД

<i>Отметка значения</i>	<i>Значение</i>	<i>Частоты (N)</i>	<i>Частоты</i>	<i>Достоверные частоты</i>	<i>Накопленные частоты</i>
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
Плохо осведомлены	1	0	0,0	0,0	0,0
	2	2	6,7	6,9	6,9
	3	6	20,0	20,7	27,6
	4	6	20,0	20,7	48,3
	5	3	10,0	10,3	58,6
	6	8	26,7	27,6	86,2
Хорошо осведомлены	7	4	13,3	13,8	100,0
	9	1	3,3	Пропуск	
	Итого	30	100,0	100,0	

- \* **Цель** построения вариационного ряда — подсчет ответов респондентов, в которых приводятся различные значения переменной.
- \* Относительную частоту различных значений переменной выражают в процентах и называют **частотями**.
- \* Подсчет распределения частот значений переменной дает возможность построить таблицу, с указанием частоты, частости и накопленных частостей для всех значений этой переменной.

# ВАРИАЦИОННЫЙ РЯД

Позволяет получить ответы на вопросы:

- \* Какое количество потребителей товара определенной марки можно считать лояльными по отношению к ней?
- \* Каково соотношение между разными группами потребителей того или иного товара или услуги (например, авиаперелетов): много летающие, средне, мало и совсем не летающие?
- \* Какое количество потребителей хорошо осведомлены о предлагаемом новом товаре? Сколько потребителей поверхностно знакомы, сколько — что-то слышали, а сколько вообще ничего не знают о данной торговой марке? Какова средняя степень осведомленности о товаре? Сильно ли различается степень осведомленности потребителей о новом товаре?
- \* Что представляет собой кривая распределения дохода для приверженцев данной марки товара? Смещено ли распределение доходов в сторону группы потребителей с низкими доходами?

# ВАРИАЦИОННЫЙ РЯД

- \* помогает определить долю неответивших респондентов, а также указывает долю ошибочных ответов;
- \* можно установить наличие выбросов, т.е. случаев с экстремальными значениями;
- \* Распределение частот также определяет форму эмпирического распределения значений переменной.
- \* Частотные данные можно использовать для построения *гистограмм* или вертикальных столбчатых диаграмм, на которых по оси  $X$  откладывают значения переменной, а по оси  $Y$ — абсолютные (частоты) или относительные (частости) значения.
- \* По гистограмме можно проверить, соответствует ли наблюдаемое распределение предполагаемому маркетологом распределению значений переменной.

# СТАТИСТИКИ, СВЯЗАННЫЕ С РАСПРЕДЕЛЕНИЕМ ЧАСТОТ

## Показатели центра распределения (measures of location)

- \* Статистики, которые характеризуют значение признака, вокруг которого концентрируются наблюдения, или, как говорят, показывают центральную тенденцию распределения.

## Показатели вариации (изменчивости) (measures of variability)

- \* Статистики, показывающие меру разброса (вариабельность) значений переменной.

## Показатели формы распределения

# Показатели центра распределения

## Среднее арифметическое, выборочное среднее (mean)

- \* Эта величина получается делением суммы всех имеющихся значений переменной на количество значений

$$\bar{X} = \sum_{i=1}^n X_i / n,$$

## Мода (mode)

- \* Значение переменной, которое чаще всего встречается в выборочном распределении.

## Медиана (median)

- \* Значение переменной, которое приходится на середину распределения частот, т.е. одна половина всех значений больше медианы, а другая — меньше



# Показатели вариации (изменчивости)

## Размах вариации (range)

- \* Разность между наибольшим и наименьшим значениями переменной в вариационном ряду

## Межквартильный размах (interquartile range)

- \* Размах вариации распределения, охватывающий центральные 50% всех наблюдений

## Дисперсия (variance)

- \* Среднее из квадратов отклонений переменной от ее средней величины

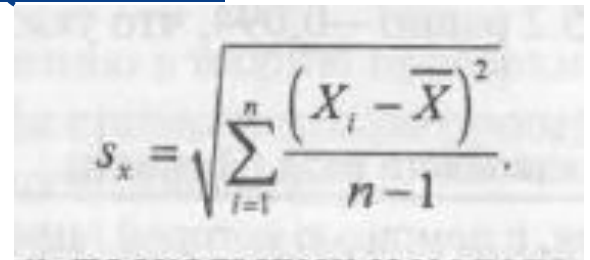
## Среднеквадратичное (стандартное) отклонение (standard deviation)

- \* Корень квадратный из значения дисперсии

## Коэффициент вариации (coefficient of variation)

- \* Величина относительной изменчивости переменной, представляющая собой отношение ее стандартного отклонения к ее среднему значению

$$CV = s_x / \bar{X}.$$


$$s_x = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}}$$

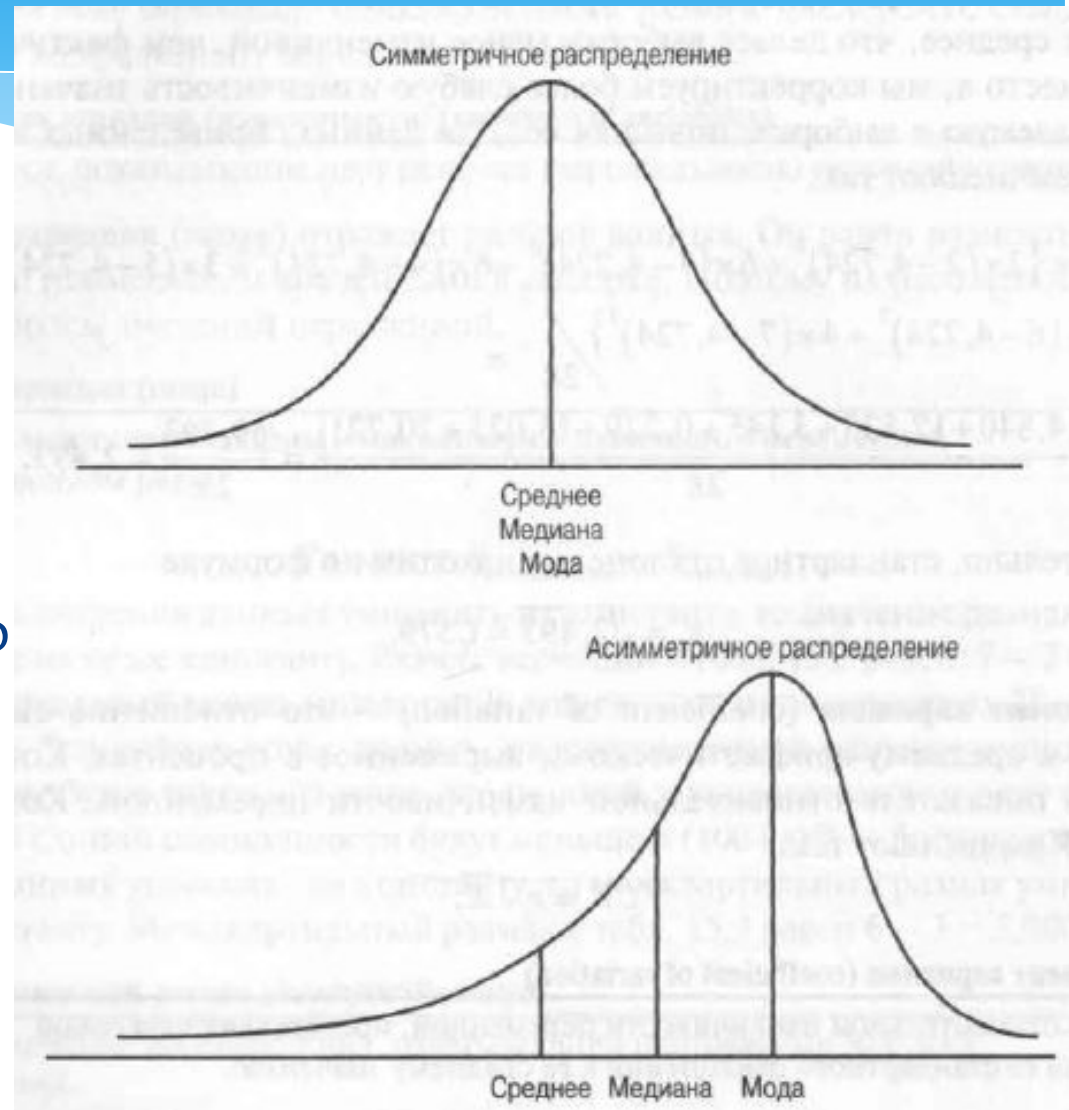
# Показатели формы распределения

## Асимметрия, скошенность вариационного ряда (skewness)

- \* Характеристика распределения, с помощью которой оценивается симметрия расположения значений переменной относительно средней.

## Экссесс (kurtosis)

- \* Мера относительной крутости кривой распределения



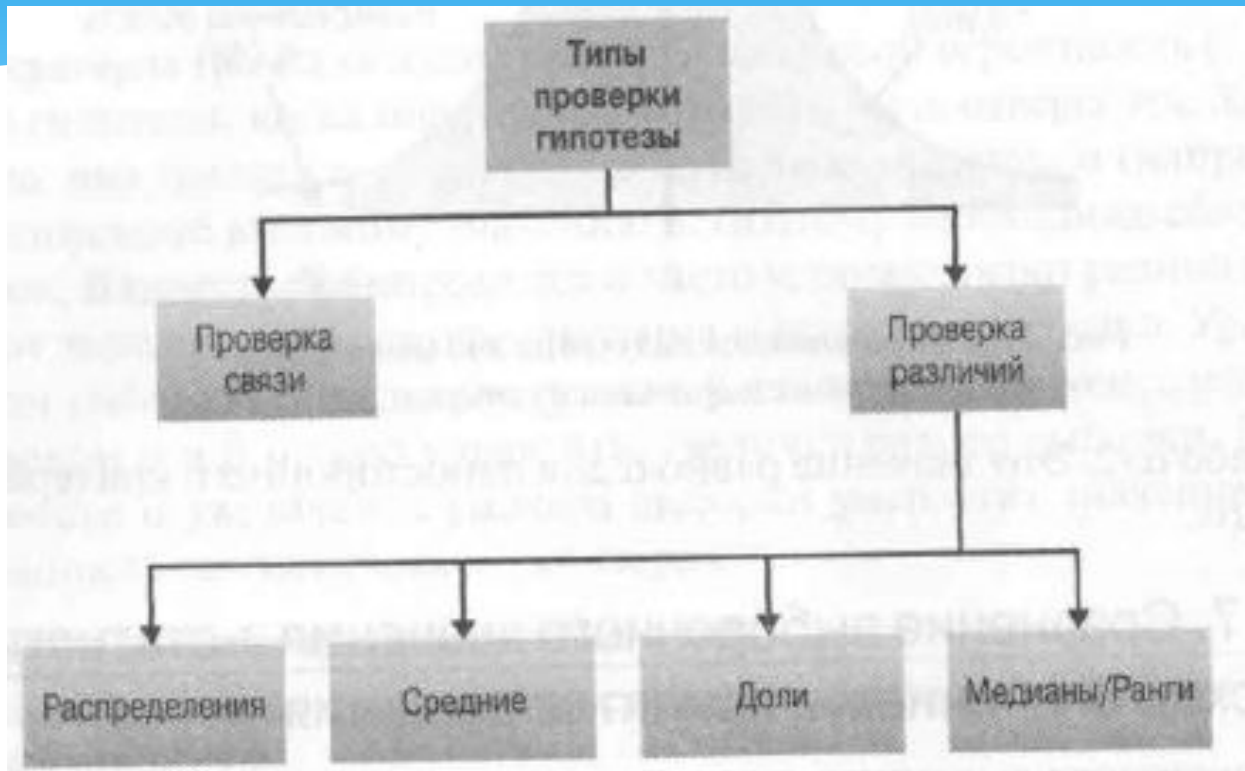
# ОБЩАЯ СХЕМА ПРОВЕРКИ ГИПОТЕЗЫ

Для проверки гипотезы необходимо выполнить следующие этапы:

- Сформулировать нулевую гипотезу  $H_0$  и альтернативную гипотезу  $H_2$
- Выбрать подходящий метод статистической проверки гипотезы (статистический критерий) и соответствующую статистику критерия (выборочную статистику, тест-статистику).
- Выбрать уровень значимости  $\alpha$ .
- Определить размер выборки и собрать данные. Вычислить значение выборочной статистики.
- Определить вероятность статистики критерия при выполнении нулевой гипотезы. Альтернативный вариант данного этапа: определить критическое значение статистики, которое делит интервал на область принятия и непринятия нулевой гипотезы.
- Сравнить полученную вероятность для тест-статистики с заданным уровнем значимости. Альтернативный вариант данного этапа: определить, попадает ли выборочное значение тест-статистики в область принятия или отклонения нулевой гипотезы.
- Принять статистическое решение, касающееся того, принять или отвергнуть нулевую гипотезу.
- Выразить статистическое решение с точки зрения проблемы маркетингового исследования



# КЛАССИФИКАЦИЯ ТИПОВ ПРОВЕРКИ ГИПОТЕЗЫ



На основании результатов проверки статистической гипотезы следует сделать заключение о стоящей перед нами проблеме маркетингового исследования.

Проверка различий может относиться к распределениям, средним, долям, медианам или рангам. Для проверки связей между переменными необходимо построить таблицы сопряженности признаков или провести кросс-табуляцию.

# СХЕМА ПРОВЕРКИ ГИПОТЕЗЫ

Характеристика этапов:

- (изучается факультативно)

# ПОСТРОЕНИЕ ТАБЛИЦ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

## Построение таблиц сопряженности признаков, кросс-табуляция (cross-tabulation)

Статистический метод, который одновременно характеризует значения двух или больше переменных и заключается в создании таблиц сопряженности признаков, отражающих совместное распределение двух или больше переменных с ограниченным количеством категорий или определенными значениями

Кросс-табуляция представляет собой процесс объединения распределений частот значений двух или больше переменных в одну таблицу. Она объясняет, как одна переменная, например лояльность по отношению к торговой марке, связана с другой переменной, такой как пол.

# ПРИЧИНЫ ПОСТРОЕНИЯ ТАБЛИЦ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

- \* менеджеры, которые недостаточно владеют статистическими методами, легко интерпретируют и понимают процедуру кросс-табуляции и ее результаты;
- \* очевидность трактовки результатов анализа ясно свидетельствует о возможных управленческих действиях;
- \* ряд операций кросс-табуляции позволяет лучше понять сложное явление, чем это сделал бы один многовариантный анализ;
- \* кросс-табуляция облегчает проблему разбросанных ячеек, которая затрудняет дискретный многовариантный анализ;
- \* анализ методом кросс-табуляции прост для выполнения и поэтому охотно используется исследователями, менее искушенными в вопросах математической статистики

# Таблицы сопряженности признаков

## Таблицы сопряженности признаков (contingency tables)

Таблица кросс-табуляции состоит из ячеек, в которых приведены комбинации категорий двух переменных.

Наиболее часто употребляются таблиц сопряженности для

- \* **двух** и
  - \* **трех**
- переменных

Можно построить таблицу сопряженности больше, чем для трех переменных, но интерпретация полученных результатов становится достаточно сложной. Кроме того, проблематично оставить необходимое количество респондентов или случаев в каждой ячейке (обычно, чтобы вычислить статистику в каждой ячейке, должно быть не менее 5 наблюдений).

Кросс-табуляция — неэффективный способ проверки связей для ситуаций с несколькими переменными, она рассматривает просто связь между переменными, а не *причинность*. Чтобы изучить причинно-следственную связь, необходимо провести соответствующее причинно-следственное исследование



# Кросс-табуляция с двумя переменными

Использование Internet в зависимости от пола (%)

<i>Использование Internet</i>	<i>Пол</i>	
	<i>Мужчины</i>	<i>Женщины</i>
Мало	33,3	66,7
Много	66,7	33,3
Итого	100,0	100,0

- \* Общее правило, которое необходимо соблюдать, гласит: проценты необходимо вычислять для каждой категории независимой переменной (так, чтобы суммарное значение категорий зависимой переменной применительно к каждой категории независимой переменной давало 100%).

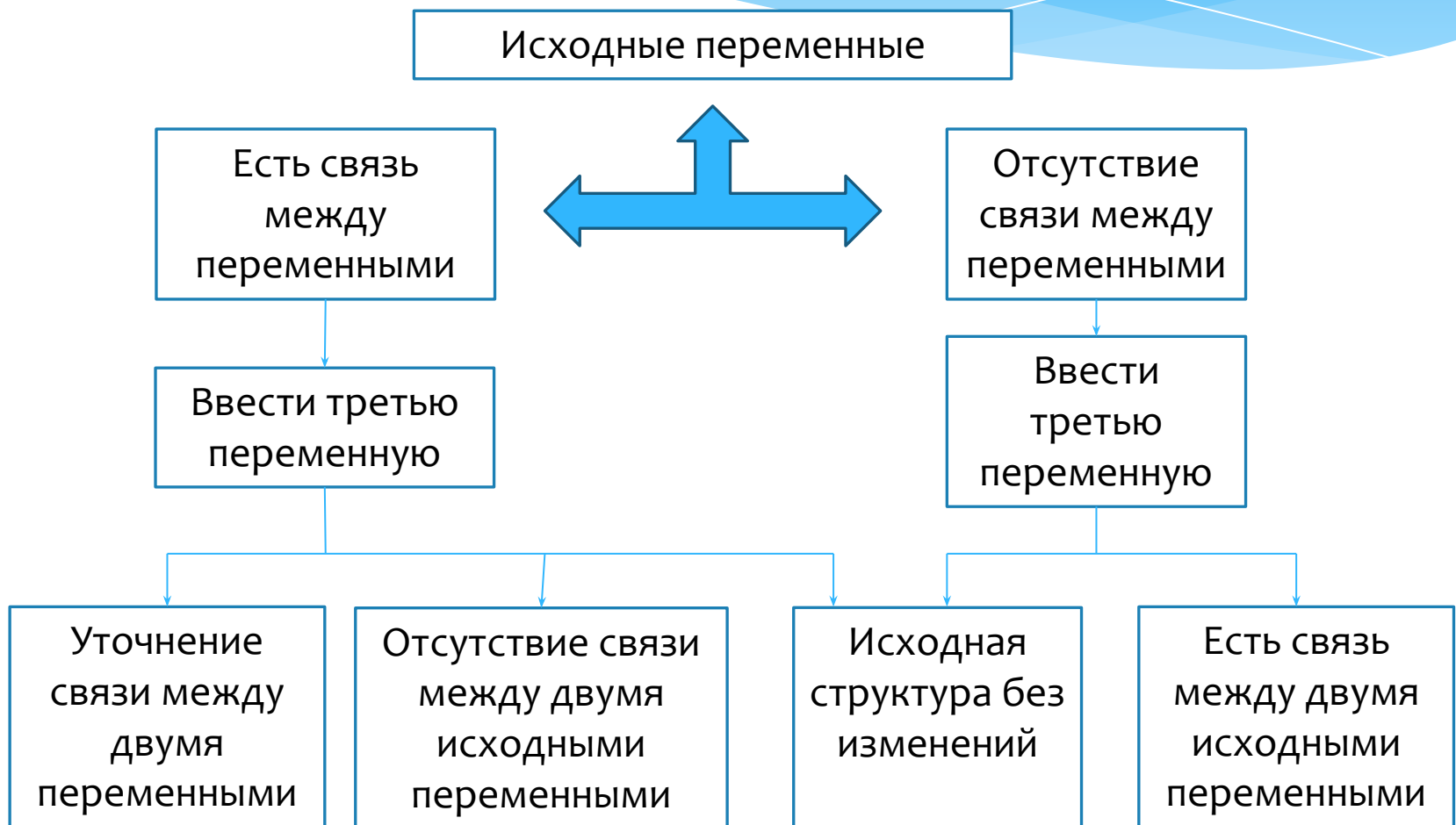
# Кросс-табуляция с тремя переменными

Введение третьей переменной часто позволяет маркетологу четче уяснить природу исходной связи между двумя переменными.

Третья переменная может предоставить четыре возможности:

- \* Уточнить связь, наблюдаемую между двумя исходными переменными.
- \* Указать на отсутствие связи между двумя переменными, хотя первоначально связь наблюдалась. Другими словами, третья переменная покажет, что исходная связь между двумя переменными была ложной.
- \* Показать некоторую связь между двумя переменными, хотя первоначально она не наблюдалась. В этом случае третья переменная показывает скрытую связь между первыми двумя переменными.
- \* Не показать никаких изменений в первоначальной связи

# Кросс-табуляция с тремя переменными



# Кросс-табуляция с тремя переменными – связь уточнена

Покупка модной одежды в зависимости от семейного положения

Покупка модной одежды	Семейное положение	
	Женат (замужем)	Не женат (не замужем)
Много	31%	52%
Мало	69%	48%
Итого	100%	100%
Количество респондентов	700	300

Третья переменная уточнила связь между семейным положением и покупкой модной одежды. Вероятность попадания в категорию покупателей, приобретающих много модной одежды, выше для несемейных респондентов по сравнению с семейными, причем она выше для женщин.

Связь покупки модной одежды с семейным положением

Покупка модной одежды	Пол			
	Мужской		Женский	
	Женат	Не женат	Замужем	Не замужем
Много	35%	40%	25%	60%
Мало	65%	60%	75%	40%
Итого	100%	100%	100%	100%
Количество респондентов	400	120	300	180

# Кросс-табуляция с тремя переменными – связь ложна

**Зависимость владения дорогими автомобилями от образовательного уровня**

<i>Наличие дорогого автомобиля</i>	<i>Образование</i>	
	<i>Выпускник колледжа</i>	<i>Без степени колледжа</i>
Да	32%	21%
Нет	68%	79%
Итого	100%	100%
Количество респондентов	250	750

Если данные по группам с высокими и низкими доходами проверить отдельно, то связь между образованием и наличием дорогого автомобиля исчезает, а это значит, что первоначально наблюдаемая связь между этими двумя переменными была ложной.

**Влияние образования и уровня дохода на наличие дорогого автомобиля**

<i>Наличие дорогого автомобиля</i>	<i>Доход</i>			
	<i>Низкий</i>		<i>Высокий</i>	
	<i>Образование</i>		<i>Образование</i>	
	<i>Колледж</i>	<i>Нет</i>	<i>Колледж</i>	<i>Нет</i>
Да	20%	20%	40%	40%
Нет	80%	80%	60%	60%
Итого	100%	100%	100%	100%
Количество респондентов	100	700	150	50

# Кросс-табуляция с тремя переменными – обнаружена связь

**Зависимость желания совершить туристическую поездку за границу от возраста**

<i>Желание совершить туристическую поездку за границу</i>	<i>Возраст</i>	
	<i>До 45 лет</i>	<i>45 и старше</i>
Да	50%	50%
Нет	50%	50%
Итого	100%	100%
Количество респондентов	500	500

Т.к. связь между желанием путешествовать и возрастом различна для мужчин и женщин и имеет противоположную направленность, то связь между этими двумя переменными была скрыта, пока данные не учитывали переменную “пол”.

**Связь между желанием совершить туристическую поездку за границу, возрастом и полом**

<i>Желание совершить путешествие за границу</i>	<i>Пол</i>			
	<i>Мужчины</i>		<i>Женщины</i>	
	<i>До 45 лет</i>	<i>45 и старше</i>	<i>До 45 лет</i>	<i>45 и старше</i>
Да	60%	40%	35%	65%
Нет	40%	60%	65%	35%
Итого	100%	100%	100%	100%
Количество респондентов	300	300	200	200

# Кросс-табуляция с тремя переменными – связь без изменений

Связь частоты посещения ресторанов быстрого питания с размером

семьи	Размер семьи	
	Небольшая	Большая
Частота посещения ресторана быстрого питания		
Большая	65%	65%
Маленькая	35%	35%
Итого	100%	100%
Количество респондентов	500	500

При введении в анализ третьей переменной представление о связи не изменилось. Никакой связи не наблюдалось, как и при анализе двух переменных

Связь частоты посещения ресторанов быстрого питания с размером семьи и уровнем дохода

Частота посещения ресторанов быстрого питания	Доход			
	Семьи с низким доходом		Семьи с высоким доходом	
	Малая	Большая	Малая	Большая
Большая	65%	65%	65%	65%
Маленькая	35%	35%	35%	35%
Итого	100%	100%	100%	100%
Количество респондентов	250	250	250	250

# СТАТИСТИКИ ТАБЛИЦ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

- \* Критерий хи-квадрат
- \* Фи-коэффициент
- \* Коэффициент сопряженности признаков
- \* V -коэффициент Крамера
- \* Коэффициент “лямбда”
- \* Другие статистики (tau b; tau c; gamma...)

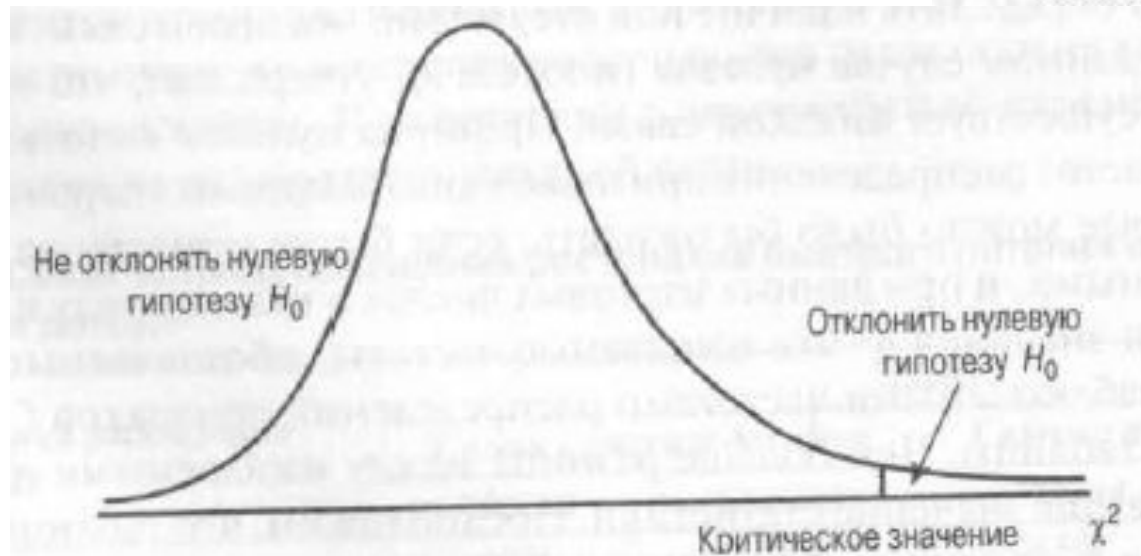


# Критерий хи-квадрат

## Критерий хи-квадрат (chi-square statistic, $\chi^2$ )

Критерий, используемый для проверки статистической значимости наблюдаемых связей в таблицах сопряженности признаков. Он помогает определить наличие или отсутствие значимой связи между двумя переменными.

$$\chi^2 = \sum_{\text{Все ячейки}} \frac{(f_0 - f_e)^2}{f_e}$$



Проверка зависимости между переменными с помощью критерия  $\chi^2$

# Критерий хи-квадрат

Он помогает определить наличие или отсутствие значимой связи между двумя переменными. .

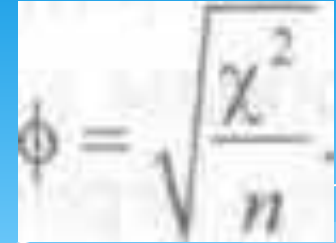
Проверка нулевой гипотезы выполняется вычислением частот распределения признаков анализируемых переменных в ячейках таблицы, которые можно было бы ожидать, если бы не существовало зависимости между переменными, и при данных итоговых числах в каждом ряду и колонке.

Ожидаемые частоты, обозначаемые  $f_e$ , сравнивают с фактически наблюдаемыми частотами распределения признаков.

Чем больше разница между ожидаемыми и фактическими частотами, тем выше значение статистики.

Чтобы определить, существует ли между переменными систематическая связь, определяют вероятность получения значения  $\chi^2$ , равного или большего, чем рассчитанное из таблицы сопряженности.

# Фи-коэффициент


$$\phi = \sqrt{\frac{\chi^2}{n}}$$

## Фи-коэффициент (phi coefficient, $\phi$ )

Мера тесноты связи переменных для конкретного вида таблицы: с двумя строками и двумя колонками (таблица 2 x 2).

Фи-квадрат принимает значение, равное 0, если связь отсутствует, на что также указывает и значение хи-квадрат, равное 0.

При сильной связи между переменными фи- коэффициент имеет значение 1 и все наблюдения находятся на главной или второй степенной диагонали.

(В некоторых компьютерных программах фи-коэффициент принимает значение  $-1$ , а не  $+1$ , когда наблюдается отрицательная связь.)

# V-коэффициент Крамера

$$V = \sqrt{\frac{\phi^2}{\min(r-1), (c-1)}} \quad \text{или} \quad V = \sqrt{\frac{\chi^2 / n}{\min(r-1), (c-1)}}$$

## V-коэффициент Крамера (Cramer's V)

Мера тесноты связи, используемая в таблицах, больших по размеру, чем 2x2

V-коэффициент Крамера получают корректировкой фи-коэффициента по количеству или строк, или колонок в таблице. Причем из двух значений выбирают меньшее.

Корректировку осуществляют так, что значения V-коэффициента лежат в диапазоне от 0 до 1. Большее значение V-коэффициента указывает на более сильную связь, но не указывает, как связаны переменные.

# Коэффициент “лямбда”

$$V = \sqrt{\frac{\phi^2}{\min(r-1), (c-1)}} \quad \text{или} \quad V = \sqrt{\frac{\chi^2 / n}{\min(r-1), (c-1)}}$$

## Асимметрический коэффициент “лямбда” (asymmetric lambda)

Мера выраженного в процентах улучшения прогнозирования значения зависимой переменной при данном значении независимой переменной.

Коэффициент “лямбда” используется в том случае, когда **переменные измерены с помощью номинальной шкалы**.

Значения коэффициента “лямбда” лежат в пределах от 0 до 1.

Значение “лямбда”, равное 0, означает, что никакого улучшения в прогнозировании не наблюдается. Значение 1 указывает на то, что прогноз может быть сделан без ошибки. Это происходит тогда, когда каждая категория независимой переменной связана с одной категорией зависимой переменной.

Также рассчитывают **симметричный коэффициент “лямбда”** (symmetric lambda) — средним значением двух асимметричных значений

# Другие статистики таблиц сопряженности признаков

( $\tau_b$ ;  $\tau_c$ ;  $\gamma$ ...)

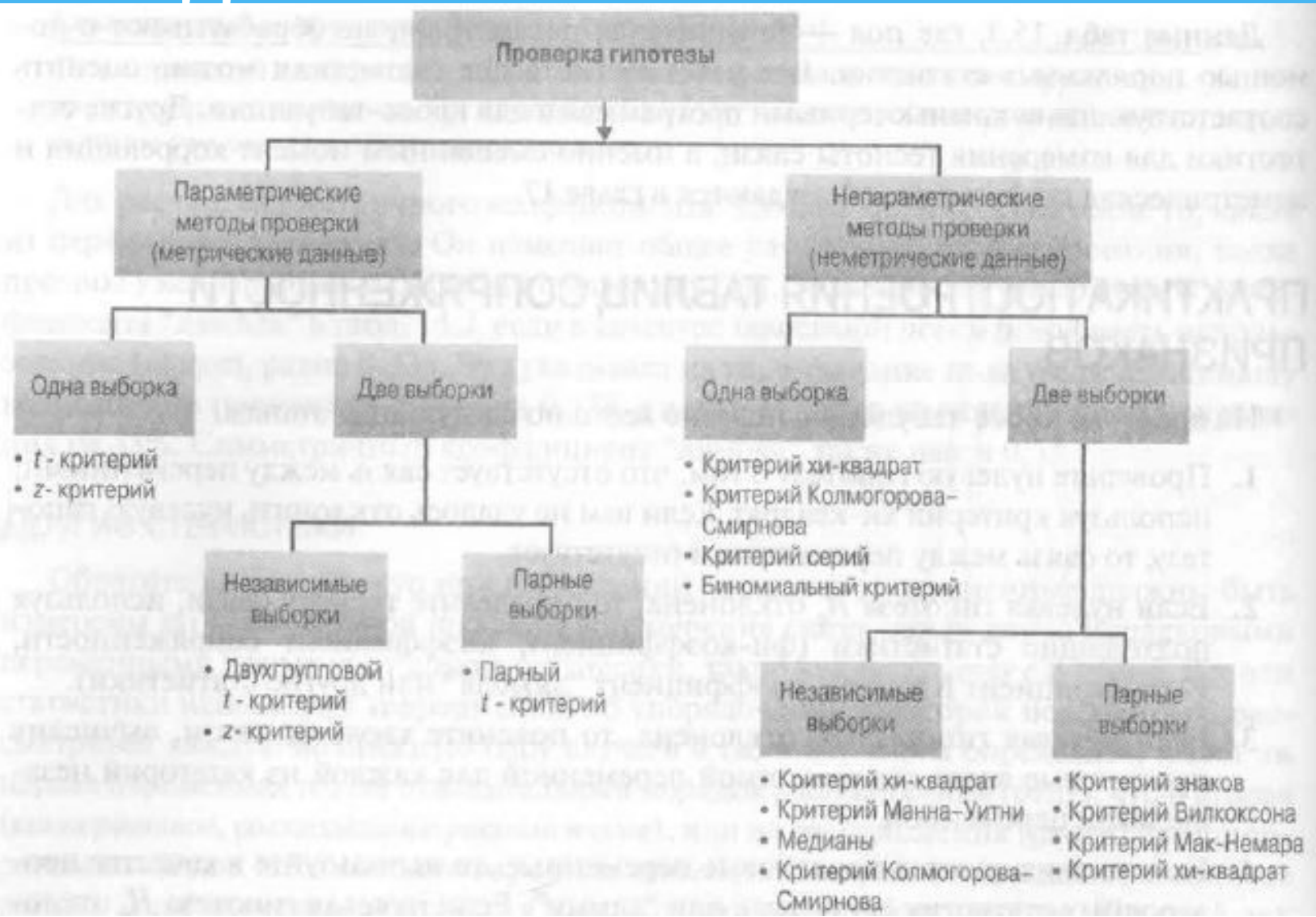
- \* Для измерения связи между двумя **порядковыми переменными** применяют другие статистики, такие как  $\tau_b$ ,  $\tau_c$  и  $\gamma$ .
- \* Все эти статистики используют информацию об упорядочении категорий переменных, рассматривая каждую возможную пару случаев в таблице, чтобы определить, имеет ли первая переменная тот же относительный порядок расположения (ранг), что и вторая (конкорданное, согласованное расположение), или их расположения (ранги) имеют обратный порядок (несогласованное расположение), или их ранги совпадают (связанные ранги).
- \* Эти статистики отличаются только способом обработки рангов. Как  $\tau_b$ , так и  $\tau_c$ , корректируют по числу связанных рангов.  **$\tau_b$**  ( $\tau_b$ ) больше всего подходит для квадратных таблиц, в которых количество строк и колонок равно.
- \* **Статистика “гамма”** ( $\gamma$ ) не учитывает ни связанные ранги, ни размер таблицы.

# ПРАКТИКА ПОСТРОЕНИЯ ТАБЛИЦ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

На практике кросс-табуляцию полезно вести по следующим этапам.

1. Проверьте нулевую гипотезу о том, что отсутствует связь между переменными, используя критерий хи-квадрат. Если вам не удалось отклонить нулевую гипотезу, то связь между переменными отсутствует.
2. Если нулевая гипотеза  $H_0$  отклонена, то определите тесноту связи, используя подходящие статистики (фи-коэффициент, коэффициент сопряженности, V-коэффициент Крамера, коэффициент “лямбда” или другие статистики).
3. Если нулевая гипотеза  $H_0$  отклонена, то поясните характер связи, вычислив процентные значения зависимой переменной для каждой из категорий независимой переменной.
4. Если анализируются порядковые переменные, то используйте в качестве проверочных статистик  $\tau_a$ ,  $\tau_b$ ,  $\tau_c$  или “гамму”. Если нулевая гипотеза  $H_0$  отклонена, то определите тесноту связи, используя величину и направление связи, а также учитывая знак проверяемой статистики.

# ПРОВЕРКА ГИПОТЕЗ: РАЗЛИЧИЯ МЕЖДУ ЗНАЧЕНИЯМИ ПЕРЕМЕННЫХ





# ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

## Параметрические методы проверки гипотез (parametric tests)

Предполагают, что изучаемые переменные измерены с помощью **интервальной** шкалы

Использование параметрических критериев позволяет сделать статистический вывод относительно среднего значения генеральной совокупности. Обычно для этой цели используют **t-критерий (t-test)**. В основе критерия лежит t-статистика Стьюдента (Student).

- ✓ **T-критерий (t-test)** - Одномерный метод проверки гипотез, использующий T-распределение. (если стандартное отклонение неизвестно и размер выборки мал).
- ✓ **T-статистика (t-statistic)** - подразумевает, что переменная имеет колоколообразное распределение, среднее известно и дисперсия ГС определена по данным выборки.
- ✓ **T-распределение (t-distribution)** - Симметричное колоколообразное распределение (центральная часть по площади — меньше, чем у нормального), используемое для выборок небольшого размера ( $n < 30$ ).

# Процедура проверки гипотезы методом t-критерия:

1. Сформулировать нулевую ( $H_0$ ) и альтернативную ( $H_1$ ) гипотезы.
2. Выбрать соответствующую формулу для вычисления t-статистики.
3. Выбрать уровень значимости  $\alpha$  для проверки нулевой гипотезы  $H_0$ . Обычно выбирают уровень значимости  $\alpha$ , равный 0,05.
4. Взять одну или две выборки и для каждой вычислить значение средней и стандартное отклонение.
5. Вычислить значение t-статистики, приняв, что нулевая гипотеза  $H_0$  верна.
6. Вычислить число степеней свободы и оценить вероятность получения большего значения статистики. (Альтернативно, вычислить критическое значение t-статистики.)
7. Если вероятность, рассчитанная на этапе 6, меньше, чем уровень значимости  $\alpha$ , выбранный на этапе 3, то отклонить нулевую гипотезу ( $H_0$ ). Если значение вероятности больше, то  $H_0$  не отклонять. (Альтернативно, если значение вычисленной на этапе 5 t-статистики больше критического значения, определенного на этапе 6, то отклонить нулевую гипотезу  $H_0$ . Если вычисленное значение меньше критического значения, то  $H_0$  не следует отклонять.) Неудачная попытка отклонить нулевую гипотезу только означает, что истинное положение несущественно (статистически незначимо) отличается от положения  $H_0$ .
8. Выразить полученный результат в виде, пригодном для решения проблемы маркетингового исследования.

# Проверка гипотезы методом t-критерия для одной выборки:

- \* Проверка утверждений о соотношении значения одной переменной по сравнению с известной или заданной величиной
- \* Эти утверждения сформулируем с точки зрения нулевой гипотезы, которую затем проверим, используя статистический критерий для одной выборки, такой как t- или z-критерий. Если маркетолог использует t-критерий для проверки значения средней, его интересует, совпадает ли значение генеральной средней со значением, задаваемым в утверждении нулевой гипотезы ( $H_0$ ).
- \* Если известно стандартное отклонение  $\sigma$  и используем его, то лучше использовать z-критерий (z-test).

$$z = (\bar{X} - \mu) / \sigma_{\bar{X}}$$

# Проверка гипотезы методом t-критерия для двух независимых выборок:

Выборки, взятые случайным образом из разных изучаемых совокупностей, называют **независимыми выборками** (independent samples).

Как и для одной выборки, проверка гипотез может проводиться относительно значений средних или долей:

- \* **Средние.** Если неизвестно, равны ли дисперсии двух совокупностей, то для проверки выборочной дисперсии используем **F-критерий**, или **критерий Фишера (F-test)**:

$$F_{(n_1-1), (n_2-1)} = \frac{s_1^2}{s_2^2}$$

$n_1$  — размер выборки 1;  
 $n_2$  — размер выборки 2;  
 $n_1 - 1$  — степени свободы для выборки 1;  
 $n_2 - 1$  — степени свободы для выборки 2;  
 $s_1^2$  — выборочная дисперсия для выборки 1;  
 $s_2^2$  — выборочная дисперсия для выборки 2.

# Проверка гипотезы методом t-критерия для двух независимых выборок:

(продолжение).

- \* Доли. Для одной выборки используют z-критерий.
- \* Если вычисленное значение меньше, чем критическое, нулевую гипотезу нельзя отклонить.

# Проверка гипотезы методом t-критерия для парных выборок:

## Парные или связанные выборки (paired samples)

В проверке гипотез наблюдения называют парными, если два набора наблюдений относятся к **одним и тем же** респондентам

- \* Различие, возникающее в этой ситуации, проверяют с помощью **t-критерия парных выборок (paired samples t-test)**.
- \* Чтобы вычислить значение t-критерия для парных выборок, вводят переменную разности, обозначаемую  $D$ , и вычисляют ее среднее и дисперсию. После этого вычисляют t-статистику :

$$\begin{aligned}H_0: \mu_D &= 0; \\H_1: \mu_D &\neq 0; \\t_{n-1} &= \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}},\end{aligned}$$

$$\begin{aligned}\bar{D} &= \frac{\sum_{i=1}^n D_i}{n}; \\s_D &= \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}}; \\S_B &= \frac{s_D}{\sqrt{n}}.\end{aligned}$$

# НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ для одной выборки

**Непараметрические методы** проверки гипотез (nonparametric tests)

Предполагают, что переменные измерены с помощью **номинальной** или **порядковой** шкал.

- \* **Критерий согласия Колмогорова-Смирнова для одной выборки (Kolmogorov—Smirnov (K—S) one-sample test)** - Непараметрический метод проверки степени согласия эмпирической функции распределения переменной с определенным теоретическим законом распределения. (Чем больше значение  $K$ , тем больше уверенности, что нулевая гипотеза ( $H_0$ ) неверна);
- \* можно выполнять проверку гипотезы по **критерию хи-квадрат**;
- \* **Биномиальный критерий (binomial test)** - Статистический критерий согласия для дихотомических переменных. Он проверяет степень согласия наблюдаемого количества наблюдений в каждой категории с количеством наблюдений, ожидаемым в условиях конкретного биномиального распределения (для **дихотомических (двузначных) переменных**)

# НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ для двух независимых выборок :

- \* **U-критерий Манна—Уитни (Mann—Whitney U-test)** - Статистический критерий для переменной, измеренной с помощью порядковой шкалы, который сравнивает различие в показателях положения двух совокупностей исходя из наблюдений, взятых из двух независимых выборок.;
- \* **Двухвыборочный медианный критерий (two-sample median test)** - Непараметрический метод проверки, который определяет, действительно ли две группы взяты из совокупностей с одной и той же медианой. (не столь мощный, как U-критерий Манна—Уитни);
- \* **(Kolmogorov-Smirnov two-sample test)** - Непараметрический метод проверки того, действительно ли два распределения являются одинаковыми. Этот критерий учитывает любые различия между двумя распределениями, включая медиану, вариацию и асимметрию.



# НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ для парных выборок:

- \* **Критерий попарных сравнений Вилкоксона (Wilcoxon matched-pairs signed-ranks test)** - Непараметрический метод проверки, с помощью которого анализируют разности между парными наблюдениями, учитывая их величину;
- \* **Критерий знаков (sign test)** - Непараметрический критерий для изучения разностей в показателях центральной тенденции двух генеральных совокупностей на основе парных наблюдений, который только сравнивает знаки разностей между парами переменных, но не учитывает величину разностей. (не столь мощный, как критерий Вилкоксона);
- \* Иногда возможно использовать критерий Мак-Немара/  
критерий хи-квадрат.

Спасибо за внимание!

