UNIVERSITY OF BERGEN

Refreshing uni- / bivariate statistics: Basic concepts

Sebastian Jentschke

UNIVERSITY OF BERGEN

The refresher lectures are split in two parts. A more theoretical one – which is this one – and a more practical one where I speak about the different uni- and bivariate statistical analyses.

Overview:

- Units, variables and values
- Population and sample
- Level of measurement – Variable levels
- Organizing your data
- Descriptive statistics
- Research hypotheses vs. statistical hypotheses
- Conceptualizations of the p-value
- Statistical significance vs. effect size

This theoretical part introduces some basic concepts:
Units, variables and values introduces central ideas about what we measure and how.
Population and sample deals with that we (typically) want to make general claims (about a population) while measuring just a small group (the sample).
When measuring, our data can have different “quality levels” called level of measurement.
I will then give some hints about organizing and documenting your data.
Then, I will speak about descriptive statistics as a mean to characterize your sample and to assess whether the sample conforms to the requirements for carrying out certain statistical analyses.

Overview:

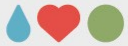
- Units, variables and values
- Population and sample
- Level of measurement – Variable levels
- Organizing your data
- Descriptive statistics
- Research hypotheses vs. statistical hypotheses
- Conceptualizations of the p-value
- Statistical significance vs. effect size

The next part deals with how we formulate good (i.e., precise and concise) research hypotheses and how these can be converted into statistical hypotheses (that we finally test).

Over time, there have been proposed different theoretical accounts about what the p-value means. I contrast two “classical” accounts – Fisher and Neyman – with a brief explanation on Bayesian statistics that became more widely used in recent years.

Finally, I will speak about that dilemma that statistical significance “scales” with sample size and the we might get significant differences that lack to be practically important. Effects size measures allow to describe and assess this aspect of practical importance.

Units, variables and values



Units, variables and values

- **units** = units of observations
  persons or objects of our research
- **variables** = measures collected from units
  variables (can vary): more than one possible value
- **values** = expressions / stages of a variable
  often numbers, sometimes categories

Before we go into details, I would like to define and clarify some terms and concepts. One central concept is that of **Units**, denoting units of observations. Units are persons or objects the research is about. Having persons as units is maybe more intuitive: We prepare an experiment, and then invite persons to participate. Objects can be, e.g., organizations: We could, e.g., explore how satisfied you (the student within a certain course, all students at UiB) you are with having digital lectures in that semester.

Units, variables and values

- **units** = units of observations
  persons or objects of our research
- **variables** = measures collected from units
  variables (can vary): more than one possible value
- **values** = expressions / stages of a variable
  often numbers, sometimes categories

From these units, we collect **variables** (as the name indicates, variables is something that can vary, i.e., a phenomenon which can have more than one value → each variable has to have at least two values to be called variable). Examples for variables are: age, gender, level of education, revenue.

Units, variables and values

- **units** = units of observations
  persons or objects of our research
- **variables** = measures collected from units
  variables (can vary): more than one possible value
- **values** = expressions / stages of a variable
  often numbers, sometimes categories

Values are different expressions or stages of a variable. Often these expressions or stages take the form of numbers, but they can also be qualitative (different categories). Examples where variables are expressed as numbers are age, body height, reaction times. Examples where variables are categories are gender, treatment vs. control, car brands, level of education, or responses in a questionnaire. Especially the latter is a bit of a hybrid: There is an assumption that even though the individual questions are ordered ranks (1 = completely agree, 2..., 3..., 5 = Completely disagree) the sum of all questions is regarded as a continuous variable.

Units, variables and values

- **units** = units of observations
  persons or objects of our research
- **variables** = measures collected from units
  variables (can vary): more than one possible value
- **values** = expressions / stages of a variable
  often numbers, sometimes categories

I already said a little about the classification into continuous and categorical variables and how we can use that to decide which statistical analysis is appropriate in the introduction lecture. But, what I said there was a bit of a simplification as both categorical and continuous variables each combine to variable levels. I will speak about that a little later in the lecture in more detail.

Units, variables and values

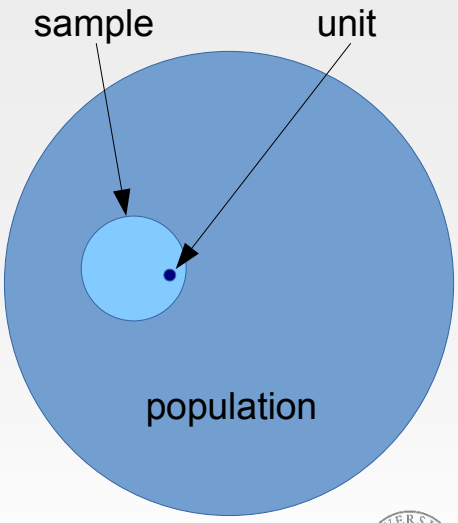
	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
1	2.479	4.208	3.938	3.958	3.458
2	2.604	3.188	3.958	3.396	3.229
3	2.813	2.896	3.417	2.750	3.500
4	2.896	3.563	3.521	3.167	2.792
5	3.021	3.333	4.021	3.208	2.854
6	2.521	3.292	3.438	3.708	2.500
7	2.354	4.417	4.583	3.063	3.333
8	2.521	3.500	2.896	3.667	3.063
9	3.104	3.813	4.063	3.771	2.833
10	2.688	3.547	3.787	3.354	3.104
11	2.625	3.458	2.896	3.458	3.375
12	2.375	3.771	3.167	3.500	3.521
13	3.063	3.417	3.771	3.813	3.125
14	3.125	2.521	2.646	3.750	3.208

At the same time determine units, variables and values how we arrange our data in our spreadsheet (table). The **units** are typically arranged as **lines**, the **variables** as **columns**, and each unit (e.g., a specific participant) assumes a particular **value** for each variable (in the example, participant 2 assumes the value 3.396 in the variable agreeable-ness).

Population and sample

Population and sample

- populations → samples → units
- aim: make statements about the population (general principles)
- Law of large numbers: with many trials the value in a sample approaches the «true» value

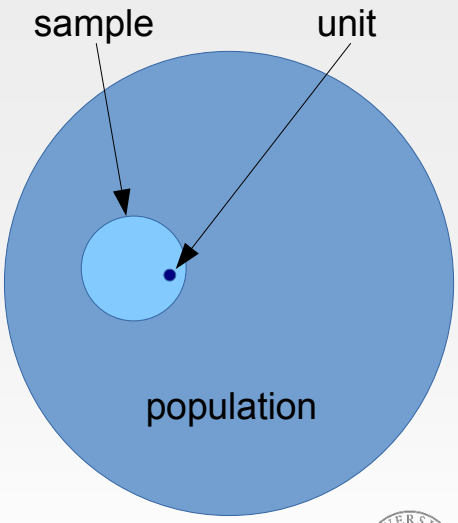


First, I will say something about units, samples and populations. As said above, units are often persons, and when we conduct a scientific study we typically collect data from a sample, i.e., a group of persons. In principle, this also applies if the units are objects (e.g., organizations). In such case a sample would be several object. For the sake of simplicity it is maybe easier to keep with the idea of persons and samples as groups of persons (often denoted as participants) for now. These samples are part of a larger population.

What we would in most cases would like to do is to make statements about the population.

Population and sample

- populations → samples → units
- aim: make statements about the population (general principles)
- Law of large numbers: with many trials the value in a sample approaches the «true» value



A guiding principle is the Law of large numbers which states that the results obtained from a large number of trials should be close to the expected value and will tend to become closer the more trials are performed. For example, if we want to know something about the average height of females Norway, we will get an estimate that is closer to the “true” body height if our sample contains 40 (instead of 20) females.

Population and sample

Types of statistics:

- descriptive statistics
  → characterize our sample
  → check assumptions
- inference statistics
  → draw conclusions for the population
  → sample has to be ***representative***

With these sample we carry out to “types” of statistics.

One is called “descriptive statistics”. Here we try to characterize our sample, e.g., with respect to central tendency (using mean, median, mode, etc.) and with respect to variation (using standard deviation, minimum and maximum, range, etc.). Descriptive statistics is also used to check assumptions that have to be fulfilled to carry out certain statistical analyses (e.g., are the data normally distributed).

Population and sample

Types of statistics:

- descriptive statistics
  → characterize our sample
  → check assumptions
- inference statistics
  → draw conclusions for the population
  → sample has to be ***representative***

The other type of statistics is called “inference statistics”. Using analyses that fall within inference statistics, we want assess whether we can draw conclusions from the sample we collected to the population (the sample was drawn from). Typically we assess using inference statistical methods whether we can conclude (make the inference) that the same relation that we observed in the sample (e.g., a difference in means between two conditions in an experiment) also exists in the population.

Population and sample

Types of statistics:

- descriptive statistics
  → characterize our sample
  → check assumptions
- inference statistics
  → draw conclusions for the population
  → sample has to be ***representative***

When selecting a sample, we need to have an overview over the target population. The sample has to include all characteristics about which the researcher wants to make claims on the basis of the experiment (e.g., if we want to draw conclusions about the whole population, we can't have a sample only consisting of women). The sample has to be suitable to describe the entire population, it has to be **representative**. If not, there will be a problem to generalize the findings.

Typically, if we randomly choose participants within for sample, this sample likely will be representative for the population.

Population and sample

Representativeness → methods for sample selection

***probabilistic*:**

- simple random sampling
- systematic sampling (e.g. every fifth)
- stratified sampling (e.g. sex, age groups)
- cluster sampling (e.g. geogr. regions)
- quota sampling (quota per subgr., e.g. 200 F / 300 M, 45 – 60 yr)

***non-probabilistic*:**

- accidental sampling (available participants; cafeteria, on the street)
- voluntary selection / self-selection (react to advert.; poster, e-mail)
- discretionary selection (researcher sel. accord. to expect. repres.)

To ensure representativeness, everyone who is part of the population we are interested in must have a certain chance of being included in the sample.

The most common forms of selection are:

(1) When using Simple random sampling, all combinations of characteristics have the same probability of being drawn.

(2) Systematic sampling (or interval sampling) relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals (e.g., each fifth person).

(3) In Stratified sampling, the population of interest is divided into subgroups or strata (e.g., by gender, age, etc.) and a random selection is drawn within each stratum.

Population and sample

Representativeness → methods for sample selection

***probabilistic*:**

- simple random sampling
- systematic sampling (e.g. every fifth)
- stratified sampling (e.g. sex, age groups)
- cluster sampling (e.g. geogr. regions)
- quota sampling (quota per subgr., e.g. 200 F / 300 M, 45 – 60 yr)

***non-probabilistic*:**

- accidental sampling (available participants; cafeteria, on the street)
- voluntary selection / self-selection (react to advert.; poster, e-mail)
- discretionary selection (researcher sel. accord. to expect. repres.)

(4) Cluster sampling selects groups that are internally heterogeneous yet (relatively) homogeneous within the group (e.g., from certain geographical regions, age groups, etc.).

(5) Quota sampling segments the population into mutually exclusive sub-groups, then participants from each segment are selected based on a specified proportion (e.g., 200 females and 300 males between the age of 45 and 60).

Population and sample

Representativeness → methods for sample selection

***probabilistic*:**

- simple random sampling
- systematic sampling (e.g. every fifth)
- stratified sampling (e.g. sex, age groups)
- cluster sampling (e.g. geogr. regions)
- quota sampling (quota per subgr., e.g. 200 F / 300 M, 45 – 60 yr)

***non-probabilistic*:**

- accidental sampling (available participants; cafeteria, on the street)
- voluntary selection / self-selection (react to advert.; poster, e-mail)
- discretionary selection (researcher sel. accord. to expect. repres.)

In addition, there is a couple of non-probabilistic sampling approaches (i.e., we used methods that are not based on a random selection). Yet, these methods aim to establish representativeness of the sample:

(a) Accidental sampling describes that the sample is being drawn from that part of the population which is close to hand (recruiting participants in the cafeteria or on the street).

(b) Voluntary selection / self-selection is where the participants decide whether they want to participate in the study (e.g. by reacting to an advertisement via poster or e-mail).

(c) For discretionary selection, the researcher selects the units according to how typical he / she assumes they are for the population.

Population and sample

- **statistics**: characteristics of measures in the sample vs.
  parameters: «true» values in the population
- **Central limit theorem**: statistics in the sample approach with increasing sample size the «true» value in the population

Finally, there is a distinction between characteristics we measure in our sample (these are called “statistics”) and the true value of these characteristics in the population (called “parameters” and typically denoted with greek letters). For example, the statistics for body height in a sample would be denoted as $\bar{x}$ or M for the mean and s for the the standard deviation whereas the parameters in the population would be denoted as μ for the mean and σ for the standard deviation.

Population and sample

- **statistics**: characteristics of measures in the sample vs.
  parameters: «true» values in the population
- **Central limit theorem**: statistics in the sample approach with increasing sample size the «true» value in the population

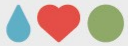
According to the Central limit theorem approach the data collected from a sample with increasing sample size a normal distribution. A more detailed description what this entails is given when explaining the z-test in the second part of this lecture.

Population and sample

- **statistics**: characteristics of measures in the sample vs.
  parameters: «true» values in the population
- **Central limit theorem**: statistics in the sample approach with increasing sample size the «true» value in the population

If you want a more comprehensive introduction into samples and populations, you can read chapter 8 in the jamovi-book (Navarro & Foxcroft, 2019). Or for a basic introduction, you can watch this video https://www.youtube.com/watch?v=eIZD1BFfw8E

Levels of measurement



Levels of measurement

- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

We categorize variables according to which range of values they can contain into four different levels. These levels are called: (1) Nominal level, (2) Ordinal level, (3) Interval level, and (4) Ratio level. What level of measurement a variable has is decisive for what kind of statistical analyses we can conduct using this variable. The higher the level, the more different tests can be conducted.

Levels of measurement

- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

At the nominal level variable values can classified into mutually exclusive categories. We can only say that values are equal and unequal with a category (i.e., we can not express them as ranks, or whether one category is better, higher or more valuable than another). Examples are gender, municipality of residence, nationality, political affiliation, etc.

Levels of measurement

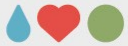
- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

For variables at ordinal level, categories can be ranked (in addition to being mutually exclusive). When comparing two units it makes sense to decide which one has the highest or lowest value of the variable. However, we can say nothing about the distance between the values. Examples are education level, level of income or socio-economic status, or responses in questionnaires (e.g., the categories between “completely disagree” to “completely agree” or “not at all” to “very much”).

Levels of measurement

- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

For variables at interval level, it is possible to measure the distance between the categories (in addition to that the categories can be ranked). We are therefore able to say, how much to values on that scale are apart from each other. Yet, we can't say anything about the relationship between the units as the scale does not have an absolute zero. Examples are the pH-scale, standard scores (e.g., in intelligence or personality tests) or temperature. Temperature does not have an absolute zero, but given that it is “normed” to have 0 (where water is freezing) and 100 (where water is boiling), we can at least say that going from 10 to 20 brings us an equally large distance from freezing to boiling as going from 20 to 30. Likewise, in a intelligence test, a test score of 85 tells us that this person is about one standard deviation below the mean (for that skill) as a test score of 115 tells us that the person is about one standard deviation above the mean.



Levels of measurement

- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

Finally, variables at **ratio level** have an absolute zero (in addition to also having equal intervals). This allows to compare the relationships between units. Examples are age, body height, body weight, etc. If one puts one person with 100 kg on a scale and two persons with 50 kg each on the other side, the scale would be in balance.

Levels of measurement

- **nominal level**: values mutually exclusive
  examples: gender, residence, nationality, etc.
- **ordinal level**: + can be ranked
  examples: education level, level of income / SES
- **interval level**: + equal intervals
  examples: pH-scale, standard scores, celsius
- **ratio level**: + absolute zero
  examples: age, body height, body weight

Please note that “measurement level” is called “measurement type” in jamovi and doesn’t contain a category for ratio level (since we can use parametric methods with either interval or ratio scale level). SPSS actually does the same (and here “measurement level” is called “Measure”). In addition, jamovi has three different categories within “Data type”: text, integer (without decimals), and decimal. These determine how the variables are stored (integer need less file space than decimals).

Measurement levels are introduced in chapter 2.2 of the jamovi-book (Navarro & Foxcroft, 2019). You can also watch this video: https://www.youtube.com/watch?v=LPHYPXBK_ks

Organizing your data

Organizing your data

- organize your data!
  directory structure – main directory
  → subfolders with participants OR
  → subfolders with measures (e.g., tests, experim.)
  → subfolder where you store (and keep) analyses
  (syntax, outputs, etc.; add a date yyyy-mm-dd)
- document your data!
  README file: general introduction, variable information, directory structure

➢ data sharing (possible requirement for publication)

Typically, when you start analysing a data set, the very first stage is to familiarize yourself with it. A word of advice: Each minute that you invest in documenting what you did when collecting the data will save you a lot of work later on and minimize that you make mistakes in your analyses (this happens quite often because you, e.g., can't remember what exactly you measured with a certain variable which may lead to wrong conclusions).

If you think of your M.Sc. thesis you typically will have one main directory that contains your summary files (i.e., files that contain the the variables you measured from all participants). Often you will have subfolder: Either one subfolder for each participant or one subfolder for each measure (e.g., a questionnaire, experimental data, etc.). Finally, create a subfolder where you keep syntax- and output-files of analyses you carried out.

Organizing your data

- organize your data!
  directory structure – main directory
  → subfolders with participants OR
  → subfolders with measures (e.g., tests, experim.)
  → subfolder where you store (and keep) analyses
  (syntax, outputs, etc.; add a date yyyy-mm-dd)
- document your data!
  README file: general introduction, variable information, directory structure

➢ data sharing (possible requirement for publication)

What I recommend you to do is to create a README file in the main directory of where your data are stored. What you should (at least) include in your README file is:

(1) a general introduction about the study: aim (possibly hypotheses), which main instruments where used (e.g., which questionnaires, which experimental paradigms, etc.), general comments (e.g., where there participants that were excluded and why, etc.)

(2) information about the variables contained: name, a verbal description what the variable contains (and possibly how it was measured), variable levels (if it is a categorical variable) and notes / comments (applying to that variable, if necessary)

(3) information about the directory structure: where are the raw data stored, how where the raw data converted into the main file (where you typically have summary variables)

Organizing your data

- organize your data!
  directory structure – main directory
  → subfolders with participants OR
  → subfolders with measures (e.g., tests, experim.)
  → subfolder where you store (and keep) analyses
  (syntax, outputs, etc.; add a date yyyy-mm-dd)
- document your data!
  README file: general introduction, variable information, directory structure

➢ data sharing (possible requirement for publication)

Another reason is that data sharing becomes a policy with more and more journals and your data will only be useful to others if they can understand what is contained and how they are organized.

Recently, this became more important in connection with the demand for replicable and reproducible research where you often have to publish your data together with your manuscript. Your data can only be useful to others if they are properly documented.

If you got a dataset from another person, you can hope that this person was making such an effort. Otherwise, it will take you a bit of time to answer the following questions that you are required to understand (e.g., for choosing appropriate statistical analyses): What is the data comprised of? How many observations? Which types and measurement levels have the variables? What are the values within the variables?

Organizing your data

- organize your data!
  directory structure – main directory
  → subfolders with participants OR
  → subfolders with measures (e.g., tests, experim.)
  → subfolder where you store (and keep) analyses
  (syntax, outputs, etc.; add a date yyyy-mm-dd)
- document your data!
  README file: general introduction, variable information, directory structure

➢ data sharing (possible requirement for publication)

Believe me, even though that might sound tedious and boring, it is worth every second you invest. A very typical case is that you prepare a manuscript. A rough estimate for the time between working on the manuscript and having to do revisions after comments from reviewers is between 6 and 12 months. How well do you believe do you know how your data are organized after such a period?

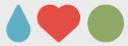
Descriptive statistics

Descriptive statistics

- summarize and visualize your data:
  central tendency: mean, median, mode
  dispersion / variation: std. dev., min. - max.
  visualization
- assure that assumptions (e.g., Normality) are met:
  extreme values / outliers?
  assumption tests

Descriptive statistics also called Explorative data analysis serves to summarize and visualize your data. What we typically do with the sample is to describe characteristics of the variables we measured. This description typically encompasses information about the central tendency as well as an indicator of variation:

If the data are on an ***interval scale level*** we typically use the ***mean*** (to describe the central tendency) and the ***standard deviation*** (as a measure of variation). If you ran an study with several conditions you will likely report at least the mean and standard deviation (or variance) for each condition.



Descriptive statistics

- summarize and visualize your data:
  central tendency: mean, median, mode
  dispersion / variation: std. dev., min. - max.
  visualization
- assure that assumptions (e.g., Normality) are met:
  assumption tests
  extreme values / outliers?

If your data are on an ***ordinal scale level***, we typically use the ***median*** (to describe the central tendency) and ***range or minimum and maximum*** (as a measure of variation).

If the data are on an ***nominal scale level***, we use the ***mode*** (the value that occurs most often) as the measure of central tendency. The measure of variation could be a ***table*** with the ***frequency of occurrences***.

If your data are on an interval scale level but not normally distributed, you may decide to use non-parametric statistics. In such case, you decide to treat these data on an ordinal scale level even though the level on which they were measured was an interval scale.

Descriptive statistics

- summarize and visualize your data:
  central tendency: mean, median, mode
  dispersion / variation: std. dev., min. - max.
  visualization
- assure that assumptions (e.g., Normality) are met:
  assumption tests
  extreme values / outliers?

In addition or alternatively to reporting these values, you typically visualize your data: For an experimental study with several conditions you will likely use a bar graph showing the mean of these conditions in comparison. If your study explored correlations, you might want to visualize the relation between these to variables using a scatter plot (please note that you need to install the module “scatr” to produce those plots).

Descriptive statistics

- summarize and visualize your data:
  central tendency: mean, median, mode
  dispersion / variation: std. dev., min. - max.
  visualization
- assure that assumptions (e.g., Normality) are met:
  assumption tests
  extreme values / outliers?

Finally, descriptive statistics is typically also used to assess whether the prerequisites for certain statistical analyses are met. For example, most so-called parametric statistics (which is the majority of tests we are using, including, e.g., t-test, ANOVA, correlation, regression, etc.) are based on the assumption that the distribution of values in your variables follows a normal distribution. You can use Descriptive statistics to check this. As you have seen in the part on population and samples, a sample might not perfectly represent the population it is coming from. That is, the way the data in the sample are distributed might not follow a normal distribution. One condition, where the sample data deviate from a normal distribution is if the sample contains extreme values. These are denoted as outliers. The check for these outliers and to visually assess whether the data are normally distributed, we use Descriptive statistics.

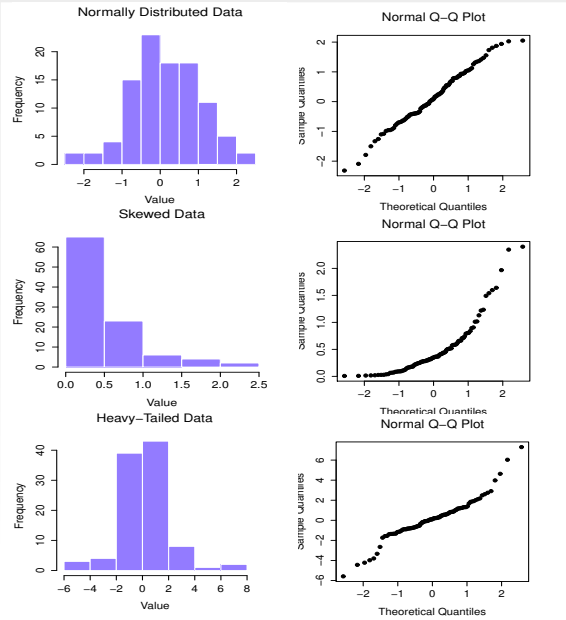
Descriptive statistics

- summarize and visualize your data:
  central tendency: mean, median, mode
  dispersion / variation: std. dev., min. - max.
  visualization
- assure that assumptions (e.g., Normality) are met:
  assumption tests
  extreme values / outliers?

Therefore, it is typically the first step before you conduct further parametric statistical analyses (like correlation, t-test, regression or ANOVA).

A second step in checking prerequisites are the so-called assumption tests. You will find a dropdown-box called “Assumption Tests” for (more or less) all analyses implemented in jamovi. This second step will be described when covering the individual tests in the lecture.

Descriptive statistics



normally distributed

heavily skewed
(positive skewness)

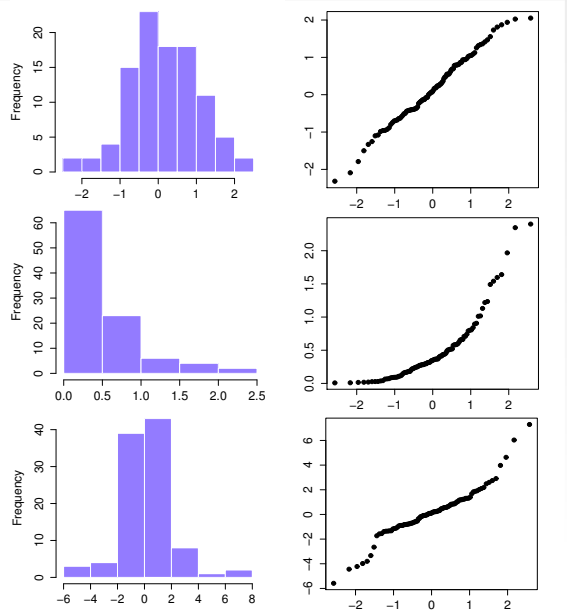
too pointy, tails to flat
(positive kurtosis)

A very convenient way of assessing whether your data are normally distributed is visually. If we look at the figures shown on this slide, we can see three conditions each with a histogram and a Q-Q-plot.

As a brief note: Here a histogram is used. I have a preference for the combination of box and violin plot (and will use that in the demonstration later).

My reason for having that preference is that basically, the violin plot gives you the same information as the histogram (even though it's turned by 90° and smoothed). In combination with the box plot, you can at one glance assess to what degree the data look normally distributed (from the violin plots) plus in addition see whether there are any outliers (from the box plots).

Descriptive statistics



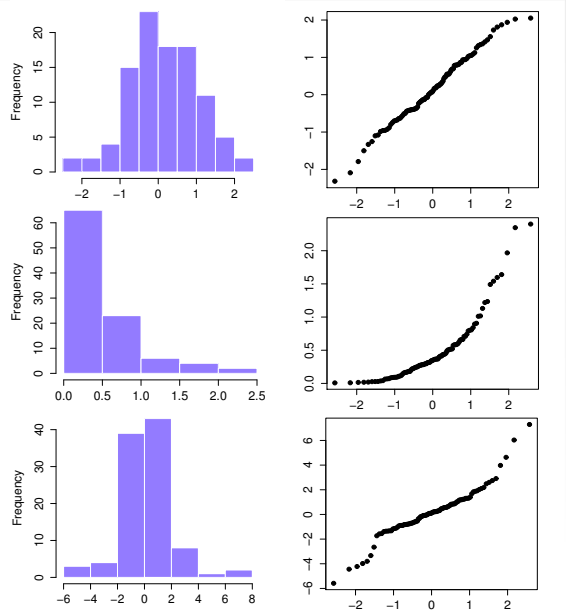
normally distributed

heavily skewed
(positive skewness)

too pointy, tails to flat
(positive kurtosis)

The first condition shows how the data look like if they (generally) follow a normal distribution. We observe the “normal-distribution-like”-shape of the histogram and that most dots in the Q-Q-plot fall close to the main diagonal.

Descriptive statistics



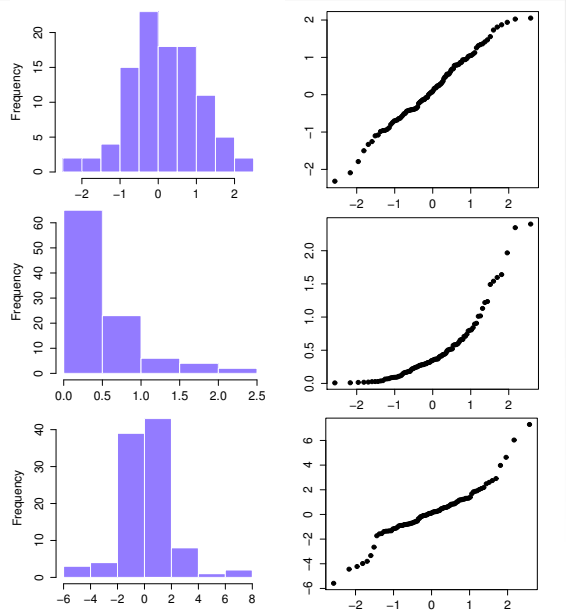
normally distributed

heavily skewed
(positive skewness)

too pointy, tails to flat
(positive kurtosis)

If the data are skewed (in that case, they are positively skewed), the have a longer tail on one side than on the other. In the histogram we see that the “peak” of the distribution (called mode: the value or that bin within the histogram that appears most frequently) is shifted from the middle to one side (i.e., the distribution is not symmetrically as it would be if it followed a normal distribution). In the Q-Q-plot, the vertical axis (showing the empirical distribution in our dataset) starts with 0 and goes only to 2. That mean that, if we compare that to a typical normal distribution, it is like cut in the middle.

Descriptive statistics



normally distributed

heavily skewed
(positive skewness)

too pointy, tails to flat
(positive kurtosis)

If the data have heavy tails, extreme values are overrepresented. Whereas in a histogram showing a “typical” standard deviation, most values (~95%) fall within plus / minus two standard deviations, here the scale goes from -6 to 8. The same can be seen on the vertical axis (showing the empirical distribution). If we look at the Q-Q-plot, the slope in the middle part appears less steep (which is a consequence of the vertical axis having a value range of -6 to 6), whereas there is a lot of additional dots on the far ends.

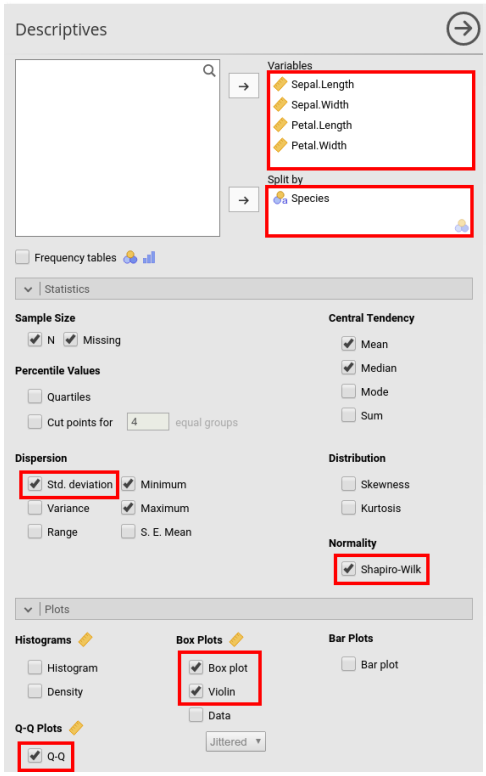
Descriptive statistics

- ☰ → Open → Data Library → Anderson's Iris Data
- Exploration → Descriptives
  Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
  → «Variables»
  Species → «Split by»
  tick: Shapiro-Wilk, Box plot, Violin, Q-Q

But let's come back to how you can use Descriptive statistics in jamovi in order to describe your data and to assess whether your data are suited for your analyses or whether you have to "clean" your data.

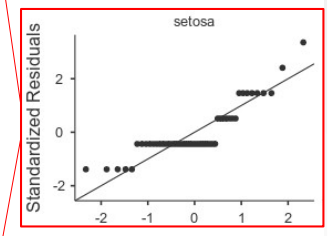
We will use the Andersen's Iris-dataset. You can open this dataset by clicking on the ☰ (top-left corner in jamovi) → "Open" → "Data library" → "Anderson's Iris data". The dataset contains four columns with continuous variables (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) and one categorical variable (Species).

We assign the continuous variables to "Variables" and Species to "Split by". We open the drop-down-box "Statistics" (click the ">"-sign) and tick "Shapiro-Wilk" (under "Normality"). Then we open the drop-down box "Plots" and tick "Box plot" and "Violin" (under "Box Plot"), and "Q-Q" (at "Q-Q-plots").

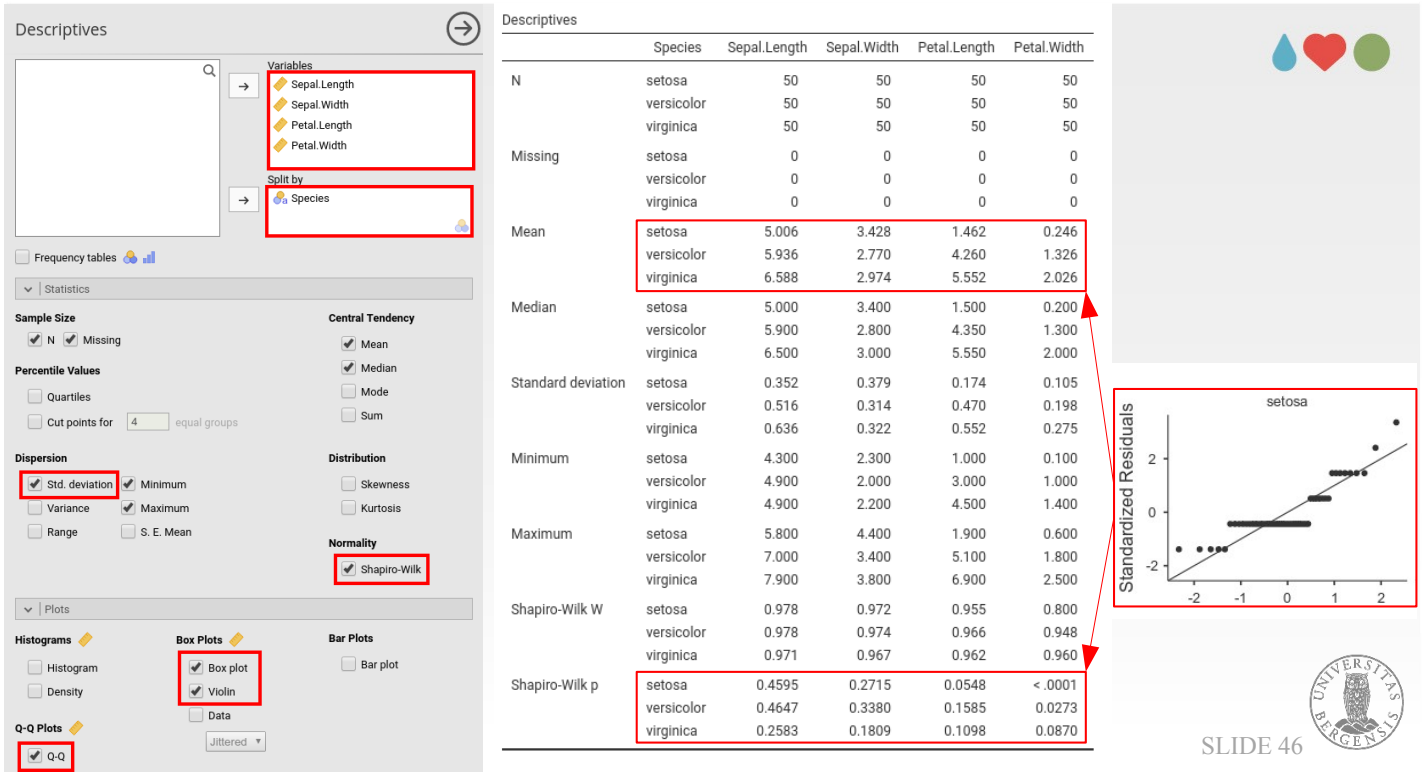


Descriptives

	Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
N	setosa	50	50	50	50
	versicolor	50	50	50	50
	virginica	50	50	50	50
Missing	setosa	0	0	0	0
	versicolor	0	0	0	0
	virginica	0	0	0	0
Mean	setosa	5.006	3.428	1.462	0.246
	versicolor	5.936	2.770	4.260	1.326
	virginica	6.588	2.974	5.552	2.026
Median	setosa	5.000	3.400	1.500	0.200
	versicolor	5.900	2.800	4.350	1.300
	virginica	6.500	3.000	5.550	2.000
Standard deviation	setosa	0.352	0.379	0.174	0.105
	versicolor	0.516	0.314	0.470	0.198
	virginica	0.636	0.322	0.552	0.275
Minimum	setosa	4.300	2.300	1.000	0.100
	versicolor	4.900	2.000	3.000	1.000
	virginica	4.900	2.200	4.500	1.400
Maximum	setosa	5.800	4.400	1.900	0.600
	versicolor	7.000	3.400	5.100	1.800
	virginica	7.900	3.800	6.900	2.500
Shapiro-Wilk W	setosa	0.978	0.972	0.955	0.800
	versicolor	0.978	0.974	0.966	0.948
	virginica	0.971	0.967	0.962	0.960
Shapiro-Wilk p	setosa	0.4595	0.2715	0.0548	< .0001
	versicolor	0.4647	0.3380	0.1585	0.0273
	virginica	0.2583	0.1809	0.1098	0.0870

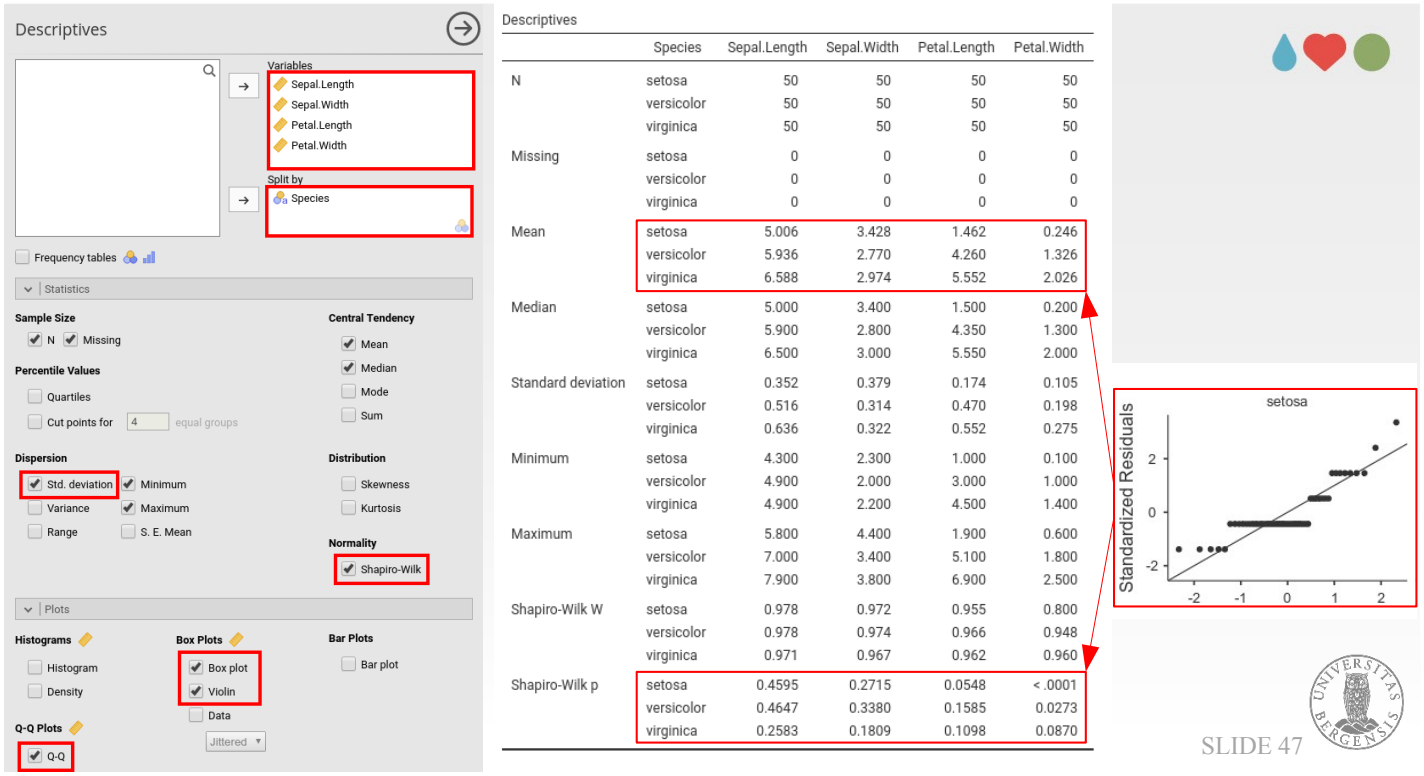


We get a very comprehensive table that we can have a look at while the figures get prepared. We first take a look at the very bottom where we find the “Shapiro-Wilk p”. A low p-value indicates a deviation from a normal distribution. As a general rule, if $p > 0.1$ that typically is fine, if p is between 0.05 and 0.10 there is reason for concern, if $p < 0.05$ we should really consider whether we need to use that variable. In any case, for p-values smaller than 0.10, we should check whether there are outliers that might explain that variation. When looking at the table again, we see that the p-values for the first and the second columns (Sepal.Length and Sepal.Width) seem fine, in the third column (Petal.Length) the values for the Species “setosa” raise a bit concern, and in the fourth column (Petal.Width) more or less all values.

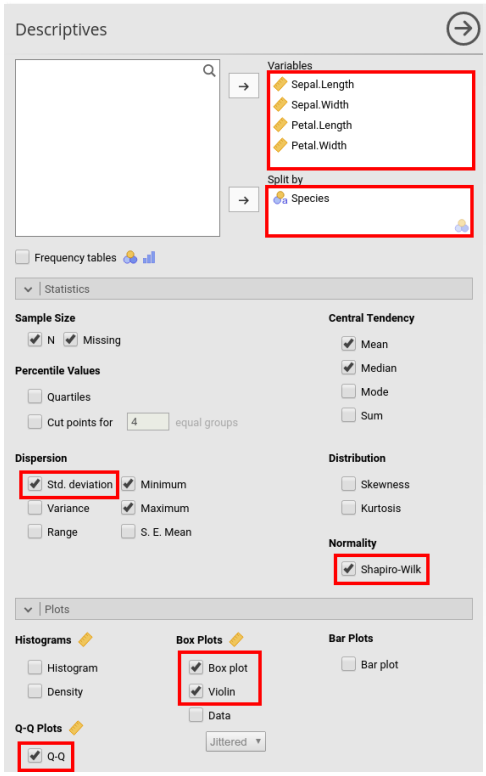


Let's now turn to the figures. Please note, that only the most important figure is included in the slide, so you either have to conduct the analysis yourself or open AndersenIris_Descriptives.html on MittUiB → Files→ Data sets → Examples4 jamovi.

For the Box- / Violin-Plot-combination, we check whether there are any outliers (black dots), how many there are and how extreme. Possibly, we have to decide to remove cases from our dataset, if that case contains an extreme value. Generally, this looks fine for that data set. There are some outliers but they appear not very extreme. The second thing we look at (still in the Box- / Violin-Plot-combination) is whether the "Violin"-part looks approximately like a normal distribution curve (turned 90°). There is neither that raises red flags.

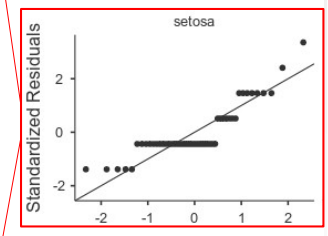


Now, we turn to the Q-Q-plots. What the do is comparing the empirical distribution of values (i.e., the actual values contained in our variables) with what would be expected theoretically (i.e., what we would expect based on a normal distribution). The rationale is as follows: When we take a sample and expect that the values are normally distributed, we expect that the majority of the values falls (relatively) close to the mean and extreme values are more uncommon. If we had an empirical distribution that were perfectly in accordance with a normal distribution, all values would fall on the black line. The more they deviate from the black line (especially in the bottom-left and the top-right corner) the more likely with have a skewed distribution, extreme outliers that prevent us from using that variable.

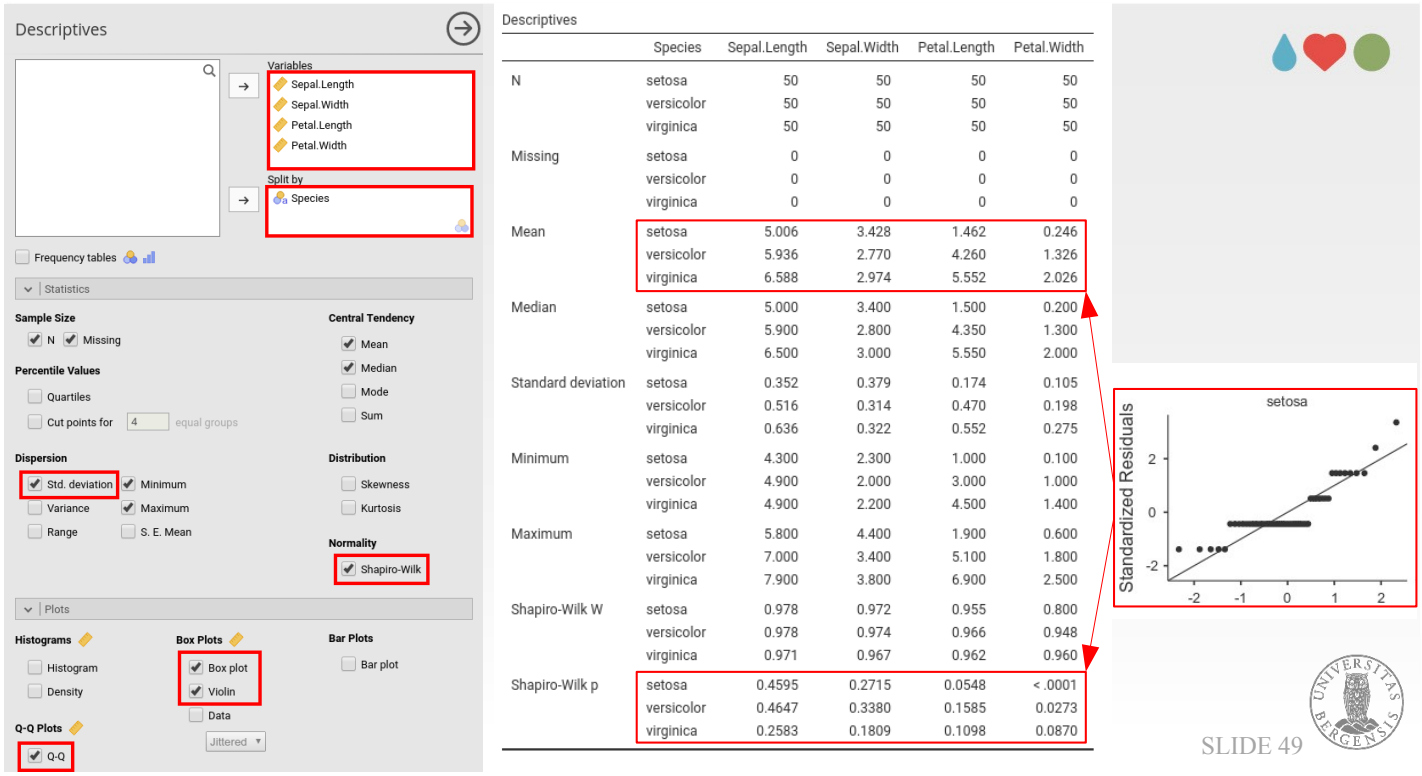


Descriptives

	Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
N	setosa	50	50	50	50
	versicolor	50	50	50	50
	virginica	50	50	50	50
Missing	setosa	0	0	0	0
	versicolor	0	0	0	0
	virginica	0	0	0	0
Mean	setosa	5.006	3.428	1.462	0.246
	versicolor	5.936	2.770	4.260	1.326
	virginica	6.588	2.974	5.552	2.026
Median	setosa	5.000	3.400	1.500	0.200
	versicolor	5.900	2.800	4.350	1.300
	virginica	6.500	3.000	5.550	2.000
Standard deviation	setosa	0.352	0.379	0.174	0.105
	versicolor	0.516	0.314	0.470	0.198
	virginica	0.636	0.322	0.552	0.275
Minimum	setosa	4.300	2.300	1.000	0.100
	versicolor	4.900	2.000	3.000	1.000
	virginica	4.900	2.200	4.500	1.400
Maximum	setosa	5.800	4.400	1.900	0.600
	versicolor	7.000	3.400	5.100	1.800
	virginica	7.900	3.800	6.900	2.500
Shapiro-Wilk W	setosa	0.978	0.972	0.955	0.800
	versicolor	0.978	0.974	0.966	0.948
	virginica	0.971	0.967	0.962	0.960
Shapiro-Wilk p	setosa	0.4595	0.2715	0.0548	< .0001
	versicolor	0.4647	0.3380	0.1585	0.0273
	virginica	0.2583	0.1809	0.1098	0.0870



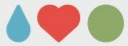
The Q-Q-plot for the first two variables look fine, for the third variable we observe a little strange plot for the species setosa (where we had concerns about the normal distribution) and for the fourth variable we observe a very similar pattern (lines of dots like as if the the data were layered). We got an explanation for why this happened from looking at the means of these four occasions where we had concerns about Normality (species “setosa” for Petal.Length, all species for Petal.Width). The means here are all relatively small, which indicates that in a lot of cases the values for Petal.Length and Petal.Width don’t vary very much since they are so small. The values are therefore “behaving” more like as if they belonged into categories than as if they were continuous. That is the reason for why they are not normally distributed. If we had the orginal data, we could use more decimals so that we get a more fine-grained set of values.



In the jamovi-book (Navarro & Foxcroft, 2019), you can read chapter 4 for a more comprehensive introduction into descriptive statistics and chapter 11.8 for a more in-depth discussion of how to assess normality visually.

If you would like a demonstration from a different angle, there are some of Barton Poulson's videos that you can watch: go to https://datalab.cc/jamovi/, click on the Hamburger button (three vertical lines; top-right in the video). Start from video 17 "Exploration: Chapter overview" and watch the following videos up to video 24 "Bar plots").

Research vs. statistical hypotheses



Research vs. statistical hypotheses

research hypothesis = precise and concise form of a research question (like: if X then Y)

examples:

- "Listening to music reduces your ability to pay attention to other things." (claim about the relation btw. meaningful concepts)
- "Intelligence is related to personality." (correlational, not causal; little too broad – choose a dimension)
- "Intelligence is speed of information processing." (invalid, ontological claim + confoundation)

When we start a scientific project we typically begin with a research question, i.e., a question that we would like to answer in the experiment. Often, we begin with a very general formulation of this question. From there, we have to work to make this question more concise. Once you arrived at such a concise and precise description overall scientific goal (typically a theoretical description of some predictor or cause and its influence on the expected outcome), you go further to a process called operationalization. This process involves that you consider how you could measure what is at the core of the research question (i.e., the assumed causes / predictors and the expected outcome).

That means that this transformation process involves taking a general question or a thought in everyday language into a plan or a design of an experiment (i.e., a way of assessing and measuring what is behind that question).

Research vs. statistical hypotheses

research hypothesis = precise and concise form of a research question (like: if X then Y)

examples:

- "Listening to music reduces your ability to pay attention to other things." (claim about the relation btw. meaningful concepts)
- "Intelligence is related to personality." (correlational, not causal; little too broad – choose a dimension)
- "Intelligence is speed of information processing." (invalid, ontological claim + confoundation)

At the core of the operationalization is not only how we implement a study and how we measure predictors and outcome but also two types of hypotheses: One describes the aim of your research in rather everyday language and is called research hypothesis. Examples for such research hypotheses are:

(1) "Listening to music reduces your ability to pay attention to other things."
This is a claim about the causal relationship between two psychologically meaningful concepts. When operationalizing this hypothesis in an experimental design, one group would get a treatment (music; M) whereas the other doesn't (no music; NM). Then a certain outcome is measured. Given that the research question is about attention we design a task assessing that (e.g., the capability of paying attention and detect or recognizing objects that briefly flash in your field of view).

Research vs. statistical hypotheses

research hypothesis = precise and concise form of a research question (like: if X then Y)

examples:

- "Listening to music reduces your ability to pay attention to other things." (claim about the relation btw. meaningful concepts)
- "Intelligence is related to personality." (correlational, not causal; little too broad – choose a dimension)
- "Intelligence is speed of information processing." (invalid, ontological claim + confoundation)

(2) "Intelligence is related to personality."
This is a weaker relational claim about two psychological constructs (intelligence and personality). The reason is that this question is correlational not causal. Whereas we in the first example manipulated a variable (music vs. no music) and AFTERWARDS measured an outcome (attention), is for this research hypothesis not clear whether higher intelligence is the cause or the consequence of personality. There is another weakness with this hypothesis, "personality" is a very comprehensive construct, and therefore we should better formulate: Intelligence is related to (e.g.) openness to experience (i.e., a certain personality dimension).

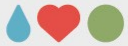
Research vs. statistical hypotheses

research hypothesis = precise and concise form of a research question (like: if X then Y)

examples:

- "Listening to music reduces your ability to pay attention to other things." (claim about the relation btw. meaningful concepts)
- "Intelligence is related to personality." (correlational, not causal; little too broad – choose a dimension)
- "Intelligence is speed of information processing." (invalid, ontological claim + confoundation)

(3) "Intelligence is speed of information processing." This statement has two issues (and I am therefore reluctant to call it hypothesis): First, it ontological claim about the fundamental character of intelligence (what is intelligence). Second, within most intelligence constructs, "processing speed" is an aspect or a factor contributing to intelligence as a whole. We can therefore say that these variables are confounded and can't be measured independent of each other.



Research vs. statistical hypotheses

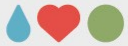
some common problems with research hypotheses:

- "*Love is a battlefield.*" (to vague)
- "*The first rule of tautology club is the first rule of tautology club.*" (can't be falsified)

In addition to these three examples illustrating the basic distinction between causal and (cor-)relational hypotheses (and research designs), there are some common problems or mistake we may make when formulating research hypotheses:
• the hypothesis can be to vague: "Love is a battlefield."

One central problem when hypotheses are not precise and concise is that we have difficulties to convert them into a research design, i.e., to "operationalize" them.

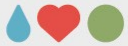
The next hypothesis is formulated in a way such that it can not be falsified: "The first rule of tautology club is the first rule of tautology club."
A statement that is used as a research hypothesis needs two possible outcomes, i.e. there must be a possibility that the claim is correct or wrong.



Research vs. statistical hypotheses

- research hypotheses = scientific claims → statistical hypotheses = claims about data
- "Listening to music reduces your ability to pay attention to other things." → "The group who listens to music will on average have a lower score in a visual attention test." (H_1: $\mu_M < \mu_{NM} \leftrightarrow H0$: $\mu_M \geq \mu_{NM}$)

From research hypotheses, which are scientific claims, we must arrive at statistical hypotheses. Whereas research hypotheses (in psychology) make claims about psychological constructs (and their relation, be it causal or correlational), are statistical hypotheses claims about data. That is, statistical hypotheses convert a description of a relation between psychological constructs into a mathematically precise description. This mathematical description has to correspond to specific claims about the characteristics of the data generating mechanism (i.e., the "population").

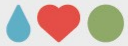


Research vs. statistical hypotheses

- research hypotheses = scientific claims →
  statistical hypotheses = claims about data
- "Listening to music reduces your ability to pay attention to other things." →
  "The group who listens to music will on average have a lower score in a visual attention test."
  (H_1: $\mu_M < \mu_{NM}$ ↔ H0: $\mu_M \geq \mu_{NM}$)

To illustrate this with an example: The research hypothesis: "Listening to music reduces your ability to pay attention to other things." is converted into the a statistical hypotheses, like: "The group who listens to music will on average have a lower score in a visual attention test.". This description formulates a (mathematically precise) description about what we are going to assess in our statistical analyses.

The statistical hypotheses typically formulates or serves as the alternative hypothesis (H_1) for our statistical analysis: "Visual attention scores in the group listening to music are lower than in the group who doesn't: $\mu_M < \mu_{NM}$". This alternative hypothesis has to be contrasted with a null hypo-thesis (H_0; claiming the opposite): "Visual attention scores in the group listening to music are greater or equal than in the group who doesn't: $\mu_M \geq \mu_{NM}$."

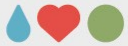


Research vs. statistical hypotheses

- research hypotheses = scientific claims → statistical hypotheses = claims about data
- "Listening to music reduces your ability to pay attention to other things." →
  "The group who listens to music will on average have a lower score in a visual attention test."
  (H_1: $\mu_M < \mu_{NM} \leftrightarrow$ H0: $\mu_M \geq \mu_{NM}$)

We spoke about population and sample before. Please note that the mathematical terms in the hypothesis μ_M and μ_{NM} indicate that, even though we are exploring these hypotheses in a sample, we would like to make a general statement (applying to the whole population, and formulating a general rule) afterwards.

The capacity to formulate good hypotheses is something that is not inborn, but a skill to be learned. As a recommendation: Especially, when you are doing this for the first time, try to formulate these hypotheses and then go to some friends or fellow students and see how much of what you have posited they understand. By doing that, you will over time being able to "refine" your hypotheses and to make them concise and precise.



Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0

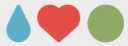
  → strong focus on only one of four outcomes

  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

There is some caveat to statistical testing illustrated in those two schemata. Statistical tests are typically only dealing with the null hypotheses (H0): We assess, how likely it is that we make an error when rejecting the null hypothesis. This probability for an error that we are willing to accept is denoted as α. Unfortunately, that makes us strongly focus on only one of four possible outcomes shown in the schema: The likelihood of making an error when rejecting the null hypothesis, the so called type-I-error. In the schema are also two possible correct outcomes: We reject the null hypothesis when it is wrong or we retain the null hypothesis when it is correct.



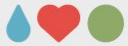
Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0
  → strong focus on only one of four outcomes
  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

Another error often goes unnoticed. It is the so-called type-II-error. We retain the null hypothesis even though it is wrong. The probability of making such an error is denoted as β and occurs if the alternative hypothesis is true (remember: it is either/or, i.e., if the null hypothesis is wrong the alternative hypothesis must be true). However, in order to reject the null hypothesis we need a certain difference (e.g., between two means: $\mu_M < \mu_{NM}$) that enables us to be reasonable sure THAT WE DON'T MAKE AN ERROR when rejecting the null hypothesis. It still might be that this difference is reflecting a true difference in the population (i.e., our alternative hypothesis is correct), but it is not sizeable enough to be sure enough that we don't make an error when rejecting H_0.



Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0

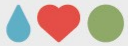
  → strong focus on only one of four outcomes

  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

A critical aspect is that the probability of making an error when measuring is dependent on the sample size. An example for such measurement error is denoted as standard error of mean, indicating how exactly we can estimate the true mean in the population when measuring a sample of the size n.



Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0

  → strong focus on only one of four outcomes

  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

This measurement error is what we base our decision about whether we can reject the null hypothesis or must retain it base upon. As I said, the sample size n is decisive for how large we estimate the standard error to be: The larger n is, the more likely it is that the statistics (i.e., the mean $\bar{x}$) in the sample that we measured, reflects the true value of that parameter (μ) in the population. The larger our sample is, the larger is also which proportion of the population it represents, and the more likely it is that the parameter we estimate for the population is correct. Therefore, we assume the larger our sample is, the smaller the error will be when estimating the parameter in the population. This measurement error is then what we use to assess whether the null hypothesis can be safely rejected.

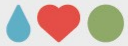
Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0
  → strong focus on only one of four outcomes
  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

As a general rule, the larger our sample is, the more likely it is to get a significant result. This leads to two consequences: If our sample size is small, we might retain the null hypothesis even though it is wrong because our sample size wasn't large enough to be certain enough that we don't make an error when we reject H_0 (remember, the measurement error is larger in small samples and we need a larger difference for safely rejecting the H_0). If you have a very sizable sample (typically, n > 100), we also might reject the null hypothesis even though it is true (remember, we accepted that we make an error according to the defined α-probability, i.e., typically in 5% of all cases).



Research vs. statistical hypotheses

- when testing statistical hypotheses, we focus on how likely error is when rejecting H_0
  → strong focus on only one of four outcomes
  → the error-likelyhood scales with sample size

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

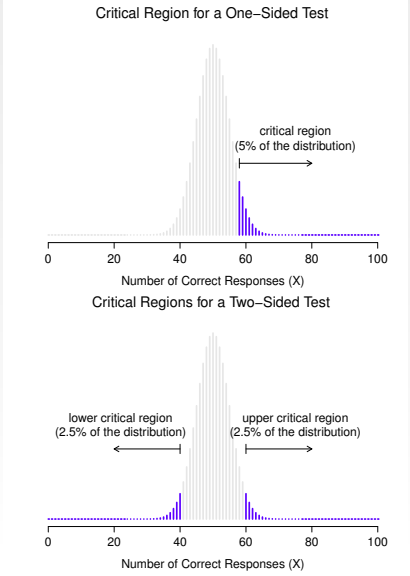
Effect sizes to some degree help to solve that dilemma: The combined information about statistical significance and effect size helps us to assess both how likely we are going to make an error when rejecting the null hypothesis AND whether the difference we obtained (e.g., μ_M vs. μ_{NM} from the example above) is large enough to be practically meaningful. Fortunately, jamovi provides effects sizes to most (if not all) statistical analyses which it includes (often, when using SPSS you had to calculate them manually).

UNIVERSITY OF BERGEN

Research vs. statistical hypotheses

general principle of statistical tests:

1) choose α level (e.g., α = .05)
2) choose test statistic (e.g., $\bar{x}$) to compare H_0 and H_1
3) determine sampling distribution if H_0 were true
4) calculate the critical region, given the α-level and one- or two-sided



A already gave a theoretical explanation about what our statistical tests are based upon: The probability of making an error when rejecting the null hypothesis which is in direct relation to the measurement error we must expect to make when collecting data from a sample of size n.

Our hypothesis test is therefore essentially complete:

(1) we choose an α level (e.g., α = .05);

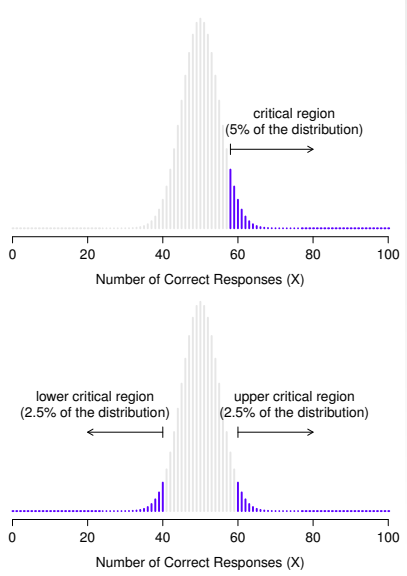
(2) we come up with some test statistic (e.g., $\bar{x}$) that does a good job (in some meaningful sense) of comparing H_0 to H_1

UNIVERSITY OF BERGEN

Research vs. statistical hypotheses

general principle of statistical tests:

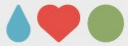
1) choose α level (e.g., α = .05)
2) choose test statistic (e.g., $\bar{x}$) to compare H_0 and H_1
3) determine sampling distribution if H_0 were true
4) calculate the critical region, given the α-level and one- or two-sided



(3) we figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true; and
(4) we calculate the critical region that produces an appropriate α level (considering whether we have a directed hypothesis, e.g., $\mu_M < \mu_{NM}$, when we test one-sided – upper figure – or an undirected hypothesis, e.g., $\mu_M \neq \mu_{NM}$, when we test two-sided – bottom figure).

Fortunately, you were born late enough that our statistics software does the heavy lifting. In the first half of the 20th century, statisticians were carrying out these calculations with paper and pen.

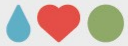
Conceptualizations of the p-value



Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

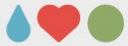
Before we go further, we should discuss the p-value which represents the main result of out statistical analysis. We have two accounts of what the p-value means, one by Sir Ronald A. Fisher the other one by Jerzy Neyman. You should include both of them in your evening prayer (for providing you with so much fun… ;-): Fisher originally came up with the idea for null-hypothesis-significance-testing [NHST] and Neyman later extended and refined that concept (e.g., adding the idea of a confidence interval).



Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

The main difference between the two is that Fisher strongly focused on the null hypothesis (without contrasting it directly with an alternative hypothesis). His line of thought was that the data we collected had to be so extremely implausible according to the null hypothesis that the null hypothesis probably was wrong. According to his reasoning, p is the probability to have observed a test statistic that is at least as extreme as the one we actually did get. His line of thought meant that the null hypothesis provided an account of the data that was so very poor that you could safely reject it. Speaking in terms of the two schemata above, Fisher's focus was mainly on the quarter with the α-probability and the type-I-error.

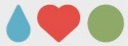


Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

Jerzy Neyman had a stronger focus on the contrast between null and alternative hypothesis. This weighing of two alternatives is reflected in his account of the *p*-value. Here, *p* is the smallest α-(error-)probability that you have to be willing to tolerate if you want to reject the null hypothesis and accept the alternative hypothesis. According to that account, the *p*-value is more of an abstract description about which "possible tests" were telling you to accept the null, and which "possible tests" were telling you to accept the alternative hypothesis.

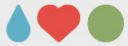
Remember that we do not know the «true» value of a parameter in the population. The confidence interval indicates that if we were to repeated our measurement on numerous samples, calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%.



Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

Null-hypothesis-significance-testing [NHST] came under attack, and voices that demand to not use it at all and replace it with Bayesian statistics became stronger in recent years. One reason is the so-called replication-crisis. It showed that not only the 5% we typically accept as threshold for which error we are likely to make when carrying out NHST can't be reproduced or replicated, but a much larger proportion of studies.

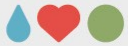


Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

A second reason are considerations (discussed above) about the likelihood of getting significant results varying with sample size and probably not reflecting a meaningful difference.

Finally, increasing computing capacity and capabilities made solving calculations of measures of Bayesian statistics: Often these calculations are carried out using an iterative approach, i.e., a statistical model is refined step-by-step so that it best represents the data you collected. Doing such iterations on paper is virtually impossible and even with computers some ten years ago it was at least very time-consuming.



Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

Bayesian statistics interprets probability as a degree of belief (e.g., a 10% chance that the null hypothesis is true given the data and 90% that the alternative hypothesis is true). Often, a measure called Bayes factor is used to quantify these probabilities: It expresses how much more likely the alternative hypothesis is compared to the null hypothesis and typically denoted as BF_{10}; the "opposite" measure BF_{01} is used much less frequently and expresses how much more likely the null hypothesis is compared to the alternative hypothesis. Please note that for the first measure the subscript "10" means 1 (alternative) vs. 0 (null hypothesis) and "01" meaning vice versa. For example, a $BF_{10} = 9$ is the same as $BF_{01} = 0.11$ (1 / 9) and express a 10% chance that the null hypothesis is true given the data and 90% that the alternative hypothesis is true (10% and 90% being in a 1:9-relation).

Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

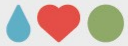
According to a Bayesian approach, the p value is a terrible approximation to the probability that H_0 is true since NHST is fundamentally a frequentist tool and as such it does not allow you to assign probabilities to both hypotheses: H_0 is either true or not, whereas the probability of the alternative hypothesis H_1 being either correct or wrong is not considered.

Conceptualizations of the p-value

- Fisher: strongly focussed on H_0 (not directly contrasting it with an H_1)
- Neyman: stronger focus on the contrast between H_0 and H_1, idea of confidence intervals
- Bayesian: probability as a degree of belief (e.g., 10% chance for H_0, 90% chance for H_1 given the data)
- p: NOT the probability that H_0 is true (but the probability that we make an error when rejecting it)

A last word regarding an interpretation of the p-value that is common but fundamentally WRONG: *p* doesn't express "the probability that the null hypothesis is true". A null hypothesis is either true or it is not, it cannot have a "5% chance" of being true. What we assess is instead how likely we make an error when rejecting H_0. In addition, such claim is also inconsistent with the mathematics of how *p* is calculated.

Significance vs. effect size



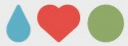
Significance vs. effect size

- good practice: combine significance and effect size
- significance: assumption – “quality” increases with sample size → small difference can get significant
- common effect size measures: Cohen’s d, r, η^2

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

I mentioned that it is good practice to give both, indicators of statistical significance (e.g., t- or F-values and the p-value assigned to that test-statistics) in COMBINATION with an indicator of effect size.



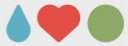
Significance vs. effect size

- good practice: combine significance and effect size
- significance: assumption – “quality” increases with sample size → small difference can get significant
- common effect size measures: Cohen’s d, r, η^2

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

A basic assumption about statistical significance testing is that the “quality” of your test increases the more measurements / participants are included. This is because, typically, the larger your sample to less likely is what you measured due to chance and reflected mathematically in that you have a smaller standard error with increasing sample size. As a consequence, the difference between (e.g.) the two conditions that you compare can be relatively small if you have a large sample of measurements. That is, sometimes, a difference can be significant without being practically meaningful (e.g., results in a students exam are increasing from 21 out of 40 to 22 out of 40). As a consequence, we also often report what is called “effect size”, where the difference between to conditions is set into relation to the variation (measured as standard deviation) within the two groups.



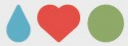
Significance vs. effect size

- good practice: combine significance and effect size
- significance: assumption – “quality” increases with sample size → small difference can get significant
- common effect size measures: Cohen’s d, r, η^2

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

To further illustrate the point with the large sample: If you make a thought experiment where you sample heads or tails when throwing a coin, you can be less certain not to make an error in claiming that your coin is “special” when getting 7 times heads in 10 throws than if you get 700 times heads in 1000 throws. You can try this out in reality as well if you like: with a “normal” coin, i.e., a coin where heads and tail are equally likely, you can try how often you have to repeat your 10 throws before you got at least 7 times heads (on average after about the 6th trial), whereas I can (kind of) guarantee you that you die before having the first success when trying to get 700 times head in 1000 throws (it is so large that computers can’t calculate it, i.e., larger than 10^{38}).



Significance vs. effect size

- good practice: combine significance and effect size
- significance: assumption – “quality” increases with sample size → small difference can get significant
- common effect size measures: Cohen’s d, r, η^2

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

There is a couple of common effect size measures. Cohen’s d is maybe the best known one. It divides the difference between the mean of two conditions by the population standard deviation. Typically, $0.2 < d < 0.5$ is regarded a small effect, $0.5 < d < 0.8$ a moderate effect, and $d \geq 0.8$ a large effect. Correlation or regression models often use r or R. Here, $0.1 < r < 0.3$ is regarded a small, $0.3 < r < 0.5$ a moderate and $r \geq 0.5$ a large effect size. We will discuss effect sizes in more complex statistical models (e.g., ANOVAs) when we introducing them. Here, often η^2 (eta-squared) measures are used. These express what proportion of the variance is explained by a certain factor (but are affected if samples sizes per condition are unequal). Often one is more interested in quantifying differences between two conditions (even for factor with more than two steps). In such case, using post-hoc-test that can output Cohen’s d is maybe a wise choice.

Significance vs. effect size

- good practice: combine significance and effect size
- significance: assumption – “quality” increases with sample size → small difference can get significant
- common effect size measures: Cohen’s d, r, η^2

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

A very readable introduction into effect size is provided in this article:
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

In the jamovi-book (Navarro & Foxcroft, 2019), effect sizes are discussed in Chapter 11.7.

Summary

Summary

- Units, variables and values
- Population and sample
- Level of measurement – Variable levels
- Organizing your data
- Descriptive statistics
- Research hypotheses vs. statistical hypotheses
- Conceptualizations of the p-value
- Statistical significance vs. effect size

Let's briefly summarize which theoretical concepts we introduced.

We started with an overview what we actually measure, using units, variables and values to describe that.

Then we turned to populations as the level we typically would like to make claims about and the sample that we select from that population to measure. We further introduced the concepts of descriptive statistics and inference statistics. Finally, we introduced methods to select samples so that they are representative for the population.

Afterwards, we spoke about different levels of measurement – nominal, ordinal, interval and ratio.

I then stressed why it is sensible to organize your data and how this could be done.

Summary

- Units, variables and values
- Population and sample
- Level of measurement – Variable levels
- Organizing your data
- Descriptive statistics
- Research hypotheses vs. statistical hypotheses
- Conceptualizations of the p-value
- Statistical significance vs. effect size

Afterwards, we used an example dataset and carried out a descriptive statistic analysis in order to describe the dataset and assess assumptions (normality) to use inference statistics with these data.

Then, we spoke about research and statistical hypotheses. What makes good research hypotheses and how do we “transform” research into statistical hypotheses that can be tested.

The p-value is very central to hypothesis testing. Therefore, different conceptualizations of the p-value were introduced: Fisher, Neyman, and Bayes.

Finally, we discussed the problem that we accept a certain probability of error when doing hypothesis tests and that with large samples, tests might (relatively) easily become significant. Effect sizes were introduced as concept to counter that and assess to what degree the observed differences have practical significance.

Thanks for now!
To be continued...

UNIVERSITY OF BERGEN