EFR summary

Applied Statistics 1, FEB11005X

2024-2025

Lecture weeks 1 to 2

Deloitte.

DeNederlandscheBank
EUROSYSTEEM

Details

Subject: Applied Statistics 1 IBEB 2024-2025

Teacher: Michel van de Velden

Date of publication: 17.01.2025

Applied Statistics 1 - IBEB
Lecture 1 - Week 1

Introduction

In science, we run various experiments and collect data - qualitative or quantitative observations of the objects we want to study. Statistics is the science of learning from this data. The goal of statistics, particularly in psychology, is to be able to make predictions about a population based on a sample.

In order to be able to do statistics, you must begin with a set of data. There are many components that make up a set of data:

- Cases are the objects described by a set of data
- A variable is a particular trait of a case
    - A qualitative or categorical variable is a variable that does not have a numeric value. They place cases into groups of categories
    - A quantitative variable is a variable that has a numeric value
- A label is a special variable used to distinguish between cases
- A value is something that a variable holds
- A distribution of a variable tells us what values the variable takes, and how often these values occur for that particular variable

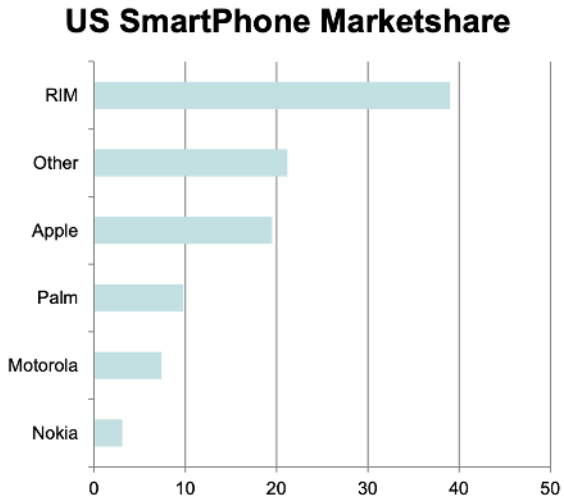
An example to explain this: If you think about a list of psychology students in Amsterdam, each student is a case; characteristics about them, such as age, sex, year of study, etc. are variables. Each student has a student number - this is a label. If a student. for example, were 24 years old, 24 would be the value associated with the variable "age."

Different variables are measured with different instruments. You need to make sure that every variable actually measures what you want it to measure. A poor variable choice can lead to wrong and misleading conclusions. If you find that the variables in your data set do not align with the goal of your research, it is also possible to create a new variable by adjusting another variable.

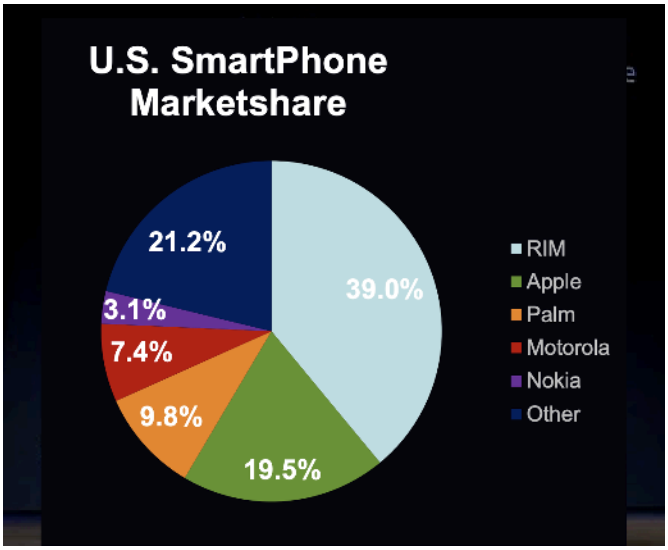
Graphing the distribution of categorical variables

For categorical variables, bar graphs and pie charts are used because of the nature of the distribution of these qualitative variables. The distribution of categorical variables lists each category and gives either the count or percentage of cases that fall in each category; these distributions can then be turned into bar graphs or pie charts.

- A bar graph lists each category (in any order) along the X-axis and the count along the Y-axis. Bars for each category are then drawn according to the count of each category. While the categories can be listed in any order, you should consider presenting your data in an order that makes sense to you and fits the purpose of your research. A bar graph whose categories are ordered from most frequent to least frequent is called a Pareto chart.



- A pie chart shows the percentages of the total count that each category takes up.Here, it is important to include every category, so that the total of the percentages is always 100%. In some cases, when specific categories have very low counts, it is acceptable to include an "Other" category on the pie chart.



Graphing the distribution of quantitative variables

For quantitative variables,stemplots (or stem-and-leaf plots) and histograms are used.
In a stemplot, each observation is separated into a stem and leaf. The stem is everything except the last digit in the value and the leaf is simply the last digit. In a vertical column, the stems of the dataset are written from least to most, and each leaf is written in the row of its corresponding stem, in ascending order. This allows us to have an overview of our dataset.

Example: For 20 employees, "De Bijenkorf" in Rotterdam collects the number of sales made by each employee during one day. For a certain day we have the following data:

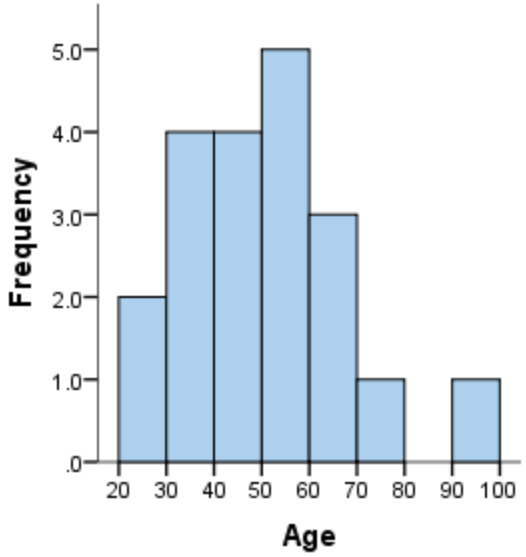
6 9 10 12 13 14 14 15 16 16 16 17 17 18
18 19 20 21 22 24

Stem	Leaf
0	69
1	02344566677889
2	0124

- If you would like to compare two distributions for the same variable, for example the IQ scores of boys vs. girls, you can construct a back-to-back stemplot, with common stems and leaves on either side of the stem.

- One can use a split stemplot to double the numbers of stems when all the leaves would otherwise fall on just a few stems. They split each stem into two: one for leaves from 0-4 and one for leaves from 5-9.

A histogram separates the range of values into classes of equal width and shows the count or percentage of each class, similar to a bar graph. In a histogram, any number of classes can be used; however it is important to use classes of equal width.



- In a histogram, we react to the area (size) of the bars in the graph. By using bars of the same width, we ensure that all of the classes are fairly represented.
- You have to find the right amount of classes and ranges to make an aesthetically representative graph. Too many classes may result in a "skyscraper" effect, while too few may lead to an overly flat graph.

While bar graphs and histograms share many characteristics, there are several notable differences:

Bar Graphs	Histograms
used for qualitative or categorical variables	used for quantitative variables
compare the counts of different items	show the distribution of counts of a variable
do not need to have a measurement scale on the X-axis	use a continuous scale along the X-axis

compare the counts of different items	do not have spaces between bars

By plotting your data, you can make statistical graphs to help you understand your data. In examining your graph, there are several features you should pay attention to. The tails of a graph refer to their extreme values of distribution. The higher values make up the right tail or high tail, and the lower values make up the left tail or lower tail.

In any graph of date one must look at the shape of the graph and try to see an overall pattern:

Center is the midpoint of the data and the spread is the range that the data covers. Once can describe the overall pattern of a histogram by its shape, center and spread.

Individual data points that fall outside the overall pattern are called outliers. These are identified by using your best judgment and it is important to search for explanations behind these outliers. Remember to look beyond just the extreme data points. In some cases, outliers are useful in pointing out mistakes that were made during the experiment, for example: errors in recording, malfunctions in equipment, or other unusual circumstances.
Modes are peaks in the data. Distributions that have one main peak are called unimodal.

- When the right and left sides of the histogram are approximately mirror images of each other then the distribution is symmetric.
- The distribution is skewed to the right (also called skewed toward large values) if the right tail is much longer than the left tail.
- The distribution is skewed to the left if the opposite is true.

It is always a good idea to collect data collected over time in chronological order. This is to avoid misunderstandings, as statistical displays that ignore time as a variable (histograms and stem plots) do not clearly show a systematic change over time.
A time plot of a variable plots each observation against the time at which it was measured.Time is always plotted on the horizontal (x) axis and the variable measured over time is plotted on the vertical (y) axis.

Statistical description of data

Central Tendency

While graphs are a good way to get an overview of your data, numerical descriptions are much more specific. It is important to remember that these numbers, like graphs, are tools to help us understand and interpret the data.

The numerical description of any dataset begins with a description of the middle. There are two common ways to describe the midpoint of a distribution; the mean and the median.

Mean

The mean is the average value of all your data points. To find the mean $\bar{x}$, for a set of observations, you simply sum all of their values and divide by the total number of observations. Thus, for a data set, $x_1+x_2+x_3\ldots+x_n$, the mean can be found using the following equation:

$$\bar{x} = (x_1+x_2+x_3\ldots+x_n)/n$$

From this we can derive a more compact expression

$$\bar{x} = \frac{\Sigma x_i}{n}$$

In this formula, $\sum$ denotes the function "sum". The bar over the x signifies the mean of all the x-values.

The main disadvantage of the mean is that it is very sensitive to extreme values in the data set, and skewed distributions will undercut the integrity and accuracy of using the mean as the midpoint of your data. Because the mean cannot help but be influenced by these extreme values, it is not a resistant or robust measure. Robust measures are not easily influenced by a few data points.

Median

The median is the literal midpoint of a distribution. Half of the observations in a dataset fall above, and half fall below the median. To find the median:

1. Order all observed values from smallest to largest.
2. If the number of observations is uneven, the median is the observation in the exact centre of the list. The median can be found by counting (n+ 1)/2 observations up from the bottom of the ordered list.
3. If the number of observations is even, the median is the mean of the two centre observations. The location of the median is again (n+1)/2 from the bottom of the list.

If a distribution is completely symmetrical, then the median and mean are the same thing. In a distribution that deviates to the left or the right, the average is located in the tail more than the median. This is because the mean is much more affected by extreme scores. The tails of a distribution consist of extreme scores.

The simplest numerical description of a distribution should consist of a measure of the midpoint (such as the average and the median), but also a measurement of the spread of a distribution.

Spread

We can describe the spread of a distribution by calculating various percentiles, the median splits the distribution exactly in half, and that is why we say that the median is the fiftieth percentile. However, there are also upper and lower quartiles on either side of the median. Each quartile is about a quarter of the data.
Quartiles can be calculated as follows:

1. First put all scores in increasing order. Then, calculate the median of the data set.
2. The first quartile (Q1) is the median of the lower half of the distribution. Its position is to the left of the location of the overall median.
3. The third quartile (Q3) is the median of the higher half of the distribution. Its position is to the right of the location of the overall median.

The p th percentile of a distribution is the value by which p percent of the scores is the same or below it.
The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile and the largest observation. So the five-number summary is:

Minimum Q1 M Q3 Maximum

These five values are clearly visible in box plots:

- The outer two edges of the box in a box plot stand for Q1 and Q3.
- The median is shown by the line in the middle of the box.
- Two lines (upwards and downwards from the box) show the maximum and minimum values.

An overview of the largest and the smallest value says very little about the variation within the data. The distance between the first and the third quartile is a more robust measure of spread. This distance is referred to as the interquartile range (IQR), and is calculated as follows: IQR= Q3 - Q1

Quartiles and the IQ are not affected by changes in the tail of a distribution; they are quite robust. However, no single numerical value of dispersion (such as the IQR) is very useful to describe the spread of skewed distributions (left or right). It is often possible to detect skewness using the five-number summary. A deviation to the left or right can be seen by looking at how far the first quartile and the lowest score are from the median (left tail) and by looking at how far the third quartile of the highest score is (right tail).
The standard deviation measures the spread of the distribution to be by looking at how far the observations are from the mean.
The variance (s^2) of a data set is the average of the standard deviations, squared.
The formula is : $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$

Another correct formula is $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

The variance and standard deviation measure the distance between the observations and the mean. Since some observations fall above and some observations below the mean, squaring all the values will make all of the variances (and consequently, standard deviations) positive. Therefore, s^2 and s will be large if observations are widely spread about the mean, and small if the observations are relatively close to the mean.

The standard deviation is particularly useful in normal distributions. The standard deviation is preferred over the variance because finding the square root of the variance ensures that spread is measured according to the original scale of the variable.

Some important properties of the standard deviation:

- Standard deviation, s, is a measure of the dispersion from the mean, and should only be used if the mean (and not the median) is chosen as a measure of midpoint.
- s = 0 when there is no spread present in a distribution. This only happens if all values are the same. If this is not so, which standard deviation is greater than zero. The more there is spread, the greater will be s.
- The standard deviation, like the mean, is not robust. The presence of outliers can make them very large. The standard deviation is even more sensitive to extreme scores than the mean.
- s has the same units of measurement as the original observations.

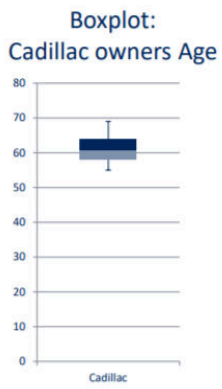
Distributions with a strong deviation (left or right) have large standard deviations. In this case, it is not very practical to calculate the standard deviation. The five-number summary is often more suitable than the average and the standard deviation when an abnormal distribution needs to be described or when a distribution has extreme outliers. The use of the mean and the standard deviation is just more convenient when there are few outliers present and if the distribution is symmetrical.

Boxplot

A boxplot is a graphical representation of the distribution of a dataset. It provides a summary of a data set's minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values. The plot is particularly useful for visualizing the spread and skewness of the data, as well as for identifying potential outliers.

- Age of Cadillac owners:

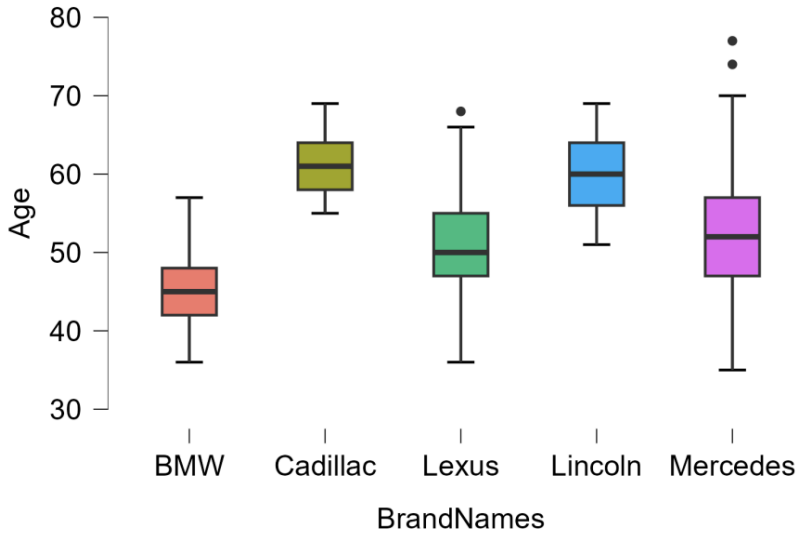
	Age
Min	55
Q1	58
Median	61
Q3	64
Max	69



Key Parts of a Boxplot:
Q1 (First Quartile): The value below which 25% of the data fall.
Q2 (Median): The middle value, dividing the dataset into two halves.
Q3 (Third Quartile): The value below which 75% of the data fall.
Whiskers: Show the spread of data outside the quartiles, typically to 1.5 * IQR.

Outliers: Data points outside the range defined by the whiskers.

Purpose and Benefits:
-Boxplots make it easy to understand the spread and symmetry of the data.
-You can use multiple boxplots side-by-side to compare different datasets.
-Outliers are easily identifiable in a box plot.



Applied Statistics 1 - IBEB
Lecture 1.2 - Week 1

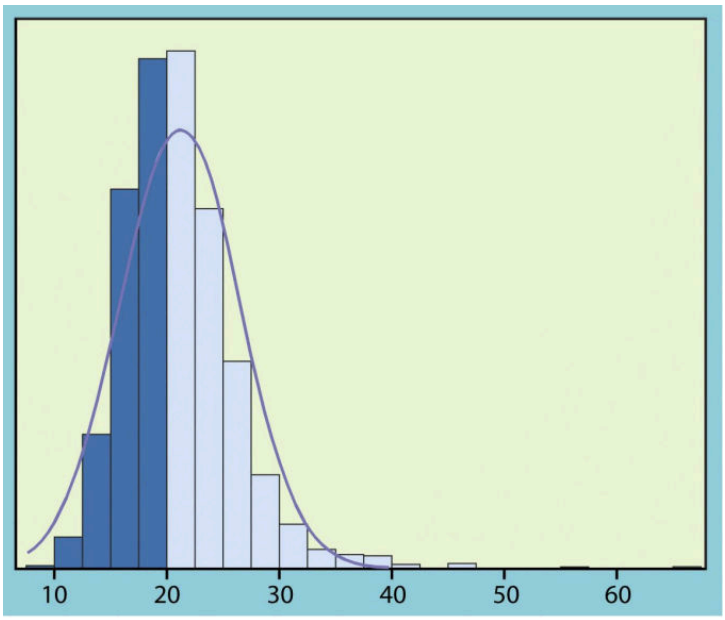
Density Curves

Definition

Because the manual creation of histograms is time consuming and impractical, scientists often use computer programs to create histograms. The advantage of using computer programs is that they can also make an appropriate curve on the basis of the histogram.

These are called density curves. Density curves "flow" with the peaks of a histogram and are a mathematical model for a distribution.

- A density curve is always made on or above the horizontal axis.
- The total area within the curve is always equal to 1.
- A density curve describes the general pattern of distribution.



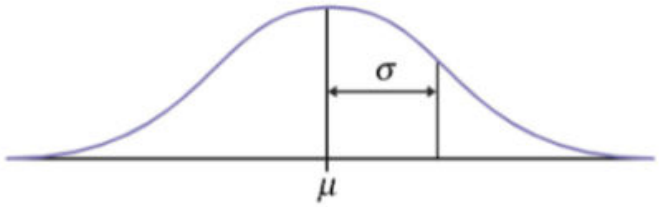
The median of a density curve is the point that divides the area under the curve in half; the equal-areas point. The mean of a density curve is the balance point at which the curve would balance if it would be made of solid material. The median and the mean are equal for a symmetric density curve. The average of a different distribution lies more in the direction of the long tail, while the median lies more in the direction of the peak.

As with distributions, density curves can have different shapes. A special variant is the normal distribution, in which both halves of the curve are symmetrical. Outliers are not described by a density curve.

Normal Distribution

Normal distributions are an important subset of density curves. They are unimodal, symmetrical, and bell-shaped. The mean and standard deviation determine the shape of a normal distribution:

The mean of a curve is indicated with the letter . Changing y (while the standard deviation is unchanged) will ensure that the position of the curve moves on the horizontal axis, while the distribution remains the same.



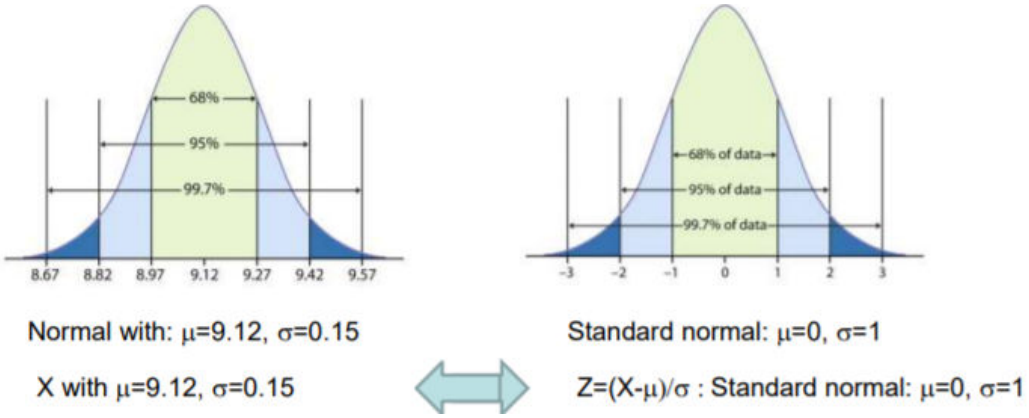
The standard deviation is represented with the symbol σ. The standard deviation is the measure of dispersion associated with a normal distribution. A curve with a larger standard deviation is wider and lower.

- Normal distributions are good descriptions of real data. Many real-life examples of data are normally distributed, including distributions of height, weight and IQ.
- Normal distributions are good approximations of the outcomes of probability calculations, for example in the case of tossing a coin.
- Normal distributions are useful because many statistical inference procedures are based on normal distributions

The 65-95-99.7 Rule: ina normal distribution with mean μ and standard deviation σ:

- Approximately 68% of the observations fall within one standard deviation (σ) of the mean (μ)
- Approximately 95% of the observations fall within two standard deviations of the means.
- Approximately 99.7% of the observations fall within three standard deviations of the mean.

The normal distribution with mean μ and standard deviation σ is written as $N(\mu,\sigma)$.



If, for example, someone has scored sixty points on a test, you do not know whether this is a high or low score in comparison to all the other scores. It is therefore important to standardize the value.If x is a score from a distribution with mean μ and standard deviation σ, then the standardized value of x is:
$z = (x-\mu) / \sigma$.

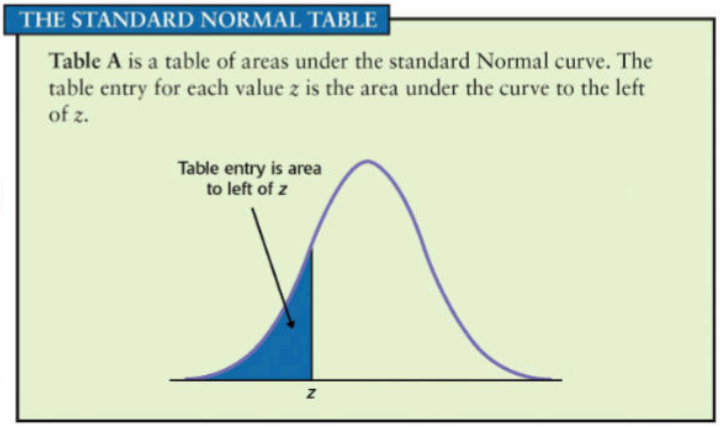
A standardized value is often referred to as a z-score. A z-score tells us how many standard deviations away from the mean a particular observation is, and in which direction. The standardized values of a standard normal distribution have a mean of 0 and a standard deviation of 1. Together, the standardized normal distribution has the N (0, 1) distribution.

The calculation of the proportions in a precise manner within the normal distribution can be done by means of z-tables or software. Z-tables and software often calculate a cumulative proportion: this is the proportion of observations in a distribution that is exactly equal to, or is below a certain value.

The Z-table can be used to determine proportions under the curve. To do this we must first have standardized scores. Suppose you wanted to know how many students had a score above or below 820 on a particular test. Assuming you have a mean score of 1026 and a standard deviation of 209:

- The corresponding z-score would be: 820-1026 / 209 = -0.99.
- Using the z-table, look up the proportion that belongs to -0.99. You will find the p-value to be 0.1611. This area refers to the area to the left of -0.99. The area to the right of -0.99 is therefore 1-0.1611 = 0.8389.
- This means that 16% of the test-takers scored below 820 and below, while 84% of the test-takers scored above 820.

Table A in your book gives probabilities for the standard Normal distribution.



Assessing the normality of data

Stem-and-leaf plots and histograms are often used to see if a distribution is normally distributed. However, the normal quantile plot is the best graphical way to discover normality. It is uncommon to make a normal quantile plot by hand, however in order to understand how software would make one, we would follow these steps:

1. All scores must be put in increasing order. The percentile that each value occupies is then recorded
2. z-values associated with these values must then be found. These are also referred to as normal scores.
3. Each data point is to be graphically connected with the corresponding normal score. If the distribution is (almost) normally distributed, then the data points will lie on an approximately straight line. Systematic deviations from the straight line indicate a non-normal distribution. Outliers are data points that are far from the general pattern of the plot.

Scatterplots

Definition

Relationships between two quantitative variables are often displayed in a scatter plot.

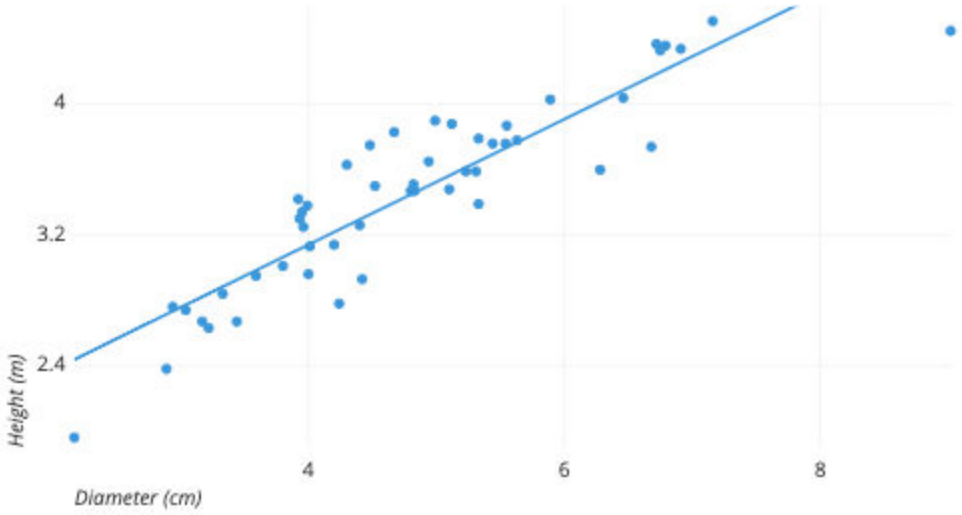
- The two variables need to be measured at the same individuals

- The values of one variable are put on the X-axis, while the values of the other variable are put on the Y-axis. Each individual in the data is processed as a point in the graph, on the basis of the scores achieved by the person on the X-axis and the Y-axis.
- The explanatory variable corresponds to the X-axis. For this reason, the explanatory variable is also referred to as the X-variable. The response (Y) variable will be put on the Y-axis.
- If there is no distinction between explanatory and response variables, then it does not matter, which variable ends up on which axis.

Time plots are a special type of scatterplots, that uses time as an explanatory or x-variable. To get a first impression of a scatter plot, it is useful to:

- Look at the general pattern and deviations.
- Describe the shape, direction, and the strength of the relationship.

Scatter plots can take on many forms and shapes. Many scatter plots show linear relationships; the points lie on a straight line. The strength of a relationship is determined by looking at the degree to which points on the graph follow a specific form such as a line.



The relationship between two variables can be positive or negative.

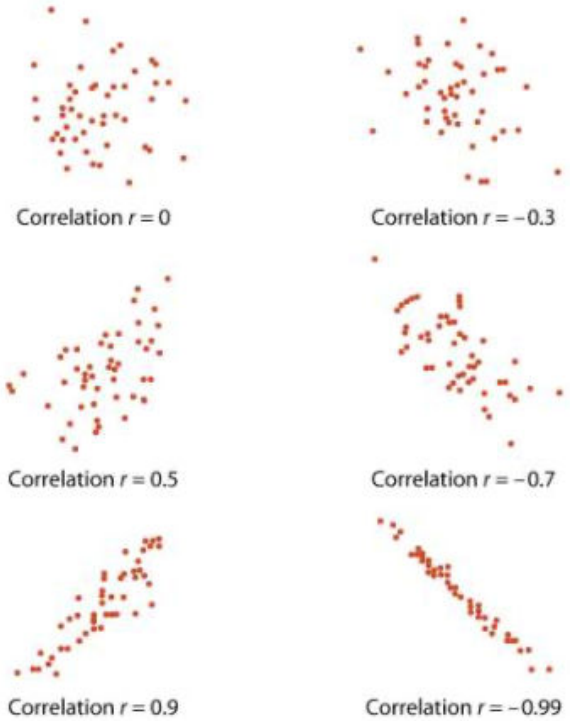
- ❖ Two variables are positively associated when high scores on one variable are associated with high scores on the other variable. An example is that a high score in height is often associated with high scores in weight.

- Two variables are negatively associated when high scores on one variable are associated with low scores on the other variable. For example, there is a negative correlation between test anxiety and performance on an exam. The more test anxiety,the lower the exam score.

Correlation

The scatterplot of a distribution describes the shape, direction, and strength of a relationship between two quantitative variables. It can be misleading to make statements about the strength of this relationship with the naked eye.

By changing the numbers on the axes, any distribution can appear to have a strong correlation. while it might not necessarily be the case. The reverse is also possible. For this reason we use the correlation measure.



The correlation measures the direction and the strength of a linear relationship between two quantitative variables. Often, the letter r is used to describe the correlation.

Suppose we have collected data for variables X and Y for n number of people. The average and standard deviation of the two variables are then € and S for the x-values and y-bar and sy for the y-values.

The correlation, r, between X and Y is: $r = \frac{1}{n-1}\sum\frac{x_i-\bar{x}}{s_x} \times \frac{y_i-\bar{y}}{s_y}$

By using this equation, all of the values for the X and Y variables will be standardised.

Covariance

Correlation is a convenient measure of linear association

$$Cov(x,y) = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Least-squares regression

A regression line is a straight line that describes how a response variable Y changes as explanatory variable X changes.

We often use a regression line to predict the value of Y for a given value of X. For regression,in contrast to correlation, however, it is important that we have specific explanatory and response variables.

The least-squares regression line of x on y is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
The least-squares regression line is:

$\hat{y} = a + bx$

With slope: $b = r\frac{s_y}{s_x}$

and intercept $a = \bar{y} - b\bar{x}$

Even with the best possible regression line, not all of the points lie precisely on the line.
Some items might therefore not be well predicted on the basis of the regression line.
The points that deviate from the regression line are called residuals.

Applied Statistics 1 - IBEB
Lecture 2.1 - Week 2

Least-squares regression: JASP

• Scatterplot:
– Descriptive statistics.
– Basic plots: Tick Correlation plots
– Customizable plots: Scatter plots
• Correlation:
– Descriptive statistics. Correlation (Use Pairwise complete observations)
• Regression:
– Classical>Linear Regression

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M_0	(Intercept)	155.466	3.287		47.295	< .001
M_1	(Intercept)	5.772	5.542		1.041	0.299
	Alcohol(ABV)	2858.198	102.370	0.905	27.920	< .001

The slope is 2858.198:

For each increase of alcohol percentage by 1 unit, the calories per 12/oz increases by 2858.198

Least-squares regression: Residuals and Outliers

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. One thing we must note is that the mean of the least-square residuals is always zero. So in formula:
Residual = observed y - predicted y = y - $\hat{y}$

One can plot the regression residuals against the explanatory variable. Such a scatter plot is called a **residual plot**. When you examine a residual plot you must look at several things:

1. A curved pattern (means relation is not linear)
2. Increasing or decreasing spread about the line

3. Individual points with large residuals
4. Individual points that are extreme in the x direction.

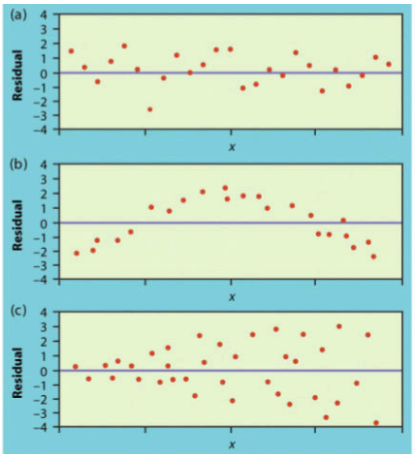
With a residual plot, it can be determined whether a regression line fits well. If the regression line fits the general pattern of the data, no patterns will be present in the residuals. An outlier is an observation that is far from the overall pattern of a residual plot.

Items that are outliers in the Y direction of a scatter plot have large residuals, but this does not necessarily apply to other residuals.

a) Equally spread residuals: the regression line fits well

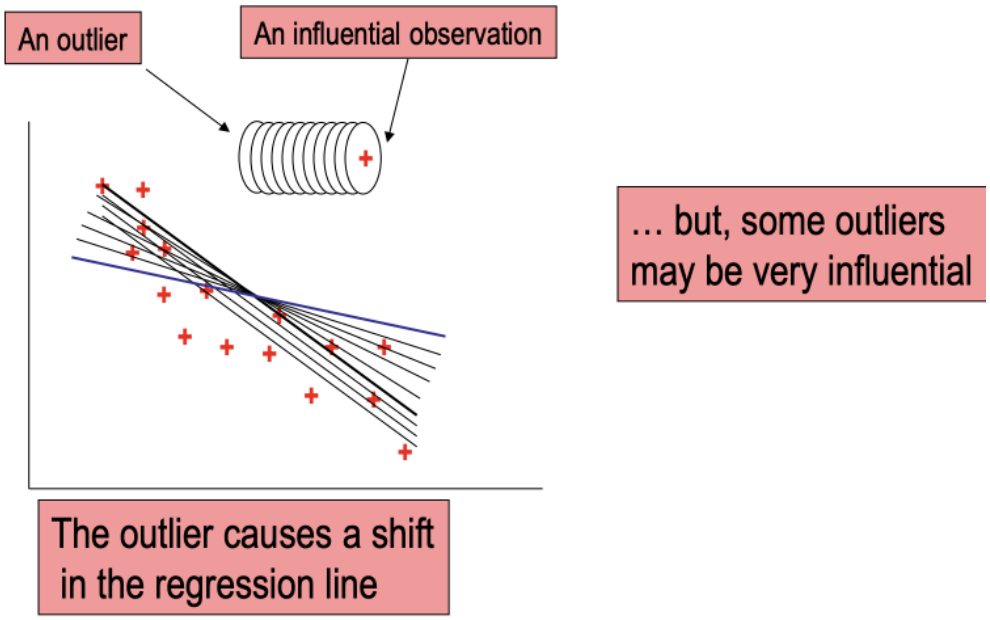
b) Curved pattern: the straight line doesn't fit well. There may be a non-linear (curved) relationship

c) There is more spread in the predictions when the values of x increase. Predictions are less accurate for higher x.



Interpretation

- **Outliers with large residuals** may indicate points that are far from the regression line, suggesting unusual or extreme values in either the x or y variable (or both).
- **Influential outliers** are those that not only have large residuals but also significantly affect the slope or intercept of the regression line when included in the model. These are points that can disproportionately influence the model's fit and predictions.
- **Outliers in x vs. y**: Outliers in the **x variable** (independent variable) can be particularly influential, even if they are not extreme in terms of the **y variable** (dependent variable), because they may exert a strong effect on the model's parameter estimation.



Cautions about correlation and regression

A comparison

	Correlation	Regression
Goal	Measure for strength and direction of relationship between two quantitative variables	Prediction from one variable by another using a straight line
Role variables	Both variables have the same role	There is one response variable y and one explanatory variable x.
	Both measures are sensitive to outliers	

Extrapolation

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line

The relationship between two variables can often be best understood by also looking at the effect of other variables. Lurking variables can make a correlation or a regression misleading.

Lurking variables

A lurking variable is a variable that is not included in the study as an explanatory or response variable, but may affect the interpretation of the relationship between these variables. A lurking variable can falsely suggest a strong relationship between x and y, or it can hide a relationship.

A (strong) relationship between an explanatory variable (X) and a response variable (Y) is not evidence that X causes changes in Y. Correlation says nothing about causality. In addition, it is important to be careful when working with regressions of averaged values.

Association ≠ Causation

Correlation does not imply causation: A strong correlation between two variables does not necessarily mean that one variable causes the other to increase.
To establish causal relationships, experiments are typically required, although this is not always feasible.

In some cases, causation can be inferred without experiments if the following conditions are met:

- A strong association between the variables
- Consistent patterns of association
- Larger values of the independent variable (x) lead to larger effects
- The cause occurs before the effect in time
- The cause is logically plausible

Relations in categorical data

Conditional distributions are the same as the marginal distribution for either variable, meaning that the distribution of one variable does not depend on the value of the other variable. Thus, There is no relation between 2 variables if the conditional distributions are the same as the marginal distribution for either variable.

Marginal Distribution: The distribution of a single variable, ignoring the effect of the other variable. It represents how the values of a variable are distributed across all observations.

Conditional Distribution: The distribution of one variable, given the value of another variable. It tells us how the distribution of one variable changes when we know the value of the other variable.

Simpson's paradox

A situation where a pattern or trend that appears when you look at different groups separately can disappear or even flip when you combine those groups into one larger dataset. In other words, the overall trend in the combined data might be completely different from the trends within the individual groups. This paradox highlights the importance of considering how data is grouped and how grouping can affect the conclusions we draw.

Producing data

Observation vs experiment

Studying samples is one type of observational study. Observational studies are studies in which individuals are observed and variables are measured. There is no intervention and the experimenter does not have an effect on the reactions of the individuals.

In contrast, an experiment is a study in which an intervention is carried out intentionally in order to see how people respond. Experiments are often preferred to observational studies, because we have more control over the variables in experiments.

Confounding

Two variables (explanatory variables or lurking variables) are confounded when their effects on a response variable cannot be distinguished from each other.

Designing samples

The whole group of individuals we want to know about is called a population. Often researchers are interested in how the population looks at certain things. In these cases, sample surveys are given to a random group of people. Sampling means that we study a part of a population to draw conclusions about the entire population.

The design of a sample survey refers to the method used to choose the sample from the population. The proportion of the original sample who actually provide usable data is called the response rate.

Voluntary response sample: consists of people who choose to participate in a survey. These kinds of samples are biased because people with strong opinions tend to respond more frequently.

In order to draw correct conclusions, it is important to apply randomisation techniques in the selection of samples. When the design of a study systematically favors certain outcomes then the study is biased.

Simple random sample (SRS): is a sample where study participants have an equal chance of being actually selected from the population. There are several different random sampling designs.

Probability sample: is a sample chosen by chance. We need to know which samples are possible and what chance each sample is associated with. A probability sample can be simple random or stratified.

A stratified random sample: is often used when there is an investigation of a large population. SRS is often not adequate enough. In order to attract a stratified random sampling the population must first be divided into groups of similar individuals. These groups are called strata. Then, separately for each stratum a SRS is done. The sum of the SRSs make up the full stratified random sample.

Bias

Definition

Bias in the design of a study refers to a systematic error that distorts the results or conclusions by favoring certain outcomes over others. Bias can lead to incorrect inferences, misleading conclusions, and poor decision-making.

There are three main types of bias:

1. Selection Bias
2. Information (Misclassification) Bias
3. Confounding Bias

Selection bias

Selection bias occurs when the sample used in a study does not accurately represent the larger population, leading to distorted or unrepresentative results.

There are several types of selection bias:

- Selection Effects: This occurs when only a specific, non-random subset of data is observed, which does not reflect the entire population. For example, only studying a certain age group might not be applicable to the general population.
- Self-Selection Bias (or Publicity Bias): This occurs when individuals volunteer to participate in a study, leading to over- or under-representation of certain groups. For instance, people without internet knowledge might avoid participating in an online survey, skewing the results.
- Nonresponse Bias: This happens when selected participants fail to respond or provide data. If the non-respondents differ significantly from those who respond, it can lead to biased conclusions.
- Texas Sharpshooter Bias: This happens when patterns are identified after data collection, and then theories or hypotheses are formulated based on those patterns. It is a form of post-hoc reasoning where researchers "find" results by selectively focusing on specific data points that seem interesting or extreme.
- Confirmation Bias: This is the tendency to search for, interpret, or recall information in a way that confirms one's pre-existing beliefs or hypotheses, while ignoring contrary evidence.

Example:

- Conspiracy theories often arise due to confirmation bias, where people selectively search for information that supports their belief, disregarding evidence to the contrary.

- Football coaches or analysts may overemphasize certain statistics (like a lucky goal or a few successful plays) while ignoring the overall performance of a team.

Information (misclassification) bias

Information bias arises when there is a systematic error in how data is collected or measured, leading to inaccurate conclusions. There are different forms of information bias:

- Response Bias: Occurs when respondents do not provide truthful answers due to factors such as social pressure, the way questions are phrased, or how the survey is conducted. For example, people may underreport undesirable behaviors (e.g., smoking or drinking) due to social stigma.
- Recall Bias: This happens when people who have been exposed to a certain factor (such as a health risk) are more likely to recall or remember their exposure than those who have not. This can skew results, particularly in retrospective studies. For example, people with a disease may more easily recall their past behaviors or exposures than healthy individuals, leading to distorted findings.

Confounding bias

Confounding bias occurs when the relationship between two variables is distorted by the presence of another, unaccounted-for variable (a "confounder"). A confounder is an external factor that is related to both the independent variable and the dependent variable, which can lead to incorrect conclusions about the cause-and-effect relationship.

- Example 1: The apparent link between coffee drinking and cancer may be confounded by smoking, as people who drink a lot of coffee may also be more likely to smoke, and it is smoking (not coffee) that increases the risk of cancer.
- Example 2: A study may find that women earn less than men on average, but this observed relationship might be confounded by level of education. If women, on average, have less education than men in the sample, the real cause of the wage gap might be education rather than gender itself.

Applied Statistics 1 - IBEB
Lecture 2.2 - Week 2

Designing experiments

Definition

The individuals that we use for an experiment are called experimental units. When these units are people, we call them subjects.

A specific experimental condition that is applied on the experimental units is called a treatment.

The distinction between explanatory and response variables for experimentation is important because we want to establish causality. Often, this will succeed only with real experiments. The explanatory variables are called factors. Oftentimes, studies look at the combined influence of several factors. In such an experiment, each treatment is formed by combining specific values or quantities of the factors. These specific values are referred to as levels.

Comparative experiments

In many laboratory experiments in science and engineering only one intervention is carried out at a time. This intervention is then applied to all experimental units. Such a set-up is called a comparative experiment and is summarized as follows:

Post-test only one group:

Problem: The placebo effect is a potential issue in this experimental design. When participants are aware they are part of an experiment, they may report

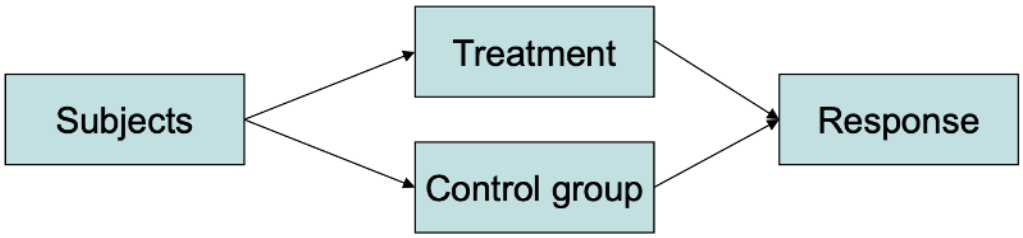
improvements or changes simply because they expect something to happen, even if they did not actually receive the treatment. This psychological response can distort the findings.

- In medical experiments, the placebo effect plays an important role in the validity of the experiments. Simply taking a pill, even if the pill does not contain any of the active ingredients being researched, often influences the behaviours of the test subjects in the placebo group.
    - ➢ Quantity matters. More pills is better!
    - ➢ "Ritual" matters (more invasive, more effective: Needles are better than pills)
    - ➢ Colour matters (Red/orange: alerting, blue/green: sedating).
    - ➢ More expensive placebo are more effective
    - ➢ Placebo's can have side effects!

On the other hand, in a post-test with only one group design, the outcomes observed could be influenced by the placebo effect, making it difficult to separate the actual effects of the treatment from the psychological effects. In this setup, the results of the placebo effect are confounded with the real treatment effects, leading to misleading conclusions.

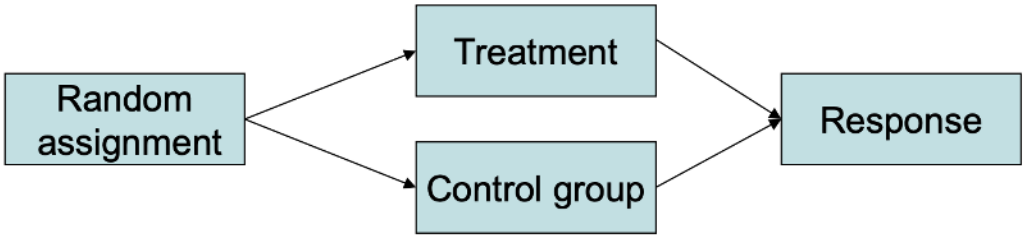
Overcoming the placebo effect

To overcome the placebo effect, researchers introduce a control group in the experiment. This group participates in the study but does not receive the treatment being tested. Instead, the control group typically receives a placebo—a treatment that has no therapeutic effect (e.g., a sugar pill or a sham procedure). By comparing the results from the experimental group (which receives the treatment) with those from the control group, researchers can isolate the true effects of the treatment from the psychological effects caused by expectations.



Issue: both the subjects and the experimenters may know who is receiving the treatment and who is receiving the placebo. This knowledge can introduce bias, as

either the subjects may report changes based on their expectations, or the experimenters may unconsciously interpret results differently depending on the group.
solution: use a double-blind design. In a double-blind study, neither the participants nor the experimenters know which group (treatment or placebo) the participants are assigned to, which helps to prevent unconscious bias from affecting the results.



Basic principles for designing experiments

- the use of a control group to account for the confounding variables
- assign the subjects randomly to the treatments (blindly)
- use many subjects

However, even when these principles are followed, it is still possible for the treatment's effect to appear much larger than expected, resulting in a statistically significant effect. This means that the observed effect is unlikely to have occurred by chance. Despite adhering to these guidelines, experiments may still not perfectly reflect real-world conditions, as controlled environments often differ from the complexities of real-life situations.

Assignment in experiments

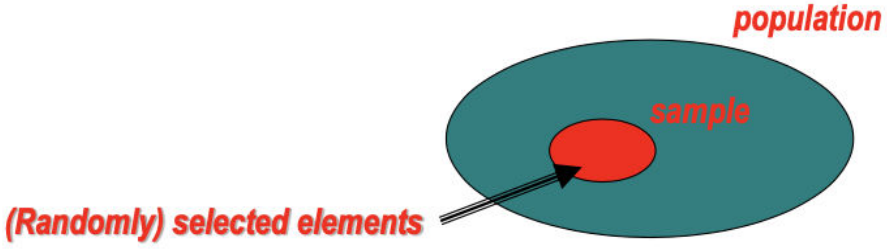
When designing an experiment, researchers need to decide how to assign participants or experimental units to different treatments.

- Completely randomized design: subjects are randomly assigned to different treatment groups. The goal is to ensure that each subject has an equal chance of being placed in any treatment group, minimizing bias and making sure the results are not influenced by any pre-existing differences among subjects.
- Matched pairs designs: two treatments are compared in subjects that are matched based on particular characteristics. This way, subjects in each pair

are similar to each other rather than unmatched subjects. The differences in their responses can then be observed and recorded and further analysed.
    - Example: If testing car tires, we might have two cars run laps on the same track under the same conditions and measure the wear on each tire. Alternatively, we could use the same car twice, each time with a different tire, so that the variation due to the car or the driver is controlled for. By comparing the wear on the two tires while holding the car and driver constant, we can more accurately attribute differences to the tires themselves.
- Block design: researchers make use of so-called blocks. A block is a group of experimental units or subjects that are similar to each other. In a block design, the random assignment of experimental units of treatments done separately for each block.

Population & samples

- The population is the entire group of individuals from which we want information.
- A sample is a part of the population that we actually gather data from. This sample is used to make inferences or conclusions about the entire population.



Statistical inference is using facts about a sample to draw conclusions or make predictions about a population.

Reference list

Van de Velden, M. (2024). Applied Statistics 1 Lecture 1.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92264134

Van de Velden, M. (2024). Applied Statistics 1 Lecture 1.2 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92365423

Van de Velden, M. (2024). Applied Statistics 1 Lecture 2.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92491895

Van de Velden, M. (2024). Applied Statistics 1 Lecture 2.2 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92567483