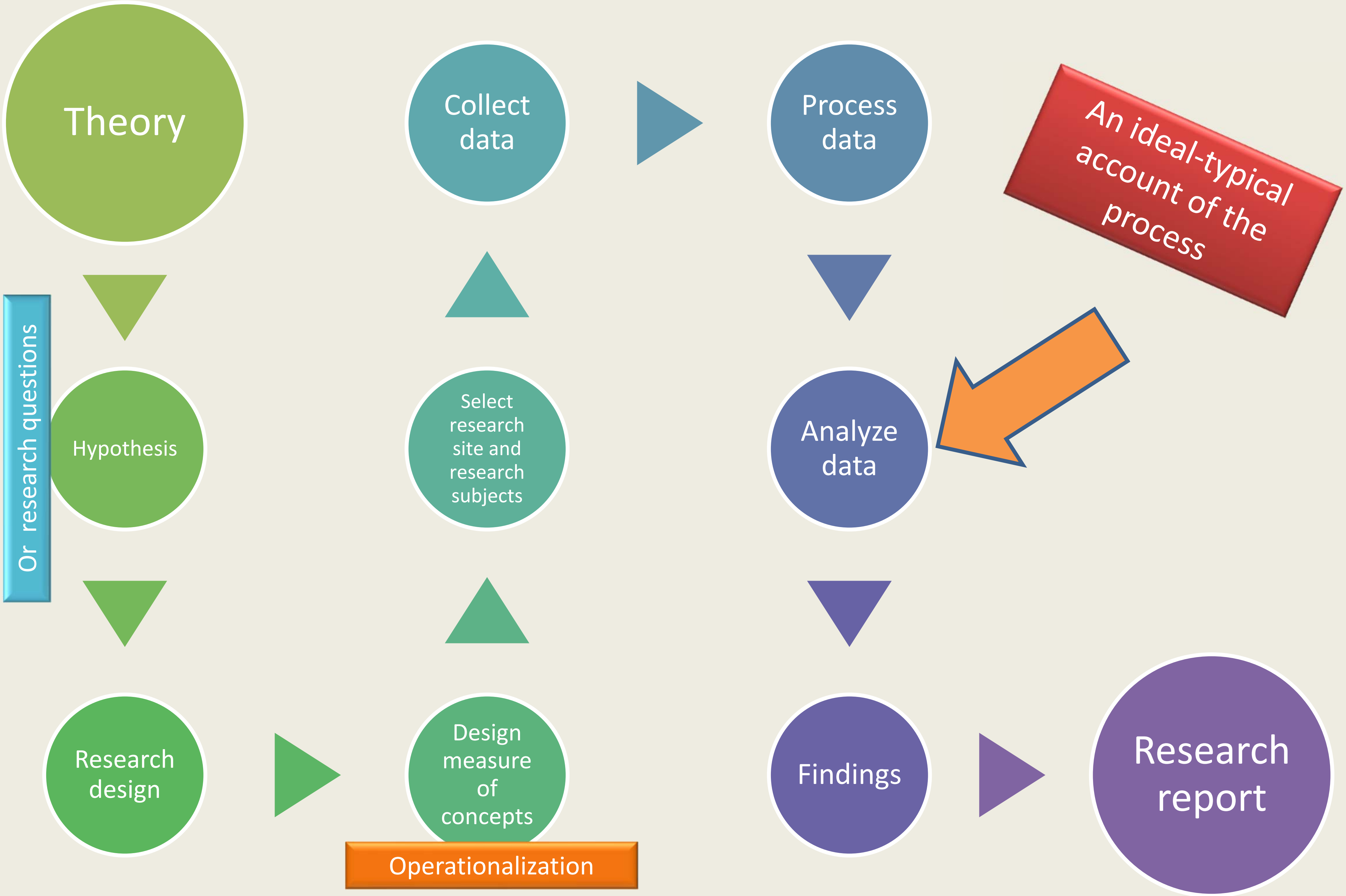
# SPSS: Data analysis
# Frequencies, measures of central tendency, measures of variability, charts

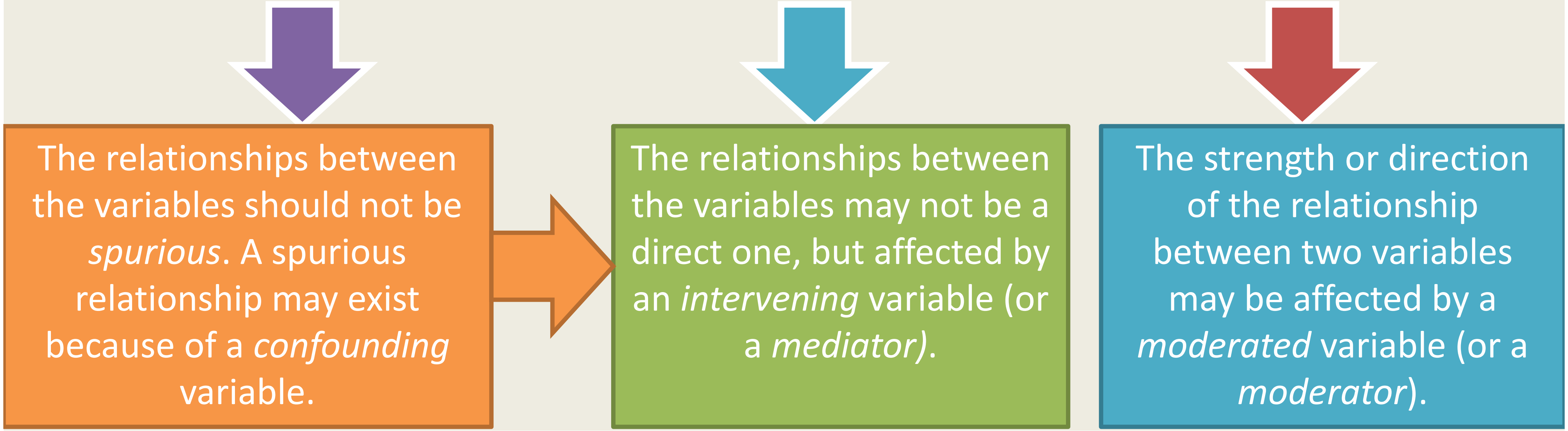
# Quantitative research process



# Types of analysis

- *Univariate analysis*: the analysis of one variable at a time.

- *Bivariate analysis*: the analysis of two variables at a time in order to uncover whether or not the two variables are related.
- *Multivariate analysis*: the analysis of more than two variables at the same time.



The relationships between the variables should not be *spurious*. A spurious relationship may exist because of a *confounding* variable.

The relationships between the variables may not be a direct one, but affected by an *intervening* variable (or a *mediator)*.

The strength or direction of the relationship between two variables may be affected by a *moderated* variable (or a *moderator*).

# Univariate analysis

## **Frequencies**

A frequency table provides the number of people and the percentage belonging to each of the categories for the variable in question.

It can be used in relation to all of the different types of variable.

(Analyze → Descriptive statistics → Frequencies)

## Example of frequency table of a categorical variable

| Gender | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Man | 42 | 46,7 | 46,7 | 46,7 |
| | Woman | 48 | 53,3 | 53,3 | 100,0 |
| | Total | 90 | 100,0 | 100,0 | |

**Valid Percent:**
When missing data are excluded from the calculations.

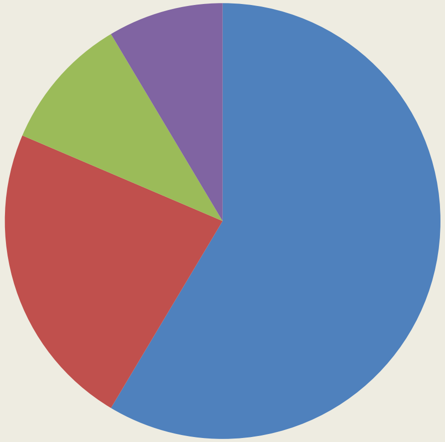
**Example of frequency table of a interval/ratio variable following value grouping**

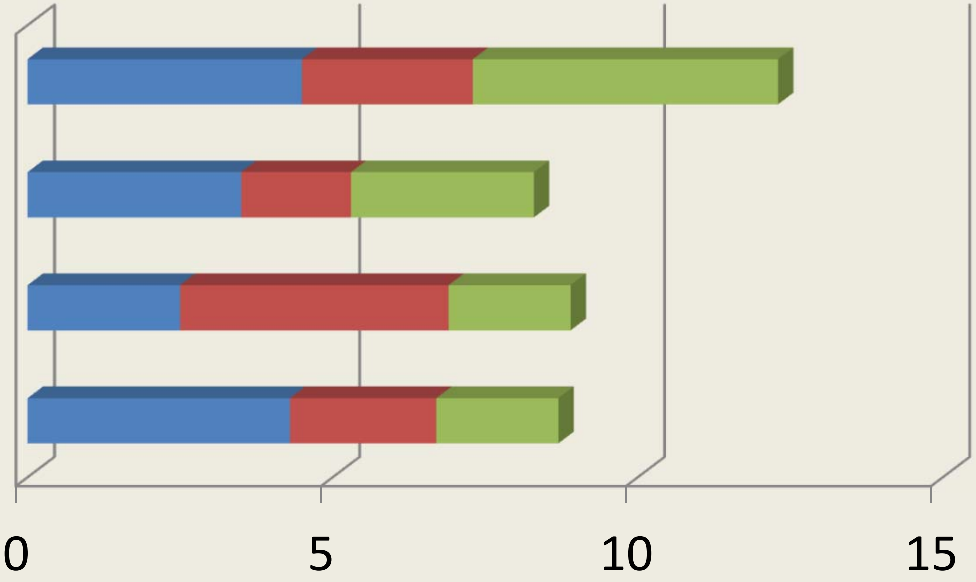
| Groups Cardiovascular exercise | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Less than 20 min | 20 | 22.2 | 22.2 | 22.2 |
| | 21 min to 40 min | 63 | 70.0 | 70.0 | 92.2 |
| | More than 41 min | 7 | 7.8 | 7.8 | 100.0 |
| | Total | 90 | 100.0 | 100.0 | |

# Diagrams - Graphs

*For nominal and ordinal variables:*

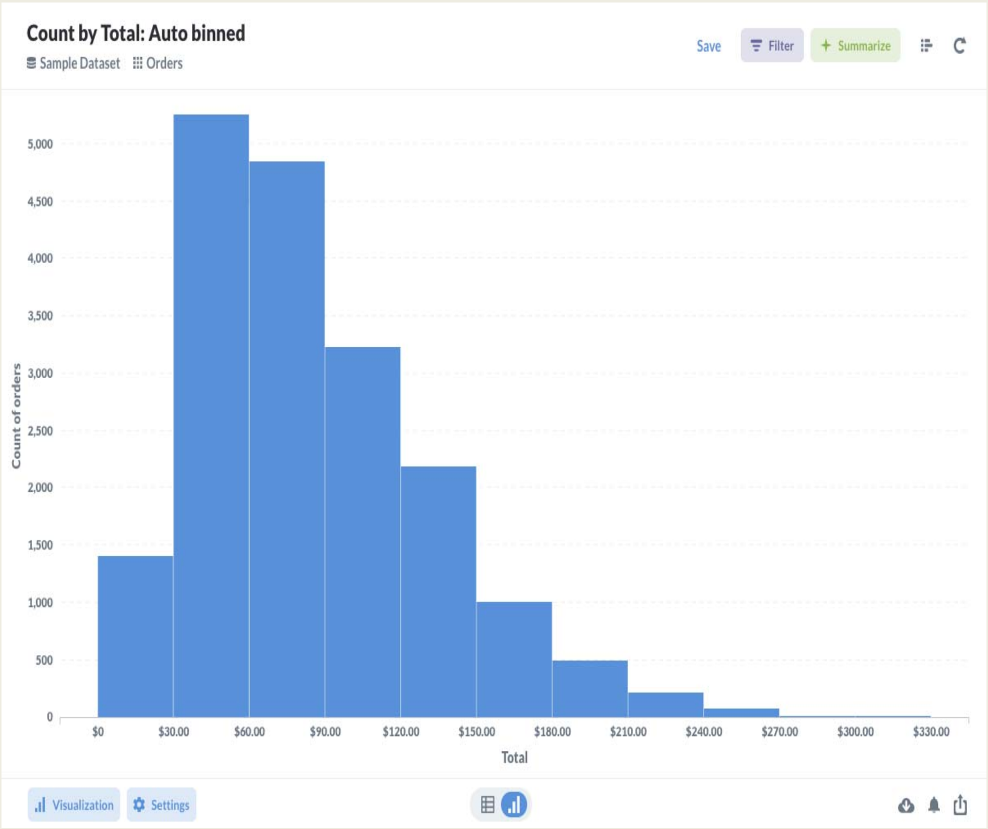
- Bar chart
- Pie chart





*For interval/ratio variables:*

- Histogram


Count by Total: Auto binned

***To produce a pie chart:***

Graphs→ Chart Builder→ Pie/Polar→ Choose from (double click or drag and drop your selection into the main box) → select the variable from *Variable* Box and drag and drop the variable into the area "*Slice by?" which is* marked blue.

***To produce a bar chart:***

Graphs→ Chart Builder→ Bar → Choose from (double click or drag and drop your selection into the main box) → select the variable from *Variable* Box and drag and drop the variable into the area "*X-Axis?" which is* marked blue.

***To produce a histogram:***

Graphs→ Chart Builder→ Histogram→ Choose from (double click or drag and drop your selection into the main box) → select the variable from *Variable* Box and drag and drop the variable into the area *"X-Axis?" which is* marked blue.

# Measures of central tendency

(Analyze→ Descriptive Statistics → Descriptive)

***Measures of central tendency capture in one figure a value that is typical for a distribution of values.***

***Arithmetic mean ( $\overline{X}$, M,μ):*** The arithmetic average which is calculated by summing up a group of numbers and dividing that sum by the sample size.

$$\text{Mean} = \frac{\sum X}{n}$$

The mean is vulnerable to **outliers**, that is extreme values. The mean can be increased or decreased due to outliers and this may affect its reliability as a measure of central tendency.

*Mean can be employed only in relation to interval/ratio variables.*

***Median:* The mid-point in a distribution of values.**

Odd numbers: We derive the median by arraying all the values of a distribution from the smallest to the largest and locate the middle point.

10,14,15,15,16,16,17,17,17,17,17,20,21,22,23

Even numbers: We derive the median by arraying all the values of a distribution from the smallest to the largest, locate the two middle points, add the points and divide by two (2).

10,14,15,15,16,16,17,17,17,17,17,20,21,22,23,25

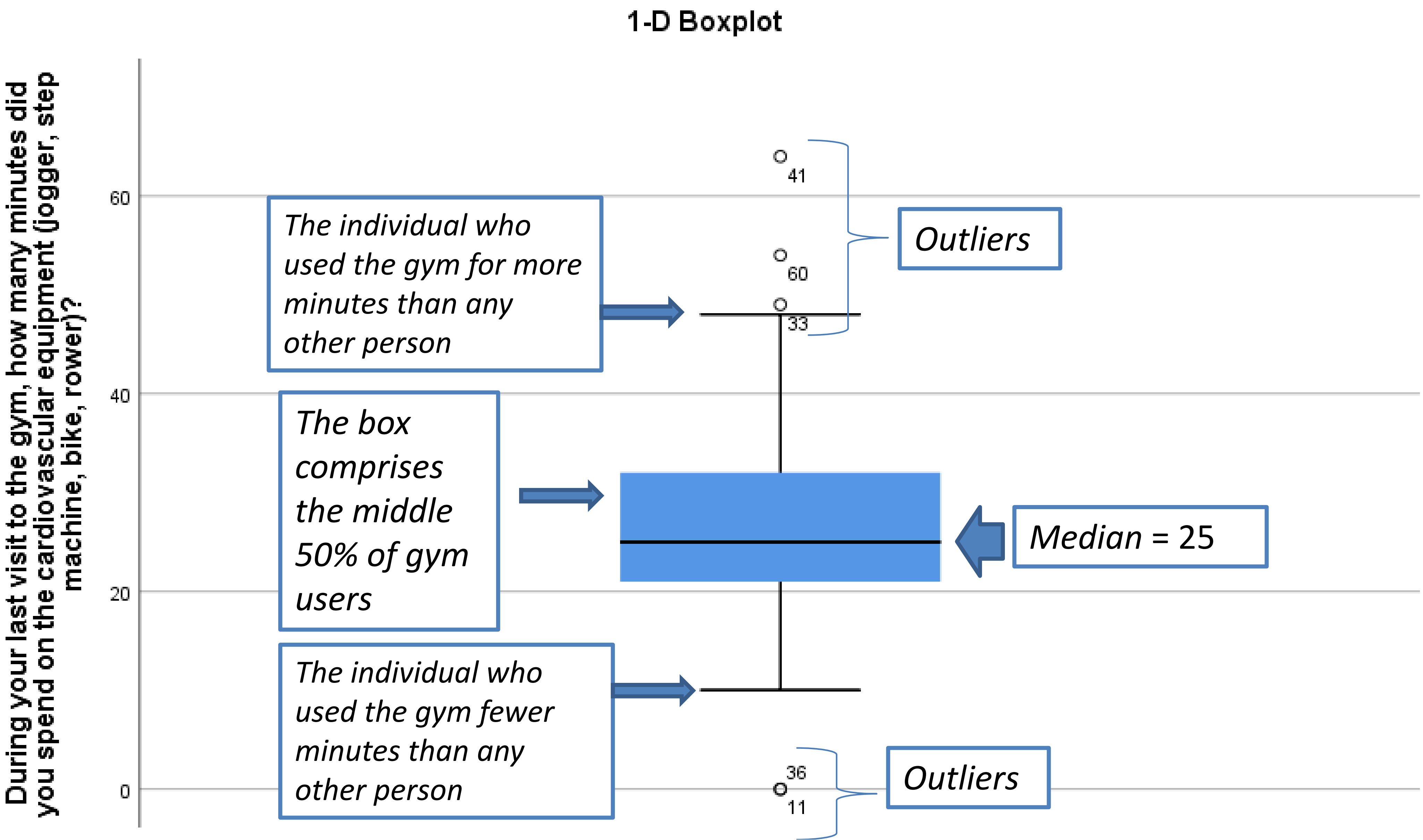
17+17= 34 34/2=**17**

**Median = 17**

*Median can be employed in relation to interval/ratio variables and ordinal variables.*

The median is not affected by the outliers.

A ***boxplot*** is a popular figure for displaying interval/ratio variables regarding central tendency and dispersion.
(Graphs→ Chart Builder→ Boxplot →*Y Axis?*)



*The individual who used the gym for more minutes than any other person*

*The box comprises the middle 50% of gym users*

*The individual who used the gym fewer minutes than any other person*

*Outliers*

*Median* = 25

*Outliers*

## *Mode (Mo):* Is the value that occurs most frequently in a distribution.

10,14,15,15,16,16,17,17,17,17,17,20,21,22,23

**17 is the most frequently occurring number because it occurs 5 times.**

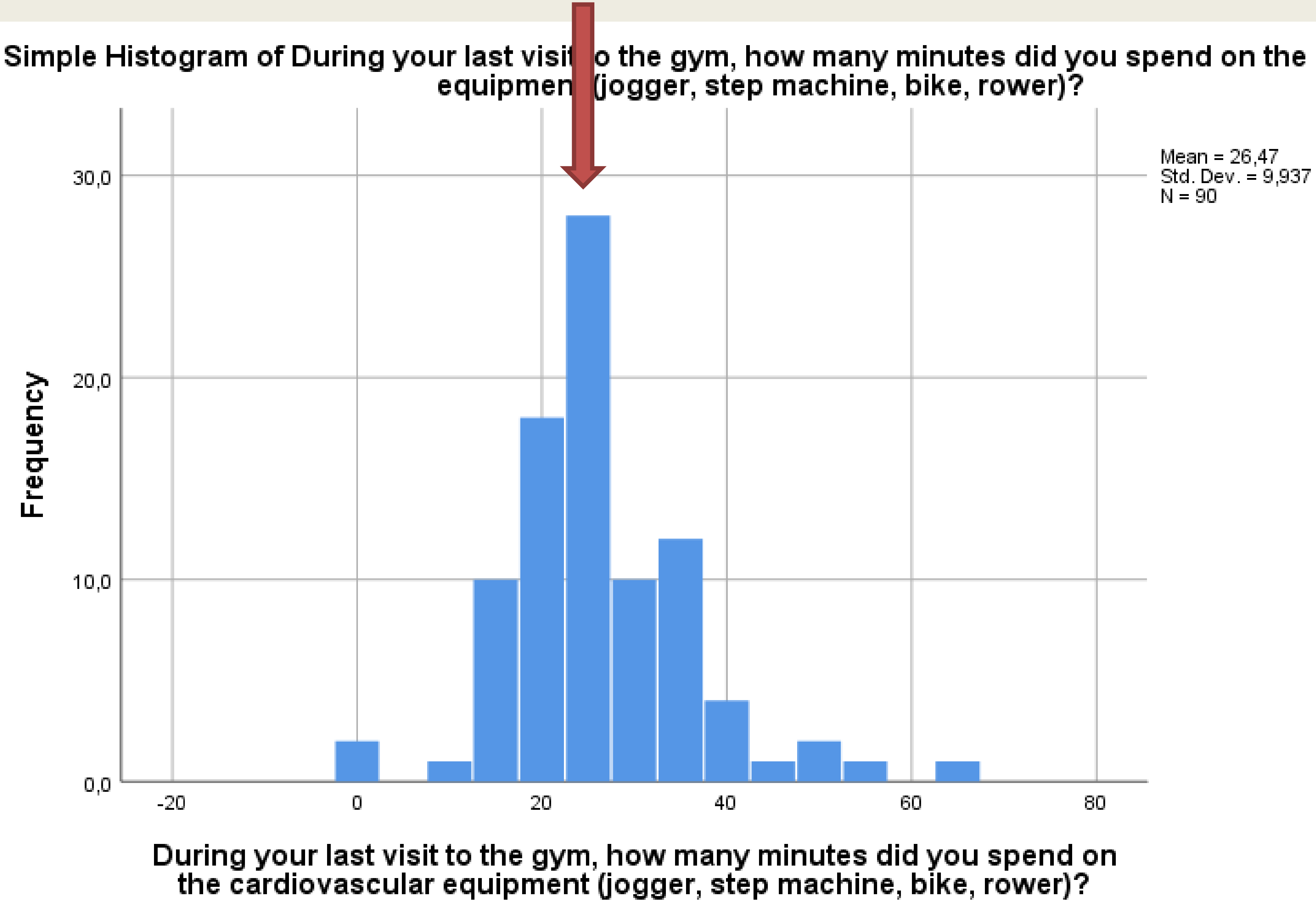
15,15,16,16,17,17,17,20,21,22,22,22,23

**Bimodal data:** When there is a tie for the most frequently occurring number. We report both modes.

**Implications of bimodal data:** The data set represents at least two different types of individuals. We rarely use the mode measure in social sciences.

Bimodal Distribution

*Mo*= 22,27



# Measures of variability

(Analyze→ Descriptive Statistics → Descriptive)

***Numerical indexes that provide information about how spread out or how much variation is present in a variable. Often called measures of dispersion***

***Range:*** **The difference between the maximum and the minimum value in a distribution.** Associated only with interval/ratio variables. The range is also influenced by outliers.

*Range = Highest value – Lowest value*

10,14,15,15,16,16,17,17,17,17,17,20,21,22,23

23-10 = 13

**Range = 13**

***Variance* (σ²) *:*** The average deviation of data values from their mean in squared units.

$$\text{Variance} = \frac{\sum (X - \bar{X})^2}{n}$$

***Standard deviation (SD,σ) :* Is an approximate indicator of the average distance that your data values are from the mean.** Is a more meaningful way of interpreting and understanding variance.

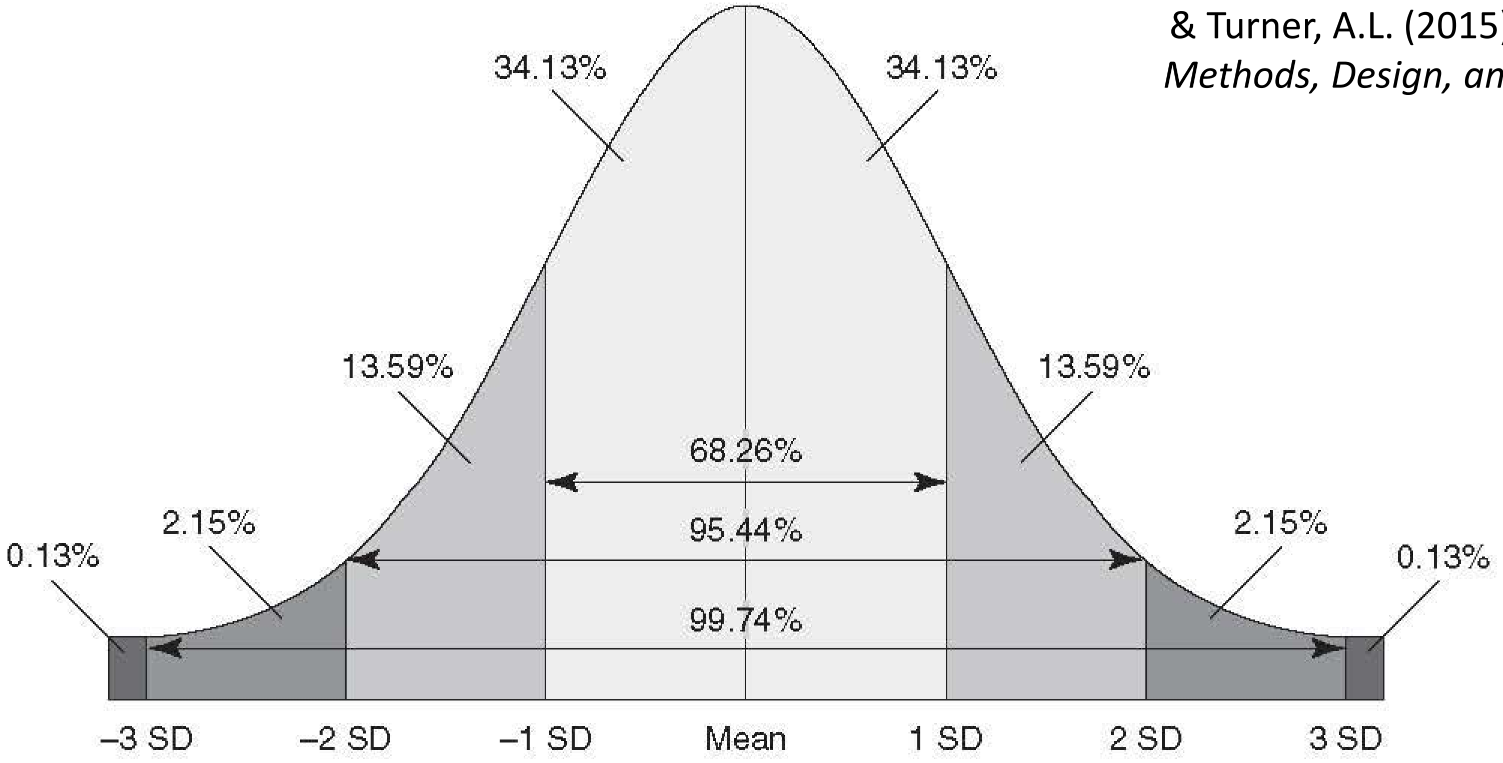
$$SD = \sqrt{\text{variance}}$$

For the variance and the standard deviation, the larger the value, the greater the data are spread out; the smaller the value, the less the data are spread out

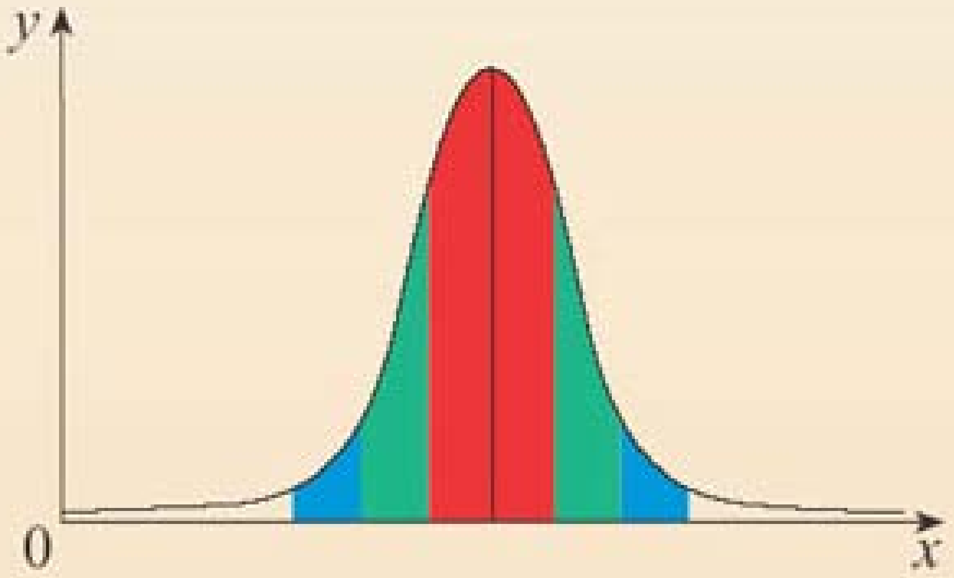
***Normal distribution and the "68,95,99.7" percent rule:*** If the data are normally distributed, 68% of the cases fall within one standard deviation from the mean, 95% fall within two standard deviations, and 99.7% fall within three standard deviations.

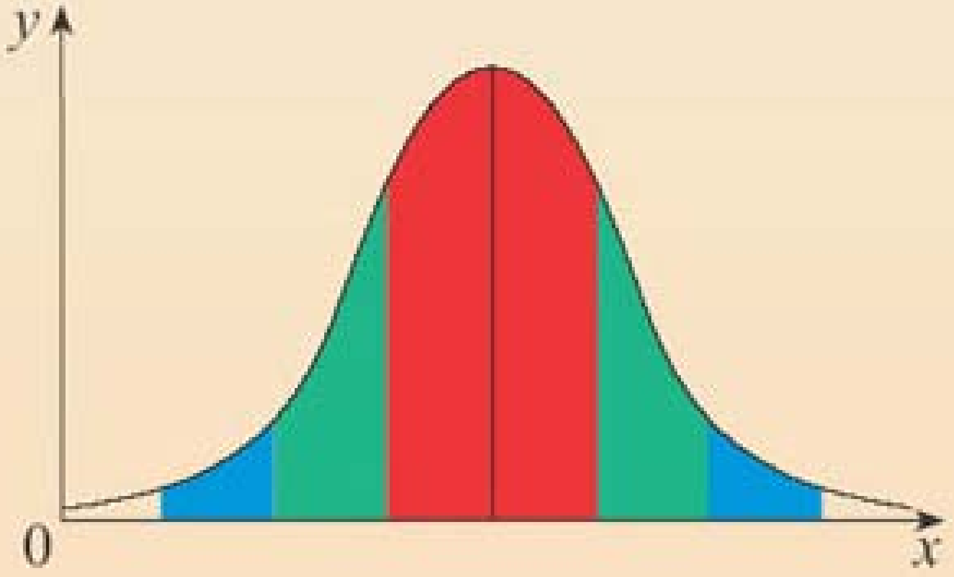
Christensen, L.B., Burke Johnson, R. & Turner, A.L. (2015). *Research Methods, Design, and Analysis.* p.408

FIGURE 14.8
Areas under the normal distribution.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *z* scores | –3 | –2 | –1 | 0 | 1 | 2 | 3 |
| Percentile ranks | 0.1 | 2 | 16 | 50 | 84 | 98 | 99.9 |





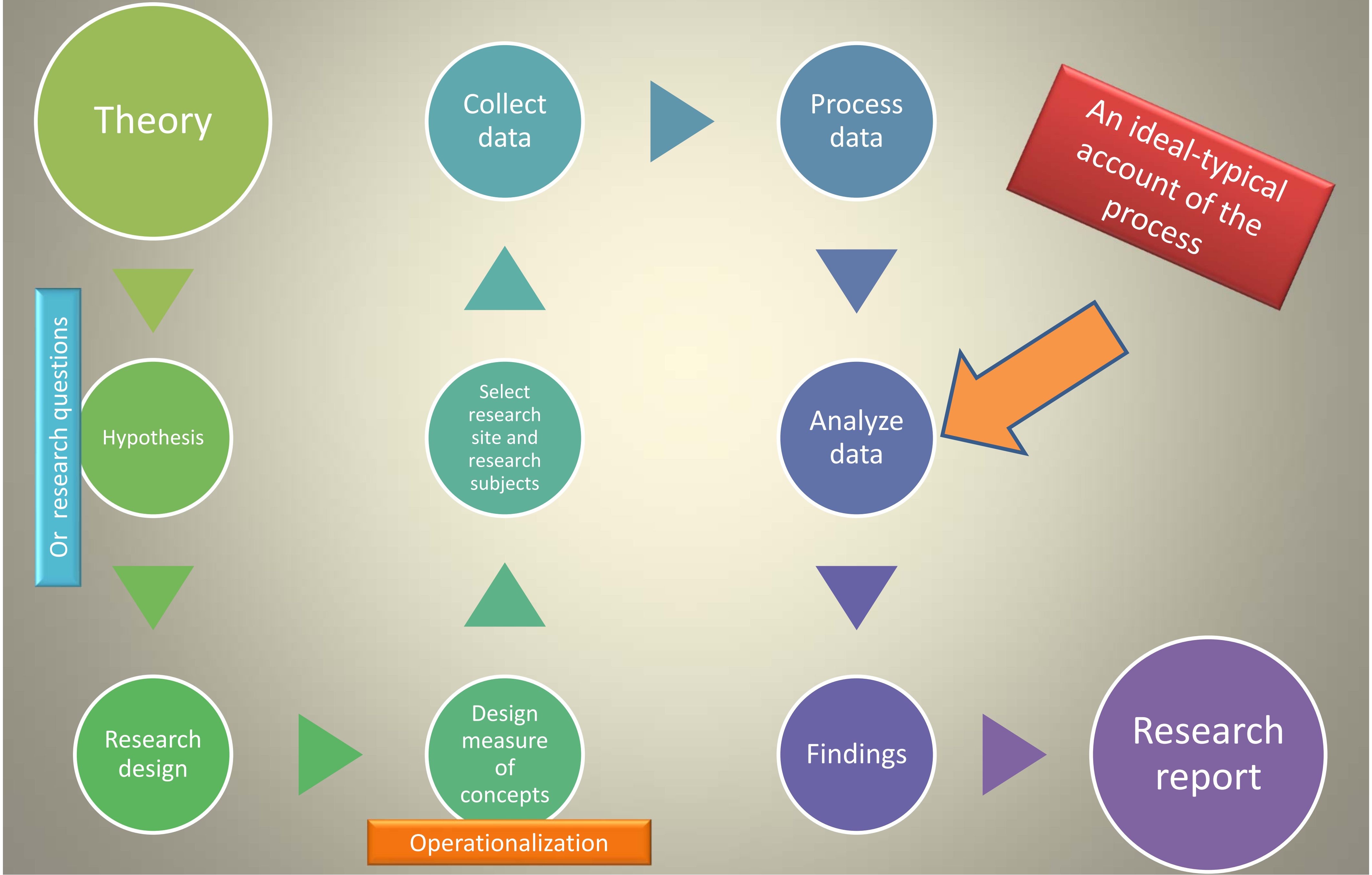
Top: Data showing a small standard deviation from the mean value.

Bottom: Data showing a large standard deviation from the mean value.

(Image: https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/using-numbers-and-handling-data/content-section-3.8.1)

# SPSS: Data analysis
# Chi-square, Correlation, T-test, ANOVA

# Quantitative research process



Theory

Hypothesis

Or research questions

Research design

Design measure of concepts

Operationalization

Select research site and research subjects

Collect data

Process data

Analyze data

Findings

Research report

An ideal-typical account of the process

# Statistical significance

In hypothesis testing, we are trying to determine whether the difference between the sample means should be viewed as random variation or as a real/true difference in the populations from which the data were selected.

*Alpha level (or level of significance, α):* The value at which one would reject the null hypothesis and accept the alternative (research) hypothesis. By convention, alpha is usually set at .05.

*Probability value (p):* The likelihood of the observed value of a statistic, if the null hypothesis were true.

*If the p value is less than (or equal to) .05, then reject the null hypothesis and accept (with caution) the alternative hypothesis*

**Statistical significance:**

Conclusion that an observed finding would be very unlikely if the null hypothesis were true

**Steps in Hypothesis Testing with Decision-Making rules** (Christensen, Johnson & Turner, 2015, p. 441)

Step 1. State the null and the alternative hypotheses.

Step 2. Set the alpha level (i.e., level of significance). (Psychologists usually set the alpha level at .05).

Step 3. Select the statistical test to be used (e.g., *t* test, ANOVA, regression analysis).

Step 4. Conduct the statistical test and obtain the *p* value.

Step 5. Compare *p* value to the alpha level (i.e., level of significance), and apply either decision rule 1 or decision rule 2.

**Decision Rule 1:**

If: *p* value is ≤ alpha level*

Then: Reject the null hypothesis and tentatively accept the alternative hypothesis.

Conclusion: The research finding is statistically significant.

**Decision Rule 2:**

If: *p* value is > alpha level

Then: Fail to reject the null hypothesis.

Conclusion: The research finding is not statistically significant.

Discovering Statistics with Andy Field: https://www.facebook.com/photo?fbid=685349330291038&set=a.402992175193423

Step 6. Compute effect size, interpret findings, and make judgment of practical significance of results.

*The issue of what to do when *p*=alpha is a matter of some controversy. We recommend the convention provided by the late Jacob Cohen that a *p* value of .00 to .050 is sufficiently small to reject the null, but values of .051to1.00 are not sufficiently small to reject the null. For example, using Cohen's rule .0504 would be statistically significant because it rounds down to .05, but .0505 is not because it rounds up to .051.

# Contingency tables

- Also called cross-tabulation
- A contingency table is like a frequency table but it allows two variables to be simultaneously analyzed so that relationships between the two variables can be examined
- Contingency tables are probably the most flexible of all methods of analyzing relationships in that they can be employed to any pair of variables. However, they are not the most efficient method for some pairs.

# Chi-square test - $X^2$

*Is used to determine whether a relationship observed in a contingency table is statistically significant*

$x^2$ (df,n)= value of chi-square, $p$=…

Sample size

Degrees of freedom

Analyze → Descriptive statistics → Crosstabs

Statistics tabs → Chi-square

Display clustered bar charts

**Chi-Square Tests**

| | Value | df | Asymptotic Significance (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 1,164[a] | 2 | ,559 |
| Likelihood Ratio | 1,545 | 2 | ,462 |
| Linear-by-Linear Association | ,004 | 1 | ,949 |
| N of Valid Cases | 90 | | |

a. 2 cells (33,3%) have expected count less than 5. The minimum expected count is ,47.

$$x^2\ (2, 90)= 1.164,\ p=.559$$

**When you go to th n, how often do you use the cardiovascular equipment (jogger, step hine, bike, rower)? * Gender Crosstabulation**

Count

| | | Gender | | Total |
|---|---|---|---|---|
| | | Man | Woman | |
| When you go to the gym, how often do you use the cardiovascular equipment (jogger, step machine, bike, rower)? | Always | 33 | 39 | 72 |
| | Usually | 9 | 8 | 17 |
| | Rarely | 0 | 1 | 1 |
| Total | | 42 | 48 | 90 |

One of the basic assumptions of chi-square test is to assume that the expected value of cells in the table should be 5 or greater in at least 80% of cells and that no cell should have an expected value less than 1.

When one of the cells has less than five observations, we should use the **Fisher's exact test**. The Fisher's exact test is used with small sample sizes.

Analyze → Descriptive statistics → Crosstabs

Statistics tabs → Chi-square

Exact → Exact (retain the test time as it is)

**Chi-Square Tests**

| | Value | df | Asymptotic Significance (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) | Point Probability |
|---|---|---|---|---|---|---|
| Pearson Chi-Square | 1,164[a] | 2 | ,559 | ,887 | | |
| Likelihood Ratio | 1,545 | 2 | ,462 | ,887 | | |
| Fisher's Exact Test | 1,128 | | | ,887 | | |
| Linear-by-Linear Association | ,004[b] | 1 | ,949 | 1,000 | ,568 | ,189 |
| N of Valid Cases | 90 | | | | | |

a. 2 cells (33,3%) have expected count less than 5. The minimum expected count is ,47.

b. The standardized statistic is -,064.

# *Reporting findings - APA*

*Reporting findings – APA:* Fisher’s exact test was used to determine if there was a significant association between gender and time spent at the gym on cardiovascular exercise. There was not a statistical significant association (*p*= .887) between men and women and the time they spent at the cardiovascular equipment.

# Correlation

*Describes the strength of an association between two variables and the direction of the relationship.* The dataset should include two or more **continuous numeric variables**, each defined as scale, which will be used in the analysis

For example: *The more time you spend on studying statistics the* more less *confused you are going to be.*

*The younger someone is, the less money he/she is likely to earn.*

*Correlation coefficient (r):* A numerical index indicating the strength and direction of a linear relationship between two quantitative variables.

The coefficient will almost certainly lie between 0 (zero or no relationship between the two variables) and 1 (a perfect relationship). This indicates the ***strength of a relationship***; The closer the coefficient is to 1, the stronger the relationship; the closer it is to 0, the weaker the relationship;

| Interval | Correlation |
|---|---|
| .00-.199 | Very weak |
| .20-.399 | Weak |
| .40-.599 | Medium |
| .60-.799 | Strong |
| .80-1.000 | Very strong |

The coefficient will be either positive or negative—this indicates the ***direction of a relationship***.

*Positive relationship*: as one variable increases, the other variable increases by the same amount (linear correlation)

*Negative relationship*: as one variable increases, the other variable decreases by the same amount (linear correlation)

Analyze → Correlate → Bivariate

**Correlate coefficients:**

*Pearson's r* (by default – interval/ratio variables, normal distribution, random/representative sample/linear relationship)

*Kendall's tau-b* (ordinal/scale variables, monotonic relationship)

*Spearman's rho* (At least one ordinal variable/ no normal distribution, monotonic relationship)

## Correlations

| ***r= -.041, p= .703*** | | Age | During your last visit to the gym, how many minutes did you spend on other activities (e.g. stretching exercises)? |
|---|---|---|---|
| Age | Pearson Correlation | 1 | -,041 |
| | Sig. (2-tailed) | | ,703 |
| | N | 90 | 90 |
| During your last visit to the gym, how many minutes did you spend on other activities (e.g. stretching exercises)? | Pearson Correlation | -,041 | 1 |
| | Sig. (2-tailed) | ,703 | |
| | N | 90 | 90 |

## Correlations

| ***r= -.228, p= .031*** | | Age | During your last visit to the gym, how many minutes did you spend on the weights machines (including free weights)? |
|---|---|---|---|
| Age | Pearson Correlation | 1 | -,228* |
| | Sig. (2-tailed) | | ,031 |
| | N | 90 | 90 |
| During your last visit to the gym, how many minutes did you spend on the weights machines (including free weights)? | Pearson Correlation | -,228* | 1 |
| | Sig. (2-tailed) | ,031 | |
| | N | 90 | 90 |

*. Correlation is significant at the 0.05 level (2-tailed).

# T-test

*It is used to determine if there is a significant difference between the means of two groups.*

***Independent samples T-test:*** It can be applied when there is at least one *independent categorical* variable (which consists of two groups) and one *dependent continuous variable* (scale).

Analyze→ Compare Means→ Independent samples t-test

Grouping variable: Independent variable

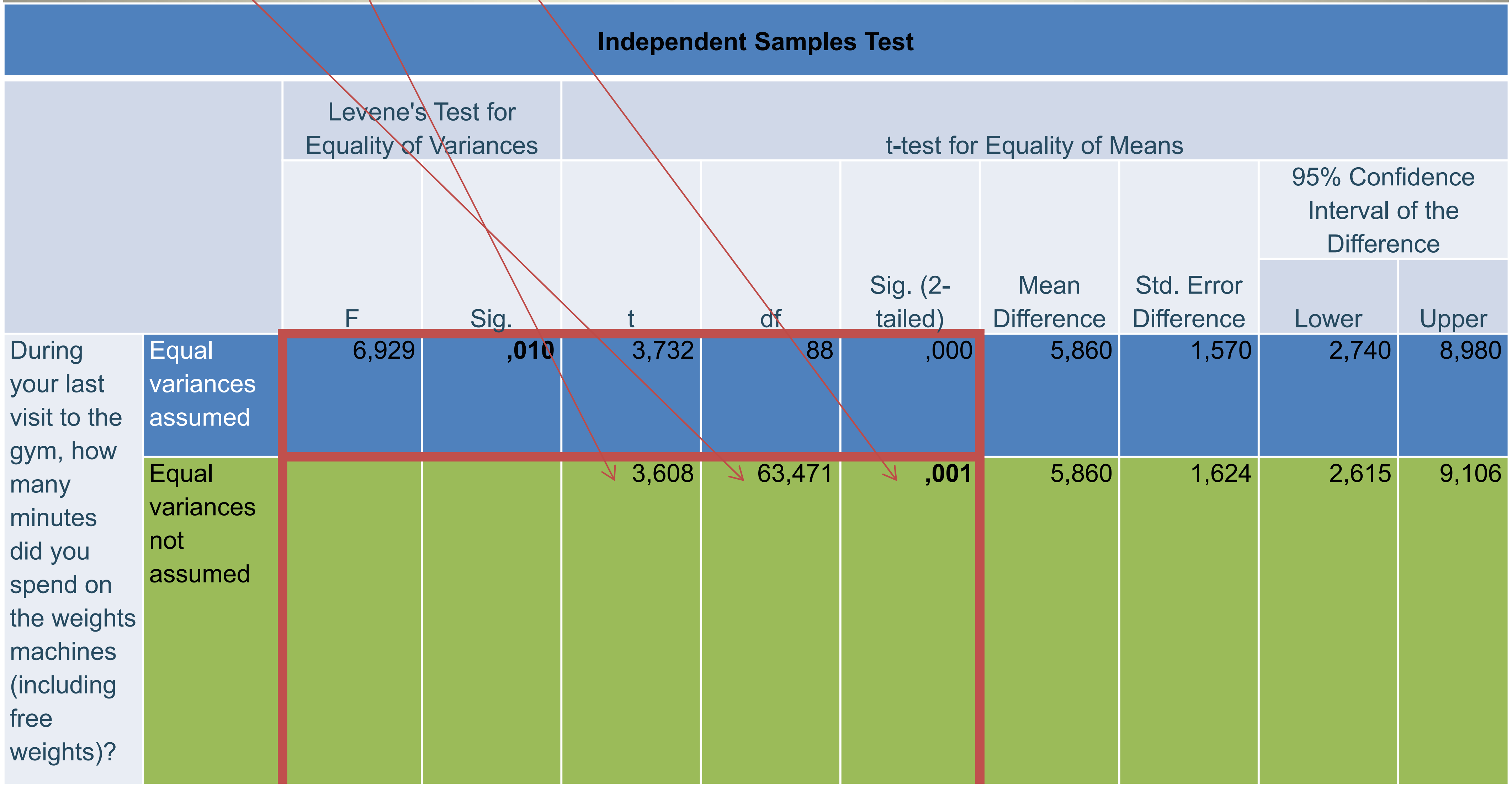
Define groups: Specify the groups values

Test variable(s): Dependent variables (scale)

If the significance level is larger than .05, you should use the first line.

If the significance level is lower than .05 use the second line

$T\ (63.471) = 3.608,\ p=.001$



**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| During your last visit to the gym, how many minutes did you spend on the weights machines (including free weights)? | Equal variances assumed | 6,929 | **,010** | 3,732 | 88 | ,000 | 5,860 | 1,570 | 2,740 | 8,980 |
| | Equal variances not assumed | | | 3,608 | 63,471 | **,001** | 5,860 | 1,624 | 2,615 | 9,106 |

# *Reporting findings - APA*

*An independent t-test found a statistical significant effect for time spent on weight machine exercise,*

*T (63.471) = 3.608, p=.001. Women reported spenting less time ($M_w$ =12.1, SD= 5.326) than men ($M_m$= 18, SD= 9.273) on weight machine exercices.*

# ANOVA

*Is used to compare two or more group means for statistical significance.*

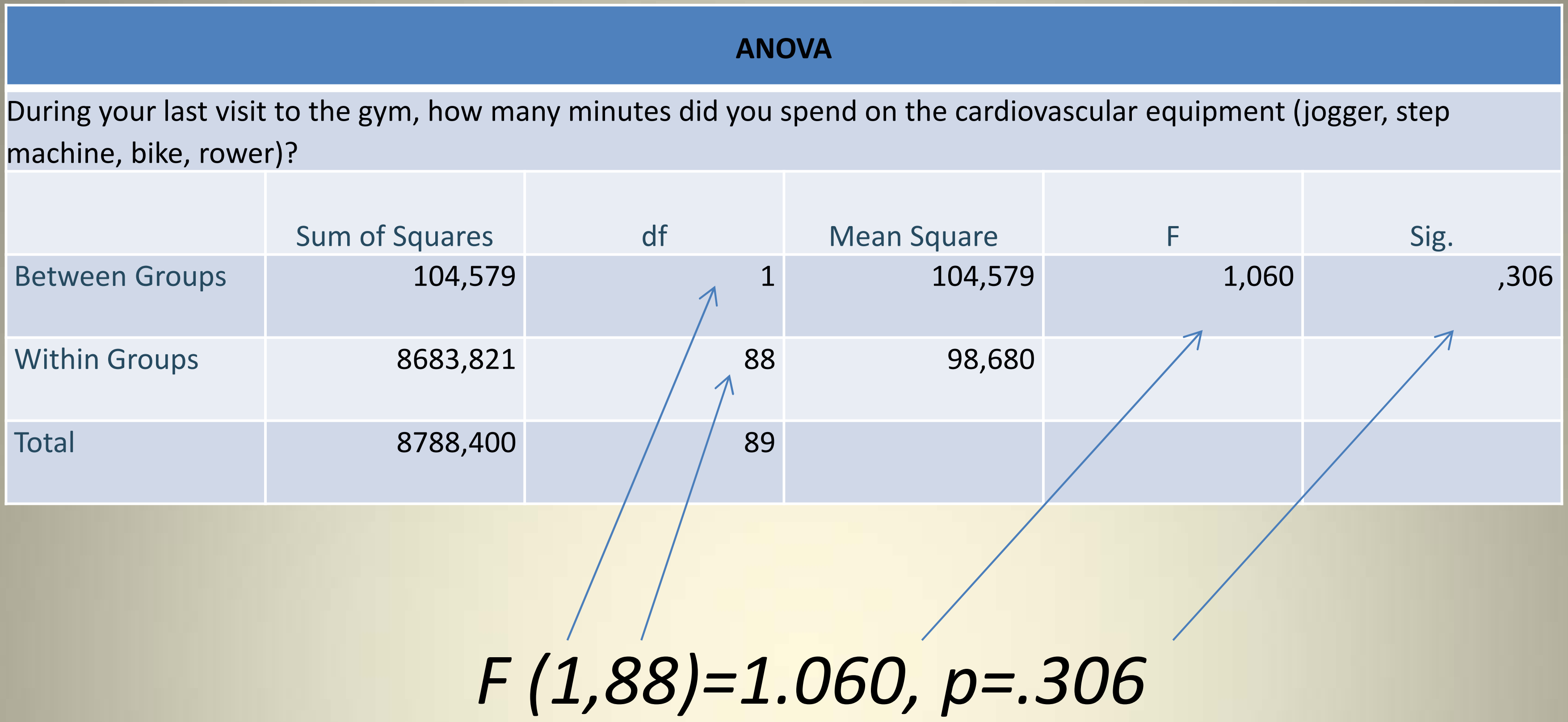
***One-way analysis of variance (one-way ANOVA):*** It is used when you have one quantitative dependent variable and one categorical independent variable (with two or more groups).

Two-way ANOVA is used when you have two categorical independent variables. Three-way ANOVA is used when you have three categorical independent variables and so forth.

Analyze→ Compare means→ One-Way ANOVA

Factor: Independent variable

Dependent list: Dependent variable(s)



**ANOVA**

During your last visit to the gym, how many minutes did you spend on the cardiovascular equipment (jogger, step machine, bike, rower)?

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 104,579 | 1 | 104,579 | 1,060 | ,306 |
| Within Groups | 8683,821 | 88 | 98,680 | | |
| Total | 8788,400 | 89 | | | |

*F (1,88)=1.060, p=.306*

A one-way analysis of variance was conducted to determine if the relationship between gender and cardiovascular exercise time was statistically significant. The ANOVA was not significant, *F(1, 88) = 1.060, p = .306. The means were* $M_m$*= 27.6 (SD=12.541) and* $M_w$*= 25.46 (SD=6.897).*

**Two-way analysis of variance (also called two-way ANOVA) is used when** you have a quantitative dependent variable and two categorical independent variables.

Analyze → General Linear Model → Univariate

Fixed Factor(s): Independent variables

Dependent variable: …

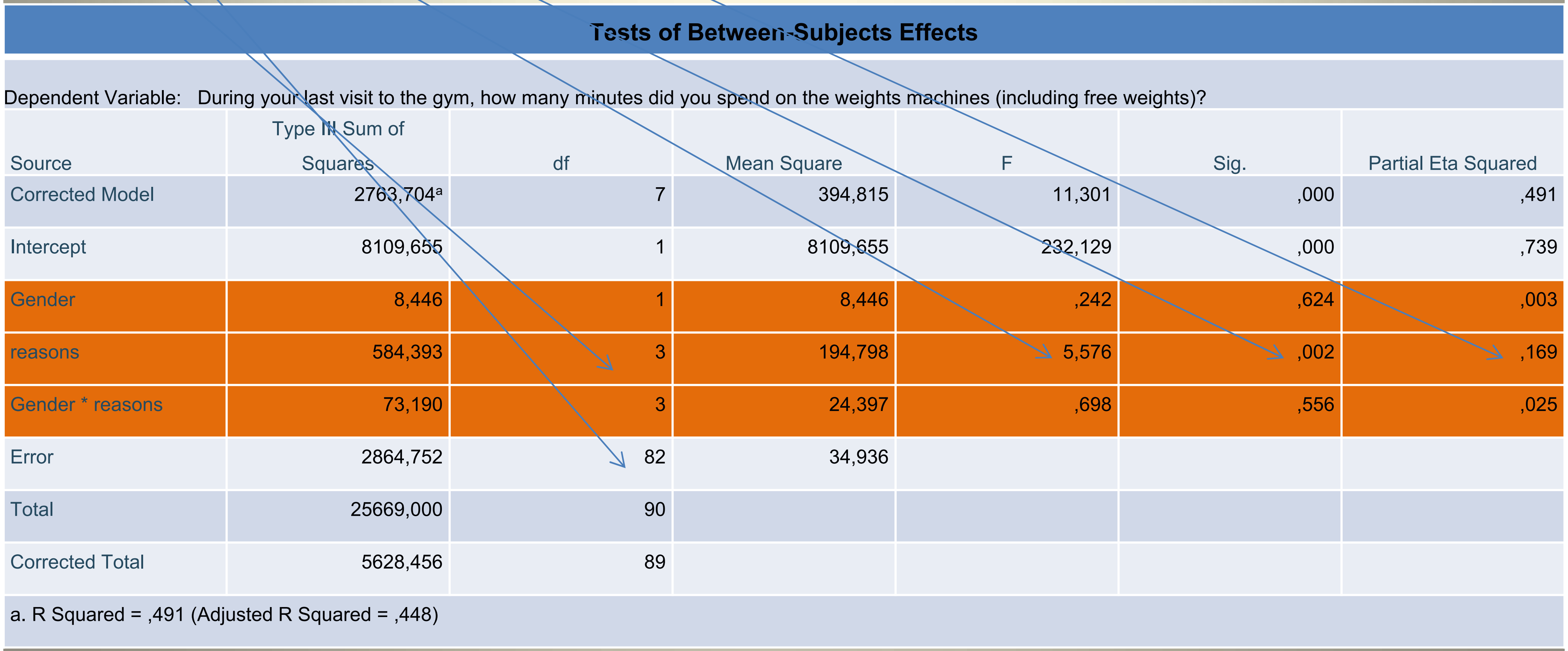
Options tab → Descriptive statistics

→ Estimate of effect size

| | |
|---|---|
| < .1 | Trivial effect |
| .1 - .3 | Small effect |
| .3 - .5 | Moderate effect |
| > .5 | Large effect |

*F*(3, 82) = 5,576, *p* = .002, *η*2 = .169

When examining two or more independent variable it is recommended to use the **eta squared effect size indicator (η²)**. The eta squared tells you how much variance in the dependent variable is explained by the independent variable. While statistical significance inform us about a real difference, effect size informs us about how important this difference is. Effect size aids the researcher to decide about the **practical significance** of the findings.



**Tests of Between-Subjects Effects**

Dependent Variable: During your last visit to the gym, how many minutes did you spend on the weights machines (including free weights)?

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 2763,704[a] | 7 | 394,815 | 11,301 | ,000 | ,491 |
| Intercept | 8109,655 | 1 | 8109,655 | 232,129 | ,000 | ,739 |
| Gender | 8,446 | 1 | 8,446 | ,242 | ,624 | ,003 |
| reasons | 584,393 | 3 | 194,798 | 5,576 | ,002 | ,169 |
| Gender * reasons | 73,190 | 3 | 24,397 | ,698 | ,556 | ,025 |
| Error | 2864,752 | 82 | 34,936 | | | |
| Total | 25669,000 | 90 | | | | |
| Corrected Total | 5628,456 | 89 | | | | |

a. R Squared = ,491 (Adjusted R Squared = ,448)

# Reporting findings - APA

A 2 X 2 ANOVA was conducted to evaluate the effects of reasons for going to the gym and gender. The analysis yielded a statistically significant “reasons for going to the gym” main effect, F(3, 82) = 5,576, p = .002, η2 = .169. The gender main effect was not significant, F(3, 82) = 8.446, p = .624, η2 = .003. There was no statistically significant interaction between reasons for going to the gym and gender, F(3, 82) = .556, p = .01, η2 = .025.