СТАТИСТИКА

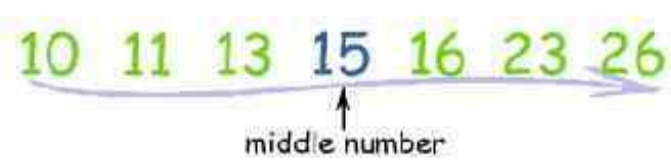
ВСЕ СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ ДЕЛЯТСЯ НА 3 БОЛЬШИЕ ГРУППЫ:

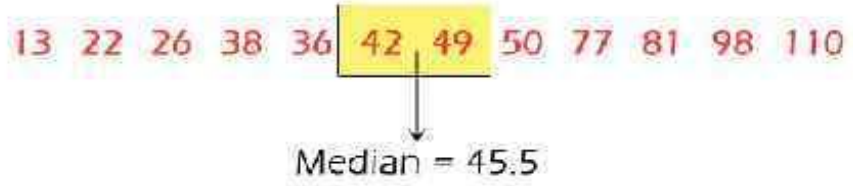
- ***Меры центральной тенденции*** - показывают расположение среднего, типичного значения признака, вокруг которого сгруппированы остальные наблюдения
- ***Меры рассеяния*** (меры изменчивости, показатели вариации) - характеризуют значения между отдельными показателями выборки. Позволяют судить о степени однородности полученного множества, и о надежности полученных результатов
- ***Меры связи*** (меры корреляции) - позволяют изучить взаимосвязь между двумя признаками/переменными

МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ (МЕРЫ ПОЛОЖЕНИЯ, МЕРЫ ЛОКАЛИЗАЦИИ)

Показывают наиболее типичное значение для данной выборки

- Среднее значение (M) - среднее арифметическое
- Медиана (Me) - средняя точка распределения
  - Если кол-во значений нечетное, то Me - среднее значение в ранжированном списке
  - Если кол-во значений четное, то Me - среднее арифметическое между двумя центральными значениями





- Мода (Mo) - наиболее часто встречающееся значение признака в выборке

МЕРЫ РАССЕЯНИЯ (МЕРЫ ИЗМЕНЧИВОСТИ, ПОКАЗАТЕЛИ ВАРИАЦИИ)

Показывают разброс значений признака в выборке

- Размах - разность максимального и минимального значения *(Недостаток: не характеризует распределение целиком, а только крайние значения)*
- Интерпроцентильный размах/интервал - значения каких-либо процентилей распределения, например, 10-го и 90-го
- Интерквартильный размах/интервал - значения 25-го и 75-го процентилей (такой интервал независимо от вида распределения включает 50% значений признака в выборке)

МЕРЫ РАССЕЯНИЯ
(МЕРЫ ИЗМЕНЧИВОСТИ, ПОКАЗАТЕЛИ ВАРИАЦИИ)

- Дисперсия - характеризует, насколько частные значения отклоняются от средней величины в данной выборке *(чем больше дисперсия, тем больше "разброс данных").* Находится как средняя арифметическая квадратов отклонений от общей средней.

- Среднее квадратическое (стандартное) отклонение (СКО, s, SD) - позволяет оценить, насколько бОльшая часть результатов данного исследования отклоняется от среднего значения (находится как квадратный корень из дисперсии)

- Стандартная ошибка (SE-standard error) - оценка возможного отличия между значением среднего в анализируемой выборке и истинным средним, характерным для всей популяции. С увеличением выборки уменьшается данная ошибка, так как чем больше наблюдений, тем больше вероятность, что полученные данные близки к истинным.

ПОНЯТИЕ О КВАНТИЛЯХ

Квантили (ед.ч. - Квантиль) - величины, разделяющие ранжированный ряд на равные части.

Разновидности квантилей:

- 1. **Медиана** - делит на 2 равные части (пополам)
- 2. **Квартили** - делит на 4 равные части
- 3. **Децили** - делит на 10 равных частей
- 4. **Перцентили** - делит на 100 равных частей

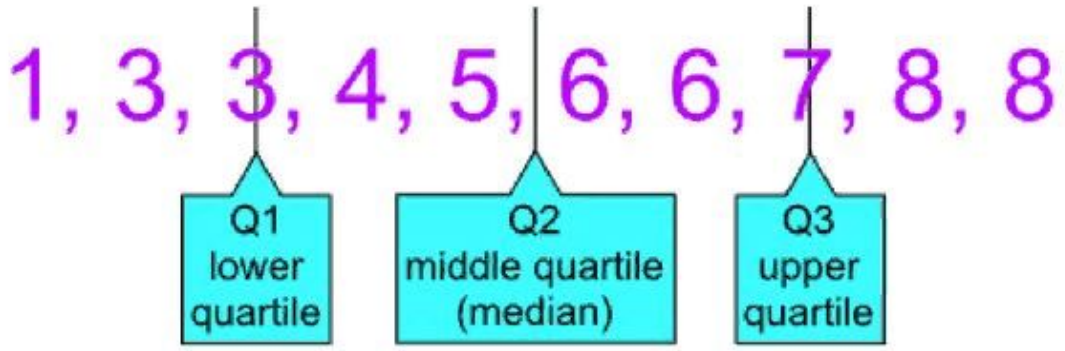
ПОДРОБНЕЕ О КВАРТИЛЯХ

Квартили делят ранжированный ряд на 4 равные части

- ***Нижний (первый) квартиль Q1*** - это медиана левой половины упорядоченного ряда. 25% значений меньше Q1

- ***Верхний (третий) квартиль Q3*** - медиана правой половины упорядоченного ряда. 25% значений больше Q3

- ***Второй квартиль Q2*** - медиана



АНАЛИЗ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Первый этап - анализ вида распределения

От вида распределения зависят:

- *Выбор способа описания центральной тенденции*
- *Выбор способа описания изменчивости значений признака*
- *Выбор методов дальнейшего анализа данных*

КАК ОПРЕДЕЛИТЬ ВИД РАСПРЕДЕЛЕНИЯ?

- ??? 4 способа с помощью программы STATISTICA, с их помощью выдвигаем одну из гипотез:
  - Нулевая гипотеза (Но) - утверждает, что распределение исследуемого признака в генеральной совокупности соответствует закону нормального распределения
  - Альтернативная гипотеза (Н1) - утверждает, что распределение исследуемого признака в генеральной совокупности **не** соответствует закону нормального распределения
- ??? 3 критерия:
  1. Колмогорова - Смирнова: применяется, если среднее значение и среднее квадратическое отклонение известны априори
  2. Лиллиефорса: применяется, когда среднее значение и среднее квадратическое отклонение **не известны** априори, а вычисляются по выборке
  3. ?Чем отличается от первого? Шапиро-Уилка: применяется так же, если известны среднее значение и среднее квадратическое отклонение априори. Данный критерий предпочтителен, так как является самым "мощным" и универсальным

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

- После использования программы STATISTICA будут получены результаты анализа распределения каждого признака - p.
- Если p < 0,05 => принимается альтернативная гипотеза -> распределение отличается от нормального -> далее будут использованы *непараметрические методы анализа данных*
- Если p ⩾ 0,05 => принимается нулевая гипотеза -> нормальное распределение -> далее будут использованы *параметрические методы анализа данных*

<u>P никак не отражает величину различий между группами, поэтому часто рассчитывают ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ (ДИ)</u>

Доверительный интервал - диапазон значений вокруг истинного значения.

ДИ с определённой вероятностью включает в себя истинные значения в генеральной совокупности.

КАКИЕ ДАННЫЕ НЕОБХОДИМО УКАЗЫВАТЬ ПРИ ОПИСАНИИ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ?

Для описания нормального распределения:

- Число наблюдений (объектов исследования)
- Среднее значение
- Среднее квадратическое отклонение (СКО)

Для описания распределения, отличающегося от нормального:

- Число наблюдений (объектов исследования)
- Медиану
- Верхний и нижний квартили

??? ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

- 1. ***Непарный t-тест (тест Стьюдента)*** - с его помощью проводят проверку гипотезы "***Но"*** об отсутствии различий средних значений переменной в двух независимых выборках
- 2. Если данные зависимые (повторные наблюдения за одним и тем же человеком или исследование людей по парам), то рекомендуется применять ***парный t-тест***
- 3. ***Т-тест Уэлча -***
- 4. ***Дисперсионный анализ -***
- 5. ***Дисперсионный анализ с повторным измерением -***

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

- Непрерывные/дискретные переменные???

СРАВНЕНИЕ ПАРАМЕТРИЧЕСКИХ И НЕПАРАМЕТРИЧЕСКИХ МЕТОДОВ

К ***преимуществам непараметрических методов*** можно отнести следующие:

- могут быть использованы, когда характеристики популяции, из которой делается выборка, частично неизвестны;
- бóльшая мощность;
- относительная несложность вычислений (в большинстве случаев);
- менее жесткие начальные допущения

Недостатками непараметрических методов являются:

- меньшая эффективность, чем у параметрических методов;
- меньшая специфичность;
- потенциальная трудоемкость при применении к большим массивам данных.

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ - МЕРА УВЕРЕННОСТИ В "ИСТИННОСТИ" РЕЗУЛЬТАТА

- Статистическая значимость определеяется значением p-уровня *(p-value)*
- Чем выше p-уровень, тем ниже уровень доверия к полученным результатам *(обратная зависимость)*

↑ p-уровень ⇒ ↓ уровень доверия

- P > 0,05 результатам нельзя доверять
- p ⩽ 0,05 статистически значимые результаты
- P < 0,01 статистически высокозначимые результаты

<u>Пример:</u> p-уровень - 5% (0,05) показывает, что сделанный при анализе вывод является случайной особенностью с вероятностью 5%. Другими словами, с вероятностью 95% вывод можно распространить на все объекты.