# Empirical Research Methods 1

# P-value. Descriptive and inferential statistics and diagrams

*Instructor: Albulene Grajcevci*

*Tutors: Shravani & Piumal*

*Session 6: December 4th, 2023*

# Agenda

- Data distribution, normal distribution, and parametric tests
- P-value
- Statistics: descriptive and inferential

- Applied descriptive statistics:
    - Frequencies
    - Measures: central tendency (mean, median, and mode) and dispersion (standard deviation and variance)
    - Types of graphs
    - Calculating descriptive statistics (frequencies, mean, median, mode, SD, etc.) and graphs on SPSS
    - Self-paced activities

# Data distribution, normal distribution, and parametric tests

# Data distribution, normal distribution, and parametric tests

- Fancy definition: “A data distribution is a function or a listing which shows all the possible values (or intervals) of the data.” (Stephanie from Statistics How To, 2017)

- Simplified definition: “A distribution is a visual way to see all the probable values of our data.” (Me, 2020)

# Normal distribution

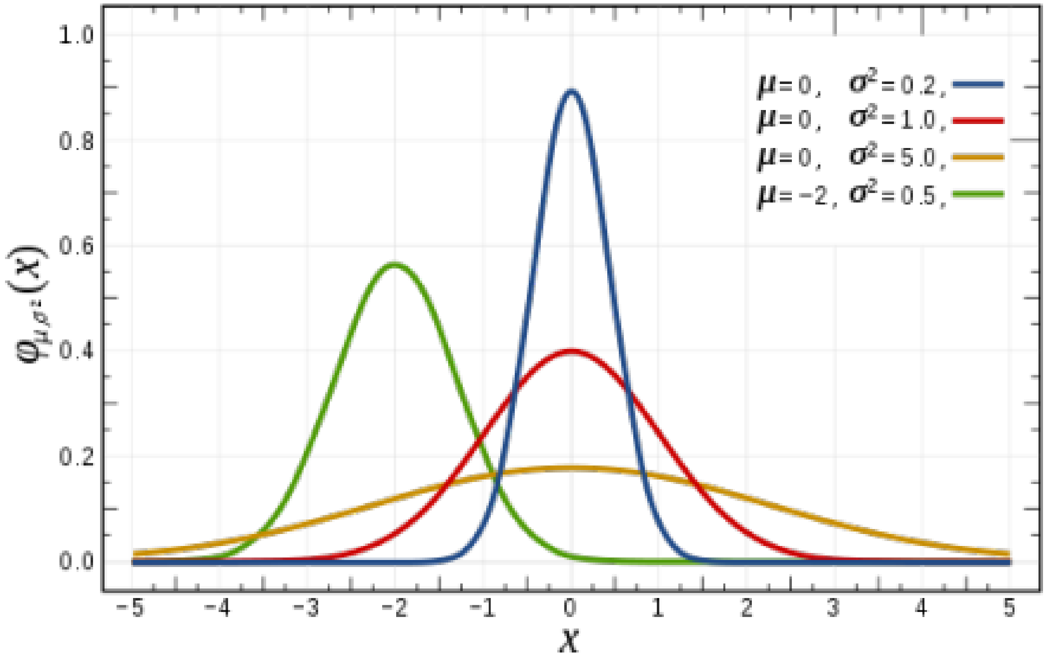
A.k.a. Gaussian or Gauss or Laplace–Gauss distribution or Bell curve

• Data that produces (or fits) a normal distribution can be called “normal data” (e.g. intelligence scores)

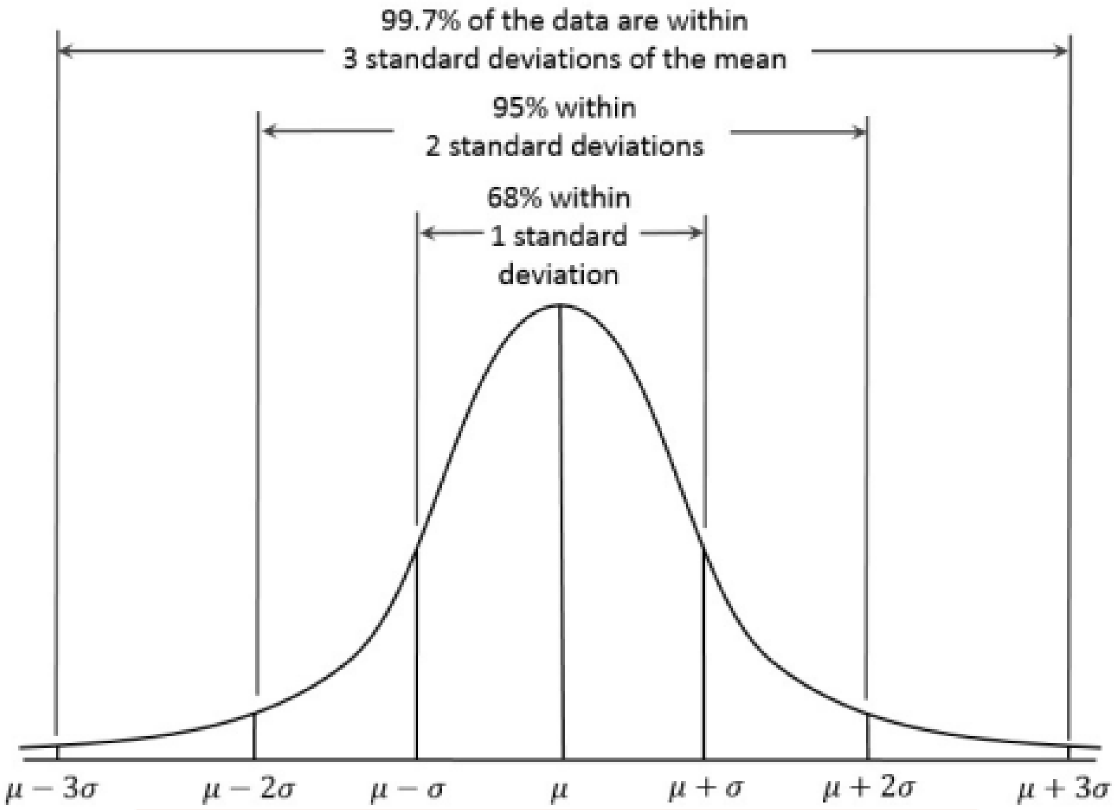
• It is a probability distribution that is symmetrical, unimodal, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

• Normality is a basic assumption/ requirement for parametric tests. Therefore, we want normal data because we want parametric tests!

# Normal distribution



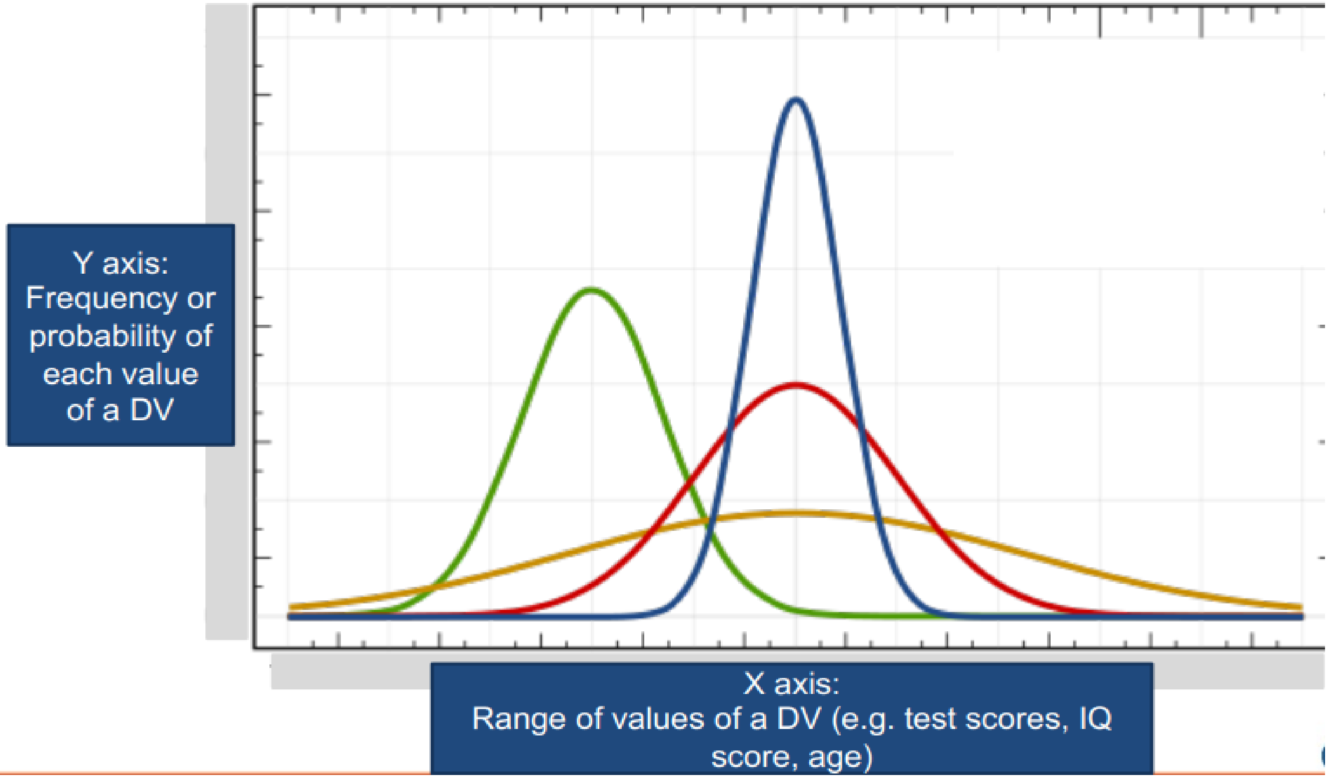

Different types of normal distributions




Standard normal distributions

# Normal distribution

- You will see lots of distributions, with different values, numbers or letters which may be confusing
- However, most data distributions represent the following information:



# Parametric tests

We mentioned that we want normal data because we want (to use) parametric tests. But what are parametric tests?

1. A parametric test is a statistical test which makes certain assumptions about the distribution of the unknown parameter of interest and thus the test statistic is valid under these assumptions.
2. Parametric tests need to ensure four basic assumptions for the test to be accurate:
    1. A normally sampling distribution
    2. Homogeneity of variance
    3. Interval or ratio data
    4. Independence

**Why do we want parametric tests? Because they have more power.**

# Parametric tests

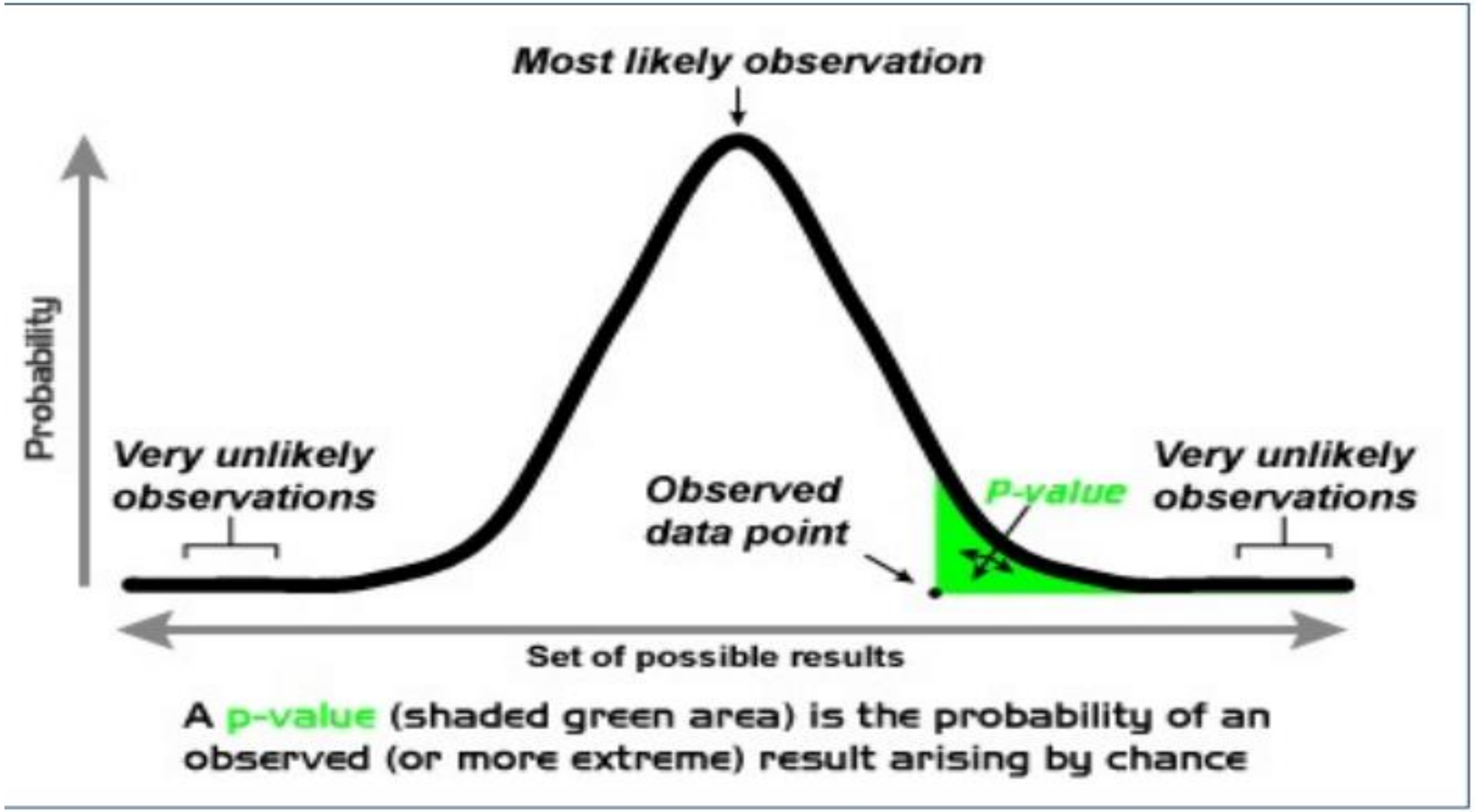
-Homogeneity of variance: the assumption that the variance of one variable is stable (i.e. relatively similar) at all levels of another variable
-Independence: the assumption that one data point does not influence another. When data come from people, it basically means that the behaviour of one person does not influence the behavior of another. Sometimes also refered as “independence of observations”

# Understanding the p-value

# P value

- A p-value is a measure of the probability that an observed difference could have occurred just by random chance. That’s why we want p-values to be as small as possible.
- A p-value indicates evidence against the null hypothesis, indicating the probability the null is correct (and the results are random).
- In social sciences, a p-value of $\leq 0.05$ normally signifies a statistically significant result. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random).
- Therefore, we reject the null hypothesis, and accept the alternative hypothesis.
- However, this does not mean that there is a 95% probability that the research hypothesis is true. The p-value is conditional upon the null hypothesis being true is unrelated to the truth or falsity of the research hypothesis.

# P value



Most likely observation
Probability
Very unlikely observations
Observed data point
P-value
Very unlikely observations
Set of possible results
A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

# P value- QUESTIONS

1. How does the alpha level and the p-value relate?
2. What does “setting the alpha to 1%” mean and what would that imply for the p-value?
3. With what you know so far about hypothesis testing (H0 and HA), alpha level/error, and p-values, why do you think we need to phrase all of our statistic result decisions in terms of accepting or rejecting the H0?
4. Do you think there are cases when we would want our p-value to be bigger than our alpha?

# Reporting p-values

- appeared to be marginally significant (p<0.10)
- approached acceptable levels of statistical significance (p=0.054)
- approached but did not quite achieve significance (p>0.05)
- approached but fell short of significance (p=0.07)
- approached conventional levels of significance (p<0.10)
- approached near significance (p=0.06)
- approached our criterion of significance (p>0.08)
- approached significant (p=0.11)
- approached the borderline of significance (p=0.07)
- approached the level of significance (p=0.09)
- approached trend levels of significance (p0.05)
- approached, but did reach, significance (p=0.065)
- approaches but fails to achieve a customary level of statistical significance (p=0.154)
- approaches statistical significance (p>0.06)
- approaching a level of significance (p=0.089)
- approaching an acceptable significance level (p=0.056)
- approaching borderline significance (p=0.08)
- approaching borderline statistical significance (p=0.07)
- approaching but not reaching significance (p=0.53)
- approaching clinical significance (p=0.07)
- approaching close to significance (p<0.1)
- approaching conventional significance levels (p=0.06)
- approaching conventional statistical significance (p=0.06)
- approaching formal significance (p=0.1052)
- approaching independent prognostic significance (p=0.08)
- approaching marginal levels of significance p<0.107)
- approaching marginal significance (p=0.064)
- approaching more closely significance (p=0.06)
- approaching our preset significance level (p=0.076)
- approaching prognostic significance (p=0.052)
- narrowly avoided significance (p=0.052)
- narrowly eluded statistical significance (p=0.0789)
- narrowly escaped significance (p=0.08)
- narrowly evaded statistical significance (p>0.05)
- narrowly failed significance (p=0.054)
- narrowly missed achieving significance (p=0.055)
- narrowly missed overall significance (p=0.06)
- narrowly missed significance (p=0.051)
- narrowly missed standard significance levels (p<0.07)
- narrowly missed the significance level (p=0.07)
- narrowly missing conventional significance (p=0.054)
- near limit significance (p=0.073)
- near miss of statistical significance (p>0.1)
- near nominal significance (p=0.064)
- near significance (p=0.07)
- near to statistical significance (p=0.056)
- near/possible significance(p=0.0661)
- near-borderline significance (p=0.10)
- near-certain significance (p=0.07)
- nearing significance (p<0.051)
- nearly acceptable level of significance (p=0.06)
- nearly approaches statistical significance (p=0.079)
- nearly borderline significance (p=0.052)
- nearly negatively significant (p<0.1)
- nearly positively significant (p=0.063)
- nearly reached a significant level (p=0.07)
- nearly reaching the level of significance (p<0.06)

# STATISTICS: DESCRIPTIVE AND INFERENTIAL

# Descriptive and inferential statistics

1. Collecting
2. Organizing and
3. Interpreting data

| | Descriptive | Inferential |
|---|---|---|
| Input values | Values that describe the characteristics of a sample or population | Values that infer results of a sample to the population from which the sample is drawn |
| Goals | Organize, simplify, summarize and describe data to help us know how our sample is and its characteristics | Provide information for testing our hypothesis/ -es and reach a decision in terms of p-value, significan and the H0<br>Provide certainty about how generalizable the results from our samply apply to its population |
| Limitations | Only allow you to make summations about the people or objects that you have actually measured | A degree of uncertainty even in terms of the accuracy of the results (i.e. Type 1 and 2 errors) |

# APPLIED DESCRIPTIVE STATISTICS

Frequencies ● Measures: central tendency (mean, median, and mode) and dispersion (standard deviation and variance) ● Types of graphs ● SPSS calculations of frequencies and graphs

# Measures: Frequencies

1. Absolute Frequency: total number of occurrences of a datapoint in our dataset
2. Relative Frequency: absolute frequency divided by the total number of occurrences

| Example. Dataset_1 contains the following datapoints for the variable 'Age': |
|---|
| 22, 23, 23, 23, 24, 24, 24, 24, 25, 25, 25, 25, 25, 26, 26 |

| Age | Absolute Frequency | Relative Frequency | Percent in % |
|---|---|---|---|
| 22 | 1 | 0.067 | 6.6 |
| 23 | 3 | 0.2 | 20 |
| 24 | 4 | 0.267 | 26.7 |
| 25 | 5 | 0.333 | 33.3 |
| 26 | 2 | 0.133 | 13.3 |
| **Total** | 15 | 1 | 100 |

# Measures: Frequencies

| Age | Absolute Frequency | Relative Frequency | Percent in % |
|---|---|---|---|
| 22 | 1 | 0.067 | 6.6 |
| 23 | 3 | 0.2 | 20 |
| 24 | 4 | 0.267 | 26.7 |
| 25 | 5 | 0.333 | 33.3 |
| 26 | 2 | 0.133 | 13.3 |
| **Total** | 15 | 1 | 100 |

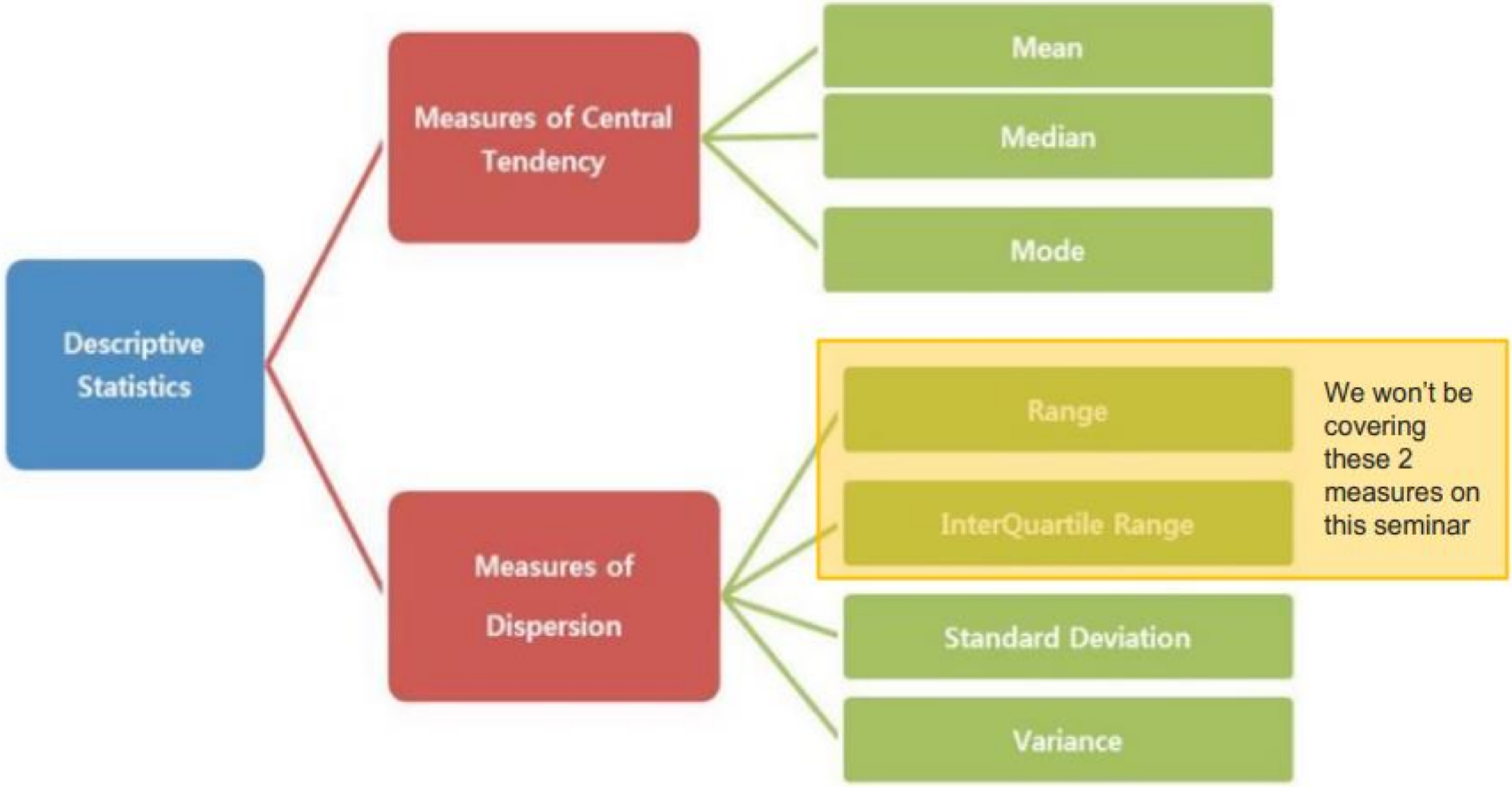
Relative frequency calculation:
→ of value “23”
3/15 = 0.2

→ of value “24”
4/15 = 0.267

Percentage calculation:
→ of value “23”
0.2 * 100 = 20%

→ of value “24”
0.267 * 100 = 26.7%

# Measures: Frequencies



# Measures of central tendency

- Most common: mean, median and mode

➢ Important symbols: Commonly found in publications

$\mu$ = Population mean
$\bar{x}$, x_bar, x bar = Sample mean

$\sigma^2$ = Population variance
$s^2$ = Sample variance

$\sigma$ = Population standard deviation
s or SD = Sample standard deviation

N = Population size
n = Sample size

# Measures of central tendency- Understanding the MEAN

- The mean is sensitive to extreme scores (outliers) in a dataset

➢ Some other symbols commonly found in the mean formula:

"μ": population mean

"x̄": sample mean

| Dataset | |
|---|---|
| Participant | Age |
| 1 | 23 |
| 2 | 22 |
| 3 | 24 |
| 4 | 23 |
| 5 | 26 |
| 6 | 25 |
| 7 | 24 |
| 8 | 26 |
| 9 | 24 |
| 10 | 25 |
| 11 | 25 |
| 12 | 23 |
| 13 | 25 |
| 14 | 25 |
| 15 | 24 |

$$\overline{x} = \frac{23 + 22 + 24 + 23 + 26 + 25 + 24 + 26 + 24 + 25 + 25 + 23 + 25 + 25 + 24}{15} = 24.27$$

# Measures of central tendency- Understanding the MEDIAN

Median = midpoint
The median is a better representative of a sample when scores are extreme, not sensitive to extreme scores

How to determine the median:

1. From all the values of a varaible in your dataset, order them from lowest to highest
2. Once ordered, the median is the value in the middle
   a) If you have an odd amount of values, the value in the middle is the median
   b) b) If you have an even amount of values, the average of the two values in the middle is the median

| Age | Absolute Frequency |
|---|---|
| 22 | 1 |
| 23 | 3 |
| 24 | 4 |
| 25 | 5 |
| 26 | 2 |
| Total | 15 |

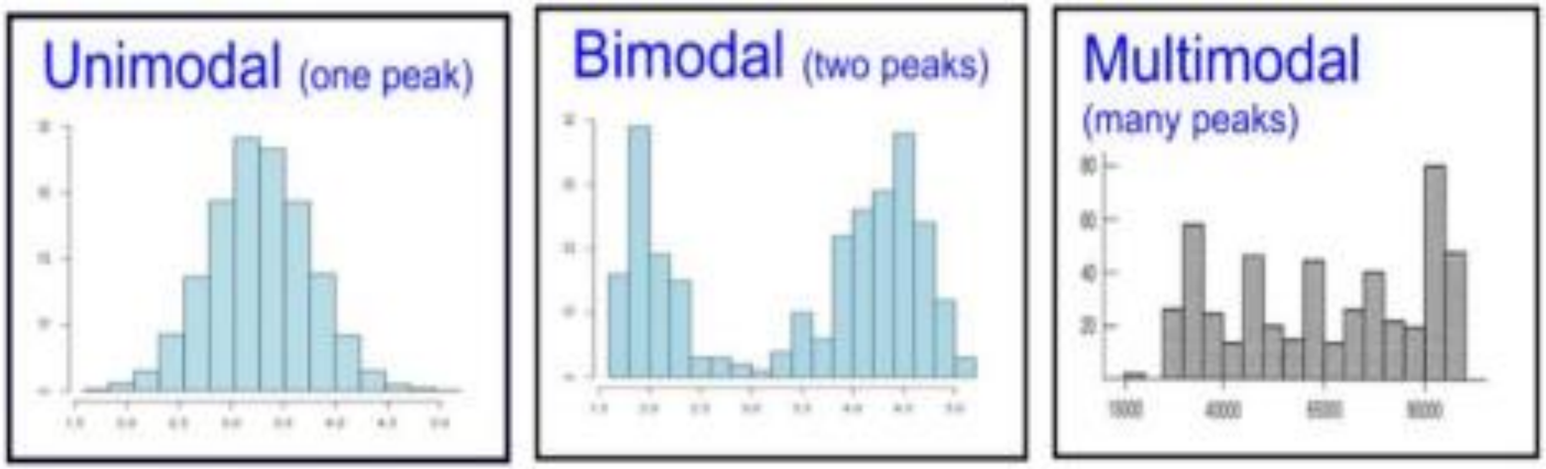
~~22~~, ~~23~~, ~~23~~, ~~23~~, ~~24~~, ~~24~~, ~~24~~, 24, ~~25~~, ~~25~~, ~~25~~, ~~25~~, ~~25~~, ~~26~~, ~~26~~

Median = 24

Adapted from the slides from Lena Schielke, Ghafiq Alkurdi and Chris Vizcarrondo, 2020

# Measures of central tendency- Understanding the MODE

- Mode: the value that appears most frequently in a dataset
- Specially useful for nominal and ordinal scale data
- **<u>Several modes? Yes!</u>**
- More than one mode is possible Visual depictions of modes: Single mode, bi modal, multi-modal (>2 peaks)




| Age | Absolute Frequency |
|---|---|
| 22 | 1 |
| 23 | 3 |
| 24 | 4 |
| 25 | 5 |
| 26 | 2 |
| **Total** | 15 |

From the variable age, what is the value with the highest absolute frequency?

Mode = 25

Adapted from the slides from Lena Schielke, Ghafiq Alkurdi and Chris Vizcarrondo, 2020

# Activity 1

- Calculate the mean, mode and median of the following data:

2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 1, 1

# Measures – dispersion: Key definitions of variance and standard deviation (SD)

- Variance ($s^2$) = sum of the squared deviations of the mean, divided by N
- Standard deviation (SD or s) = square root of the variance. Normally reported in publications
- The variance describes dispersion of data points from the mean value
- It is the average of the squared differences from the mean

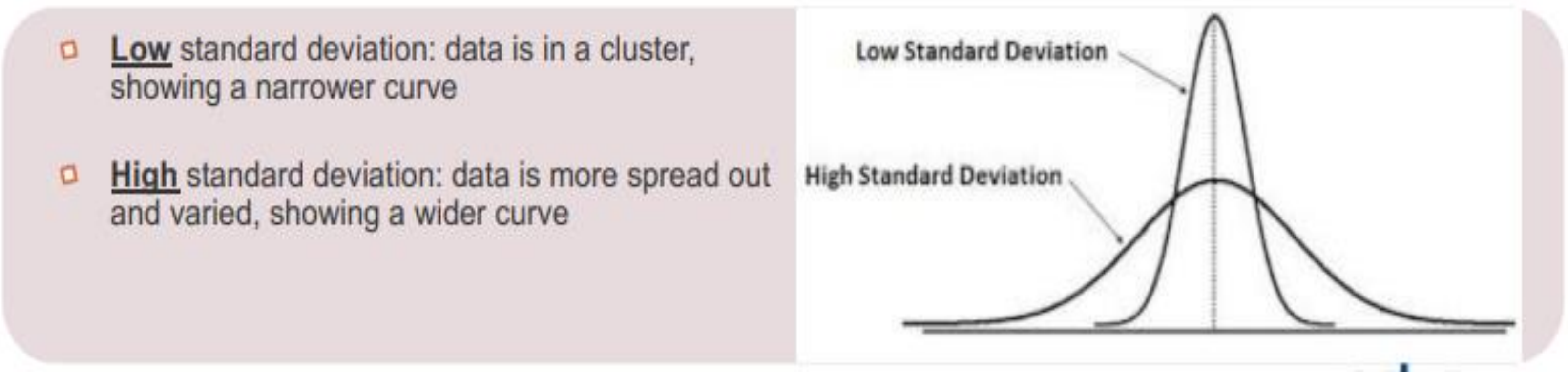
- Steps for calculating variance

1. Determine the mean of your dataset
2. For each number in your dataset (X), subtract the mean ($\bar{X}$) and square the result (squared difference). Sum the squared results
3. Divide the sum of squared results (or sum of squares) by "n-1" (for sample SD) or by "N" (for population SD)

$$s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

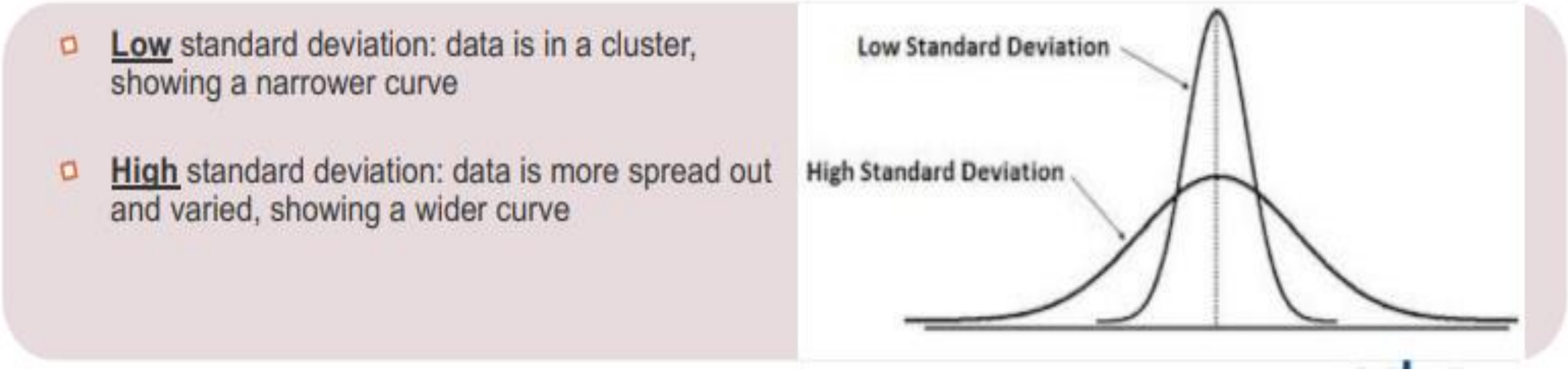
# Measures – dispersion: standard deviation

- The standard deviation describes how much variation exists from average/mean.
- Measures variability and shows how spread-out numbers are Its formula is the square root of the variance
- Standard deviation is the most frequently used measure of dispersion or variability and is sensitive to extreme values
- Is depicted as a bell curve with the center showing the normal distribution that is 1 standard deviation of the mean



# Measures – dispersion: standard deviation

- The standard deviation describes how much variation exists from average/mean.
- Measures variability and shows how spread-out numbers are Its formula is the square root of the variance
- Standard deviation is the most frequently used measure of dispersion or variability and is sensitive to extreme values
- Is depicted as a bell curve with the center showing the normal distribution that is 1 standard deviation of the mean



# Understanding the difference between variance and SD

| | Variance | Standard deviation |
|---|---|---|
| Definition | The variance is a measure of variability of values from the average | Standard deviation looks at how spread out a group of numbers is from the average |
| Purpose | Useful when doing certain math calculations | Describes variability of data of figures that are quite large; more meaningful than variance |
| Calculation | Average of the squared differences from the mean | Root of the average of the squared differences from the mean |
| Unit | Squared units | Same unit as values |

# Example

We want to calculate the mean, variance and standard deviation of the following sample data: 5, 4, 5, 2

mean = $\bar{x} = \frac{5+4+5+2}{4} = 4$

$$s^2 = \frac{\sum (X-\bar{X})^2}{N-1} = s^2 = \frac{(5-4)^2+(4-4)^2+(5-4)^2+(2-4)^2}{4-1} = \frac{1+0+1+4}{3} = \frac{6}{3} = 2$$

Variance ($s^2$) = 2

Standard deviation (s) = $\sqrt{2}$ = 1,4

# APA reporting of data

they posted. The control group ($M = 36.6$, $SD = 9.91$) and individual preparation-no script ($M = 30.86$, $SD = 9.26$) posted more, followed by argumentation scripts ($M = 19.19$, $SD = 6.84$) and the combination condition ($M = 15.26$, $SD = 6.11$). To take these differences

Tsovaltzi, D., Judele, R., Puhl, T., & Weinberger, A. (2017). Leveraging social networking sites for knowledge co-construction: Positive effects of argumentation structure, but premature knowledge consolidation after individual preparation. *Learning and Instruction*, *52*, 161-179.

*Table 1. Formal argumentative dimension by experimental group: Mean percentages and standard deviations of grounds and counterarguments.*

| | *Grounds* | | *Counterarguments* | |
|---|---|---|---|---|
| *Experimental group* | *M* | *SD* | *M* | *SD* |
| Control group | 12.08 % | 11.48 | 2.46 % | 3.67 |
| Script for the construction of single arguments | 33.80 % | 11.19 | 5.36 % | 8.07 |
| Script for the construction of argumentation sequences | 16.36 % | 17.78 | 5.99 % | 3.95 |
| Combined condition | 30.64 % | 6.10 | 13.00 % | 6.59 |

# APA reporting of data

they posted. The control group ($M = 36.6$, $SD = 9.91$) and individual preparation-no script ($M = 30.86$, $SD = 9.26$) posted more, followed by argumentation scripts ($M = 19.19$, $SD = 6.84$) and the combination condition ($M = 15.26$, $SD = 6.11$). To take these differences

Tsovaltzi, D., Judele, R., Puhl, T., & Weinberger, A. (2017). Leveraging social networking sites for knowledge co-construction: Positive effects of argumentation structure, but premature knowledge consolidation after individual preparation. *Learning and Instruction*, *52*, 161-179.

*Table 1. Formal argumentative dimension by experimental group: Mean percentages and standard deviations of grounds and counterarguments.*

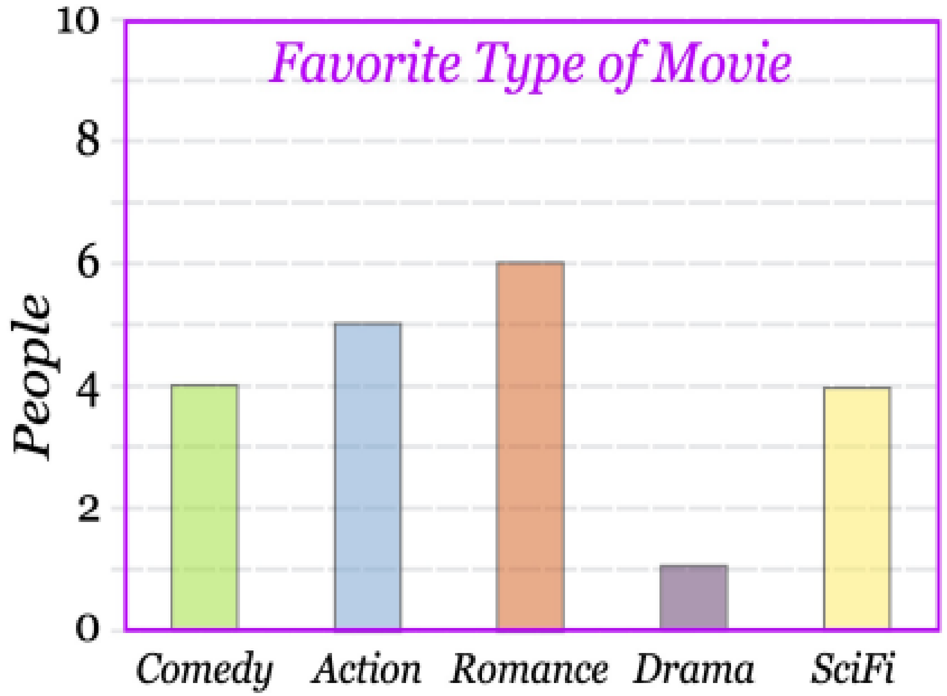
| | *Grounds* | | *Counterarguments* | |
|---|---|---|---|---|
| *Experimental group* | *M* | *SD* | *M* | *SD* |
| Control group | 12.08 % | 11.48 | 2.46 % | 3.67 |
| Script for the construction of single arguments | 33.80 % | 11.19 | 5.36 % | 8.07 |
| Script for the construction of argumentation sequences | 16.36 % | 17.78 | 5.99 % | 3.95 |
| Combined condition | 30.64 % | 6.10 | 13.00 % | 6.59 |

Weinberger, A., Stegmann, K., & Fischer, F. (2005). Computer-Supported Collaborative Learning in Higher Education : Scripts for Argumentative Knowledge Construction in Distributed Groups. In The Next 10 Years! Proceedings of the 2005 Conference on Computer Support for Collaborative Learning, CSCL ’05 (pp. 717–726).

# TYPES OF GRAPHS

# Types of graphs- Bar Graph

- Divided into different categories
- Shows amount in each category
- Categorical data on the X axis



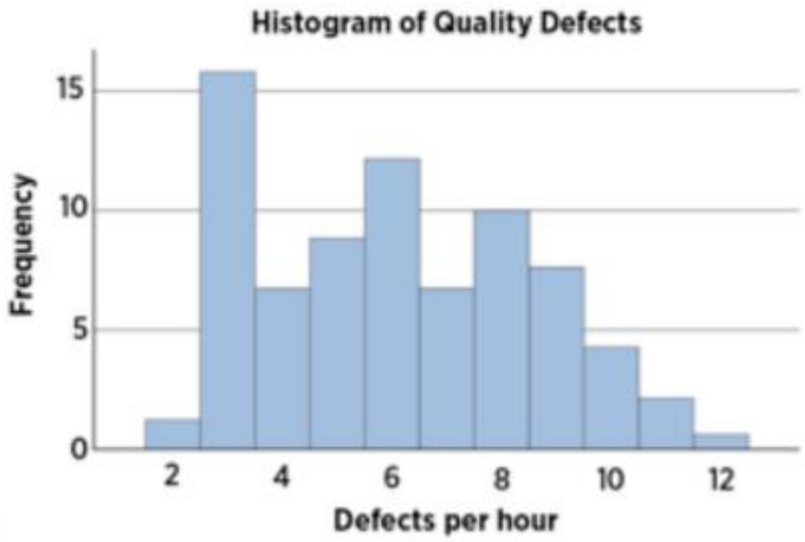
# Types of graphs- Histograms

Use a histogram when:

The data are numerical (interval or ratio)

You want to see the shape of the data's distribution, especially when determining whether the output of a process is distributed approximately normally

Helpful when wanting to communicate the distribution of data quickly and easily to others

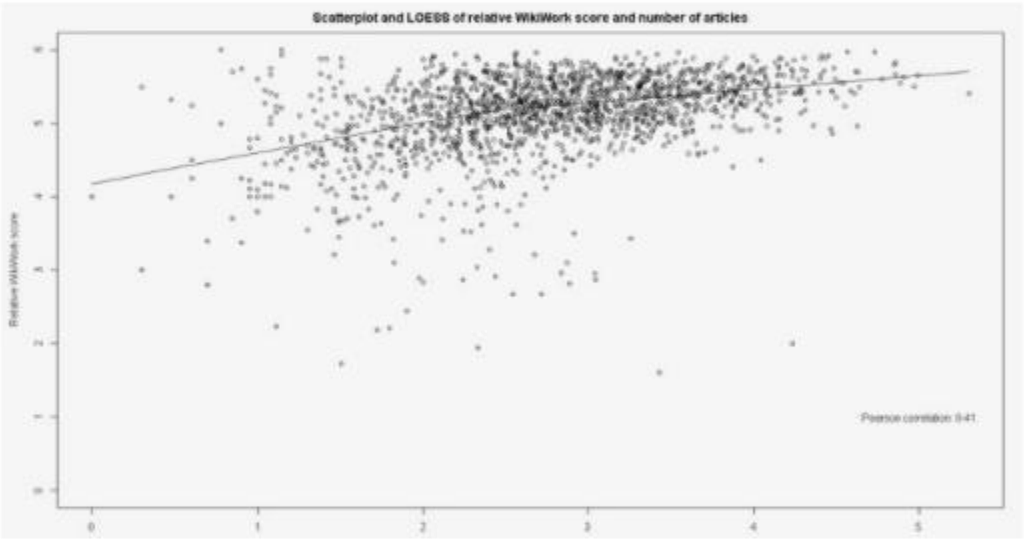


# Types of graphs- Line graph

- Graph shows data on a line
- Demonstrates trends or numbers that are interconnected

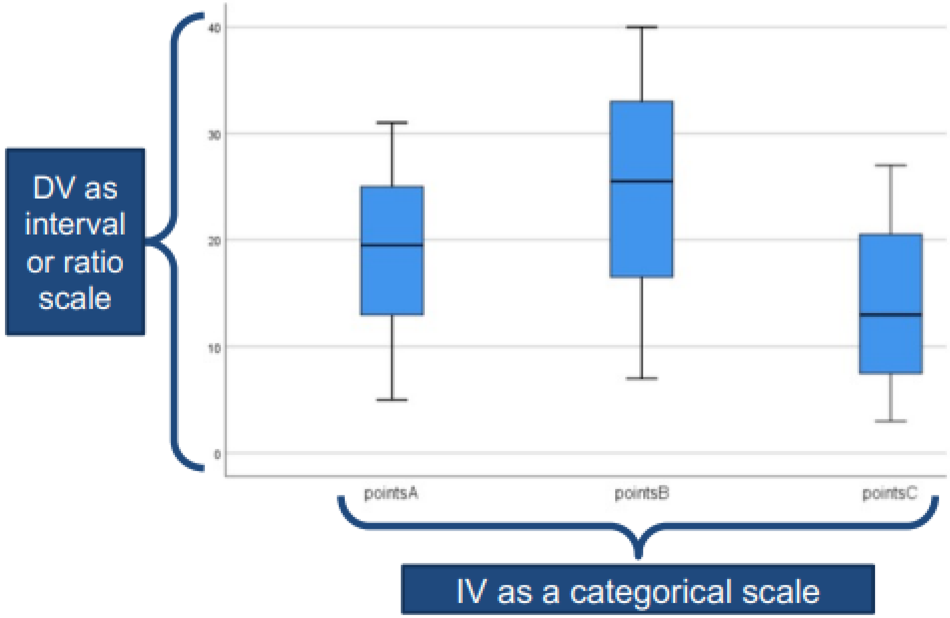
# Types of graphs- Scatterplot

- Displays data as dots that are scattered across the page
- Helps to uncover overall trends among variables and outliers



# Types of graphs- Boxplot

- The boxplot is another way to show how data is distributed
- Not recommended when having less than 20 data points per group
- Preferably to be used for visualizing multiple distributions at once.
- Also useful when having categorical grouping variable and a continuous outcome variable



# Types of graphs- Boxplot

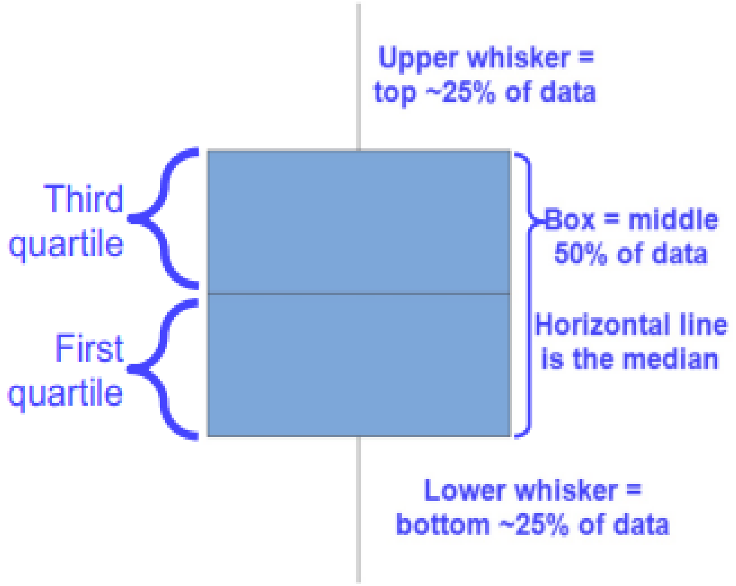
Maximum: a.k.a. $Q_4$ or $100^{th}$ percentile. Highest data point of the dataset excluding outliers. Often covered by the upper whisker

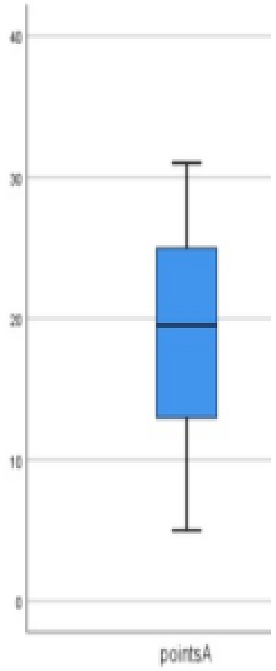
Third quartile: a.k.a. $Q_3$, $75^{th}$ percentile or upper quartile. Median of the upper half of the dataset

Median: a.k.a. $Q_2$ or $50^{th}$ percentile. Middle value of the dataset

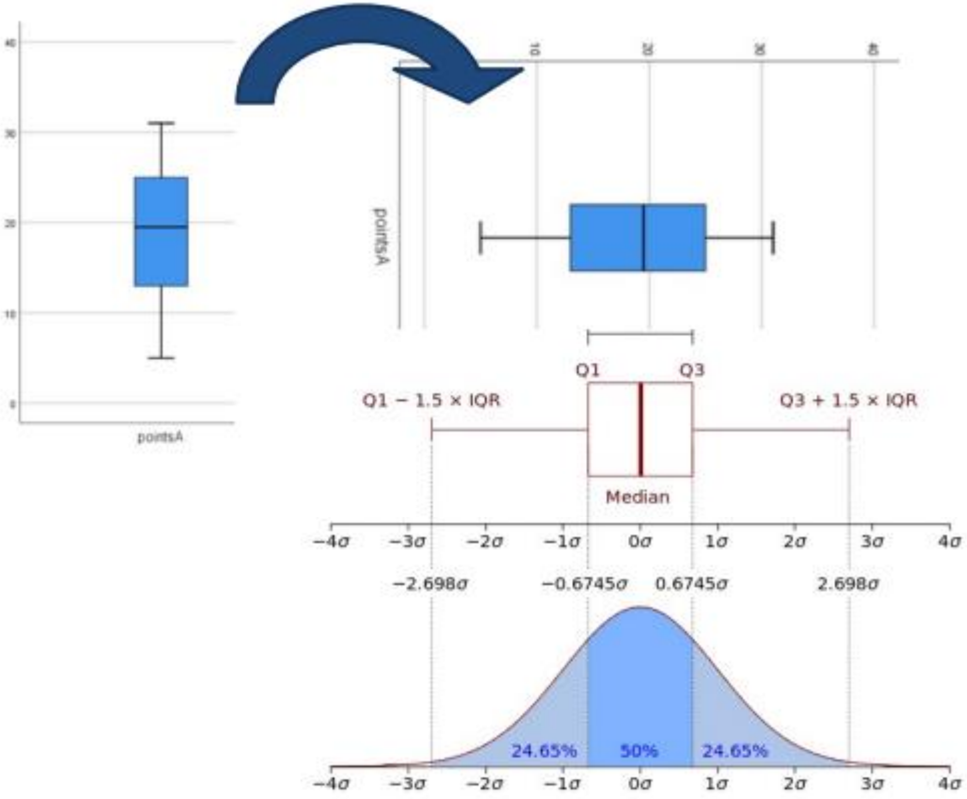
First quartile: a.k.a. $Q_1$, $25^{th}$ percentile or lower quartile. Median of the lower half of the dataset

Minimum: a.k.a. $Q_0$ or $0^{th}$ percentile. Lowest data point excluding outliers. Often covered by the lower whisker



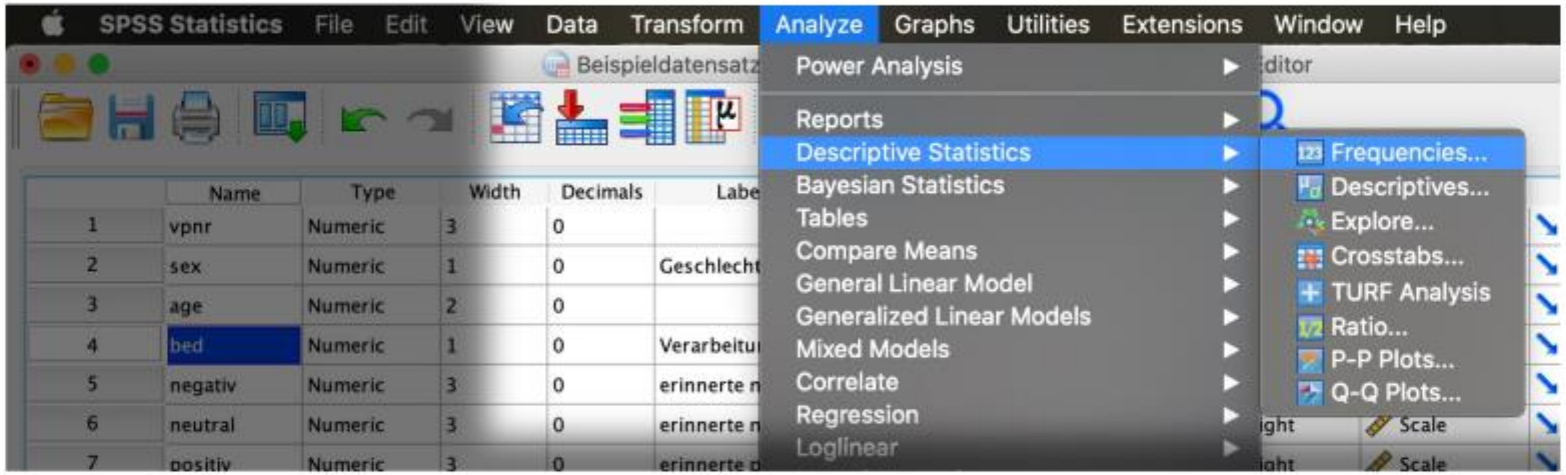
# Types of graphs- Boxplot



# How to calculate descriptive statistics in SPSS

In the main window, go to:

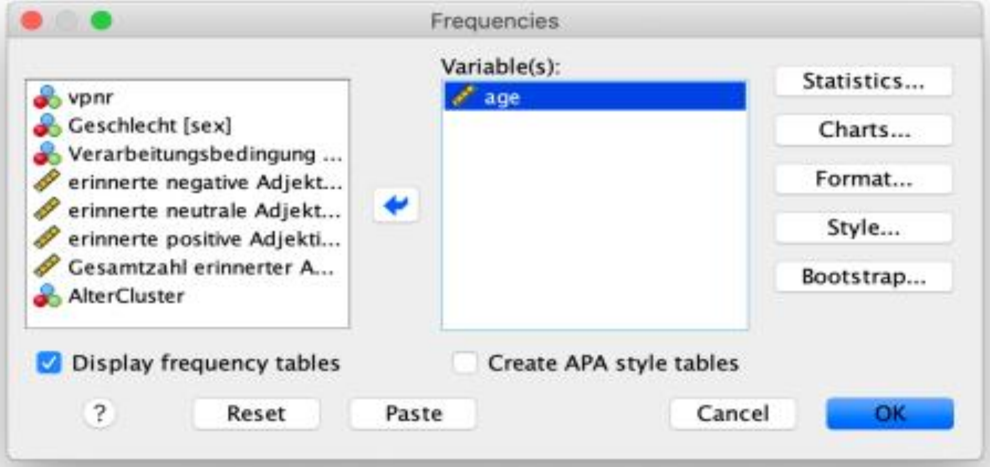
*Analyze* > *Descriptive Statistics* > Frequencies…



# How to calculate descriptive statistics in SPSS

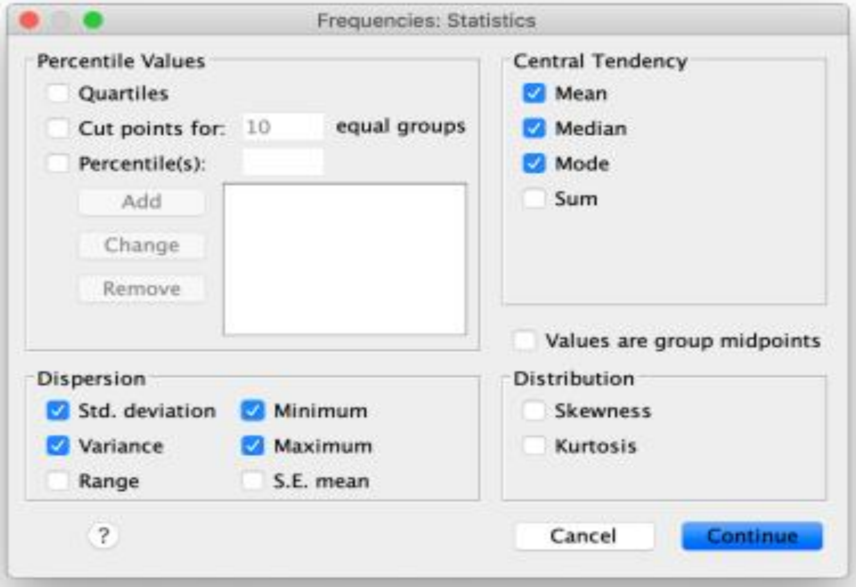
In the “Frequencies” window, move the variables you want to the “Variable(s)” box.

Afterwards, click on Statistics...

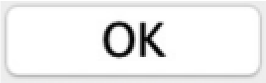


# How to calculate descriptive statistics in SPSS

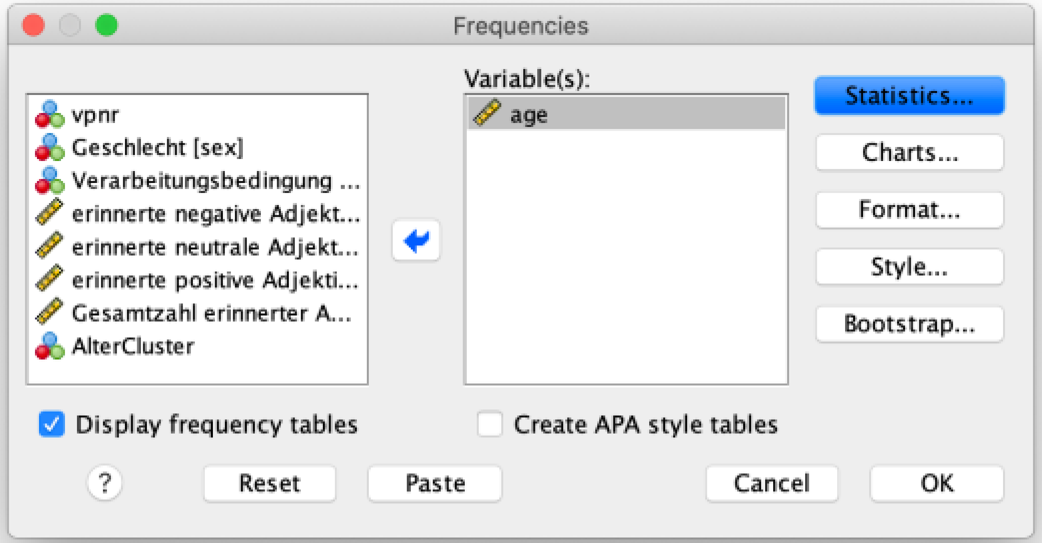
Once in the “Frequencies: Statistics” window, mark the checkboxes for the options shown and click on Continue



# How to calculate descriptive statistics in SPSS

Once back in the “Frequencies” window, click on 





# How to calculate descriptive statistics in SPSS

Results are shown in the Output Window

**Statistics**

age

| | | |
|---|---|---|
| N | Valid | 150 |
| | Missing | 0 |
| Mean | | 21.87 |
| Median | | 21.00 |
| Mode | | 20 |
| Std. Deviation | | 3.123 |
| Variance | | 9.754 |
| Minimum | | 18 |
| Maximum | | 42 |

**age**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 18 | 1 | .7 | .7 | .7 |
| | 19 | 24 | 16.0 | 16.0 | 16.7 |
| | 20 | 39 | 26.0 | 26.0 | 42.7 |
| | 21 | 22 | 14.7 | 14.7 | 57.3 |
| | 22 | 19 | 12.7 | 12.7 | 70.0 |
| | 23 | 12 | 8.0 | 8.0 | 78.0 |
| | 24 | 9 | 6.0 | 6.0 | 84.0 |
| | 25 | 8 | 5.3 | 5.3 | 89.3 |
| | 26 | 7 | 4.7 | 4.7 | 94.0 |
| | 27 | 3 | 2.0 | 2.0 | 96.0 |
| | 28 | 2 | 1.3 | 1.3 | 97.3 |
| | 29 | 2 | 1.3 | 1.3 | 98.7 |
| | 35 | 1 | .7 | .7 | 99.3 |
| | 42 | 1 | .7 | .7 | 100.0 |
| | Total | 150 | 100.0 | 100.0 | |

# How to calculate graphs on SPSS

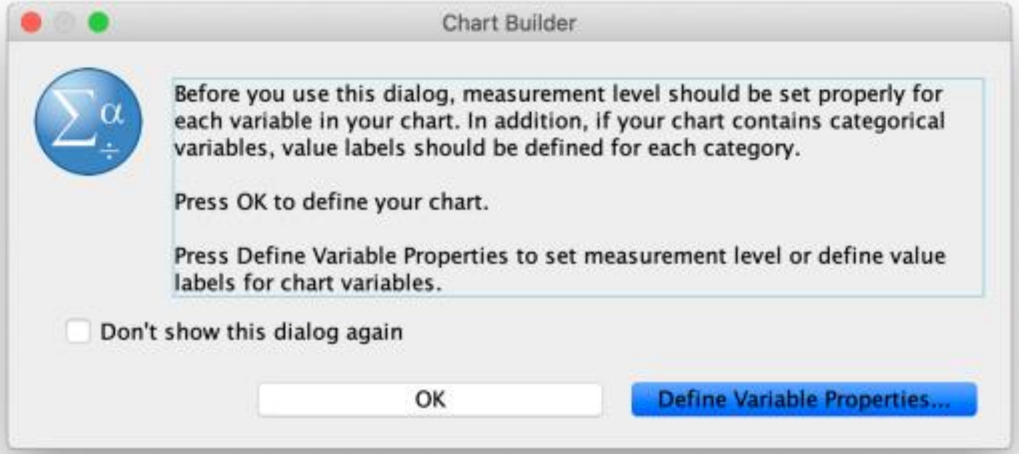
For this example we will create a histogram!

In the main window, click go to:

*Graphs > Chart Builder*

# How to calculate graphs on SPSS

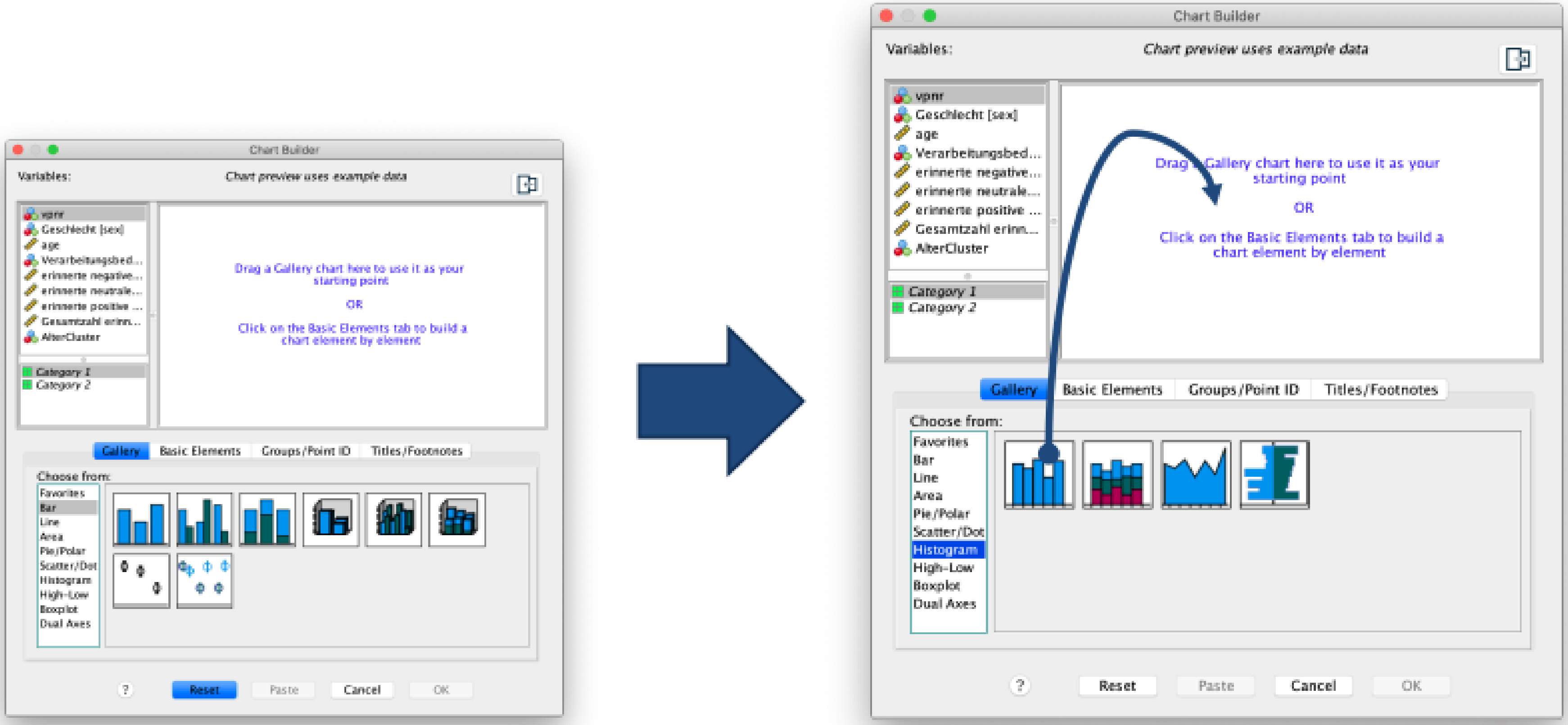
You may get this pop-up window reminding you that you need to set up your variable types correctly as well as labelling your nominal variables. Otherwise, your chart/graph won't be displayed correctly. After you have ensured your data set is correctly set-up, click on "OK"



# How to calculate graphs on SPSS

1. Once in the “Chart Builder” window, go to the Gallery tab, and from there select the “Histogram” option.

2. Click and hold the first graph icon, dragging it to the box above.

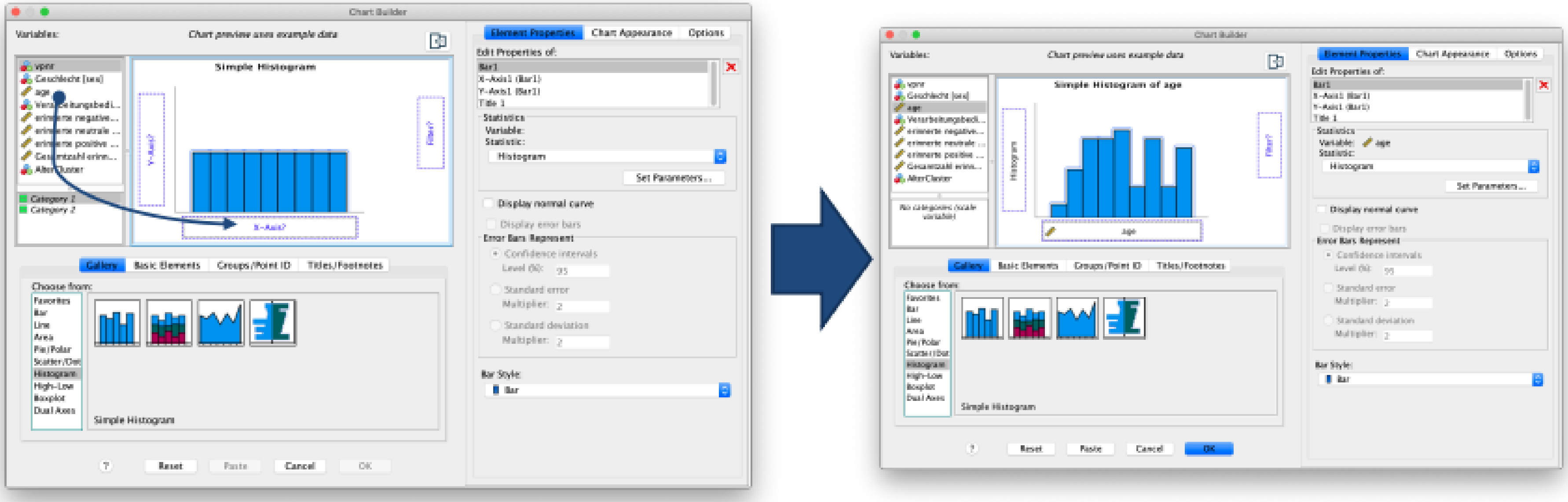


# How to calculate graphs on SPSS

1. Now with the histogram template in the box, select a variable and drag it to the "X-Axis?" box

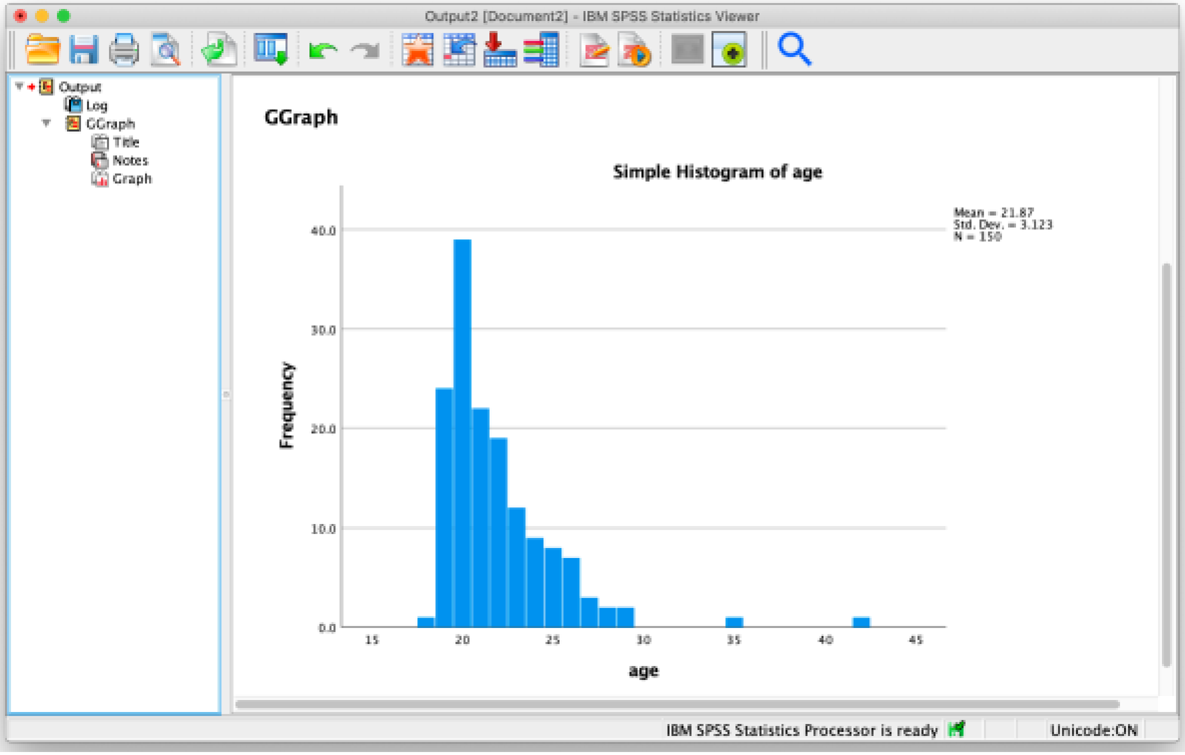
2. After that, you will notice that the "Y-Axis?" auto-fills and now says "Histogram"

3. Click on OK



# How to calculate graphs on SPSS

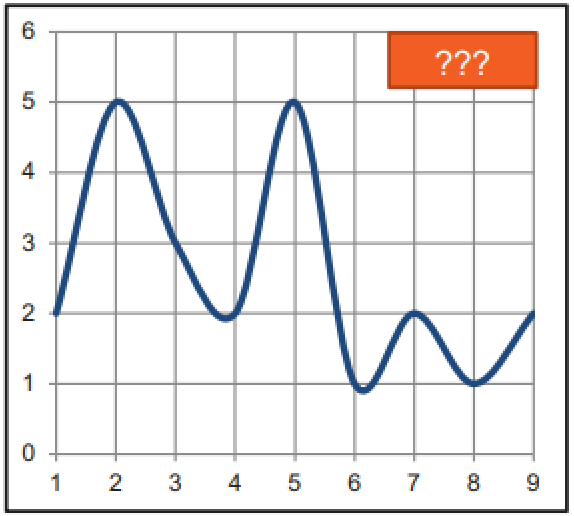
You will find the histogram in the Output Window

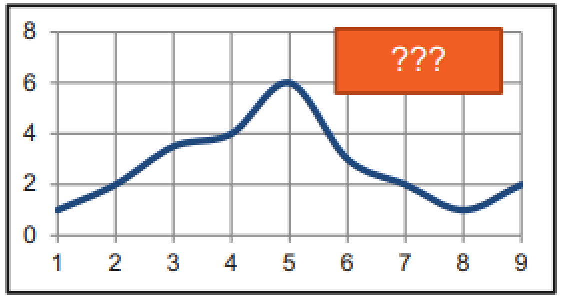


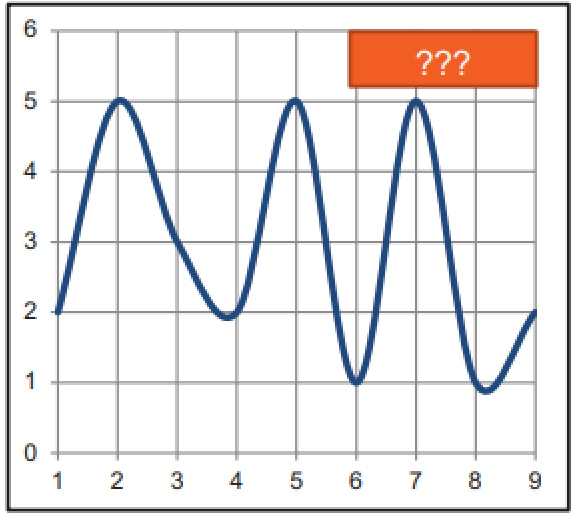
# Self-paced Activity 2: Which values are significant?

1. p = .50
2. p = .03
3. p = .27
4. p = .006
5. p = .051

# Self-paced Activity 3: Modes





**(uni-)modal, bi-modal, or multi-modal???**

**REMAINING QUESTIONS?**

**Contact me via e-mail, in teams or come to my office hours**