EFR summary

Applied Statistics 1, FEB11005X

2024-2025

Lecture weeks 1 to 7

Deloitte.

DeNederlandscheBank
EUROSYSTEEM

Details

Subject: Applied Statistics 1 IBEB 2024-2025

Teacher: Michel van de Velden

Date of publication: 22.02.2025

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

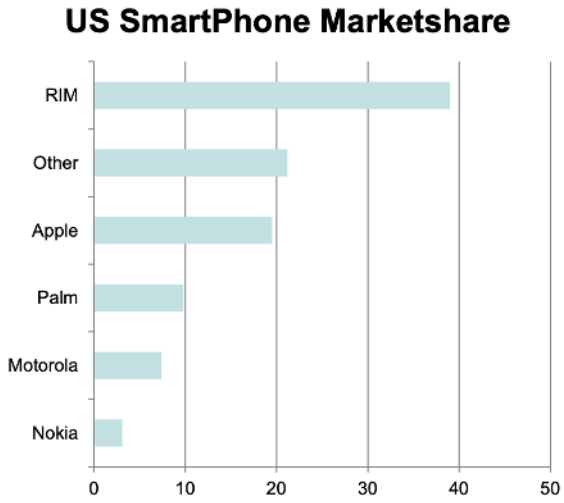
An example to explain this: If you think about a list of psychology students in Amsterdam, each student is a case; characteristics about them, such as age, sex, year of study, etc. are variables. Each student has a student number - this is a label. If a student. for example, were 24 years old, 24 would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

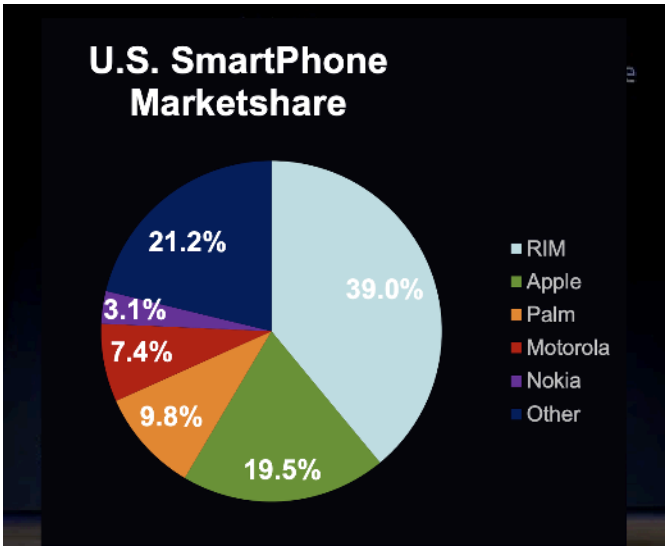
Graphing the distribution of categorical variables

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:

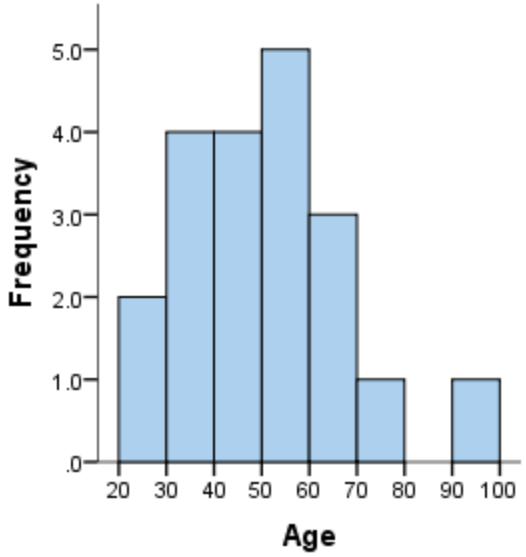
6 9 10 12 13 14 14 15 16 16 16 17 17 18
18 19 20 21 22 24

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.

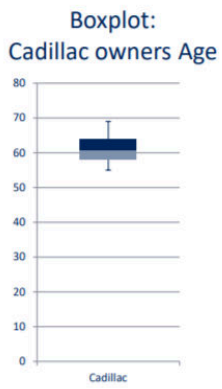
Distributions with a strong deviation (left or right) have large standard deviations. In this case, it is not very practical to calculate the standard deviation. The five-number summary is often more suitable than the average and the standard deviation when an abnormal distribution needs to be described 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:

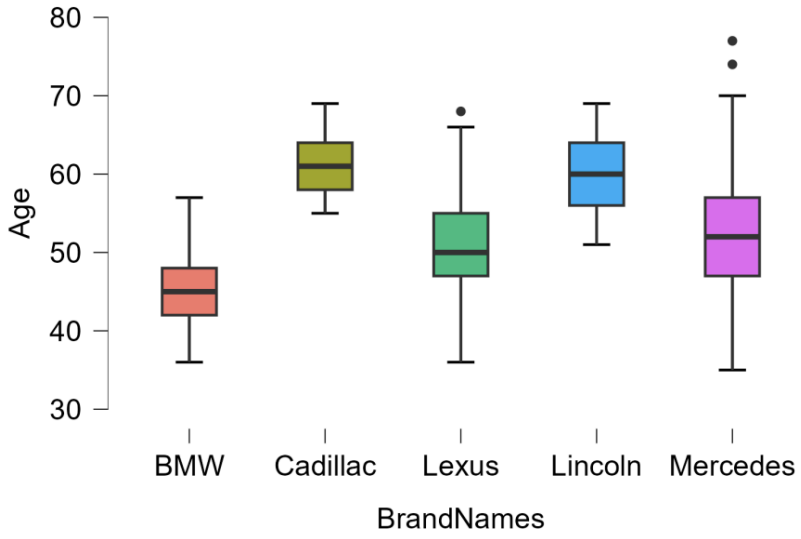
	Age
Min	55
Q1	58
Median	61
Q3	64
Max	69



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

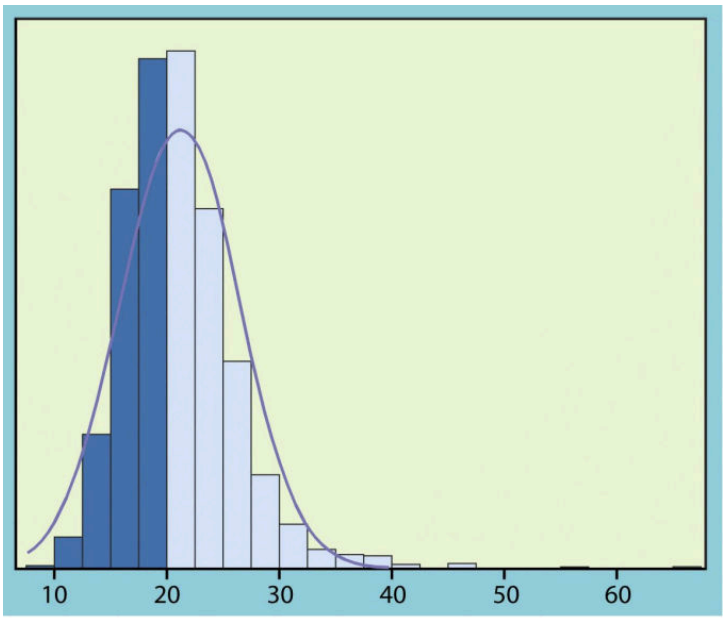
Density Curves

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.



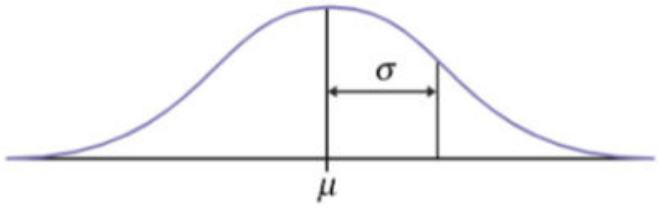
The median of a density curve is the point that divides the area under the curve in half; the equal-areas point. The mean of a density curve is the balance point at which the curve would balance if it would be made of solid material. The median and the mean are equal for a symmetric density curve. The 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.



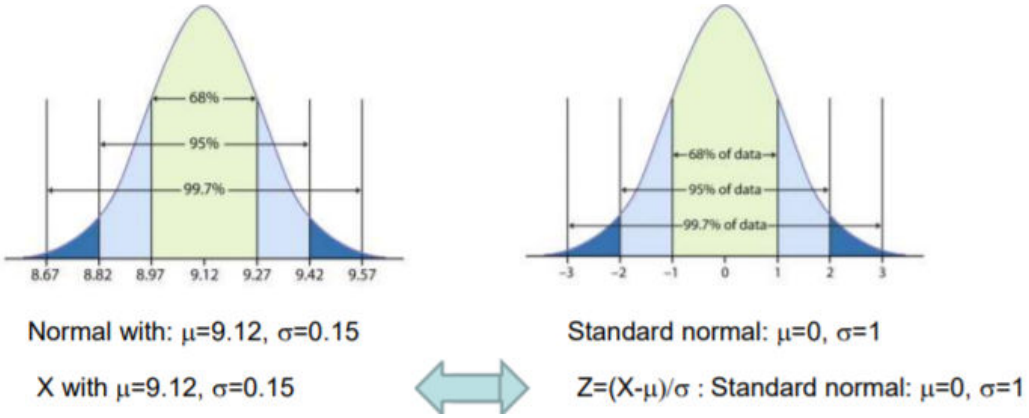
The standard deviation is represented with the symbol σ. The standard deviation is the measure of dispersion associated with a normal distribution. A curve with a larger standard deviation is wider and lower.

- 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:
$z = (x-\mu) / \sigma$.

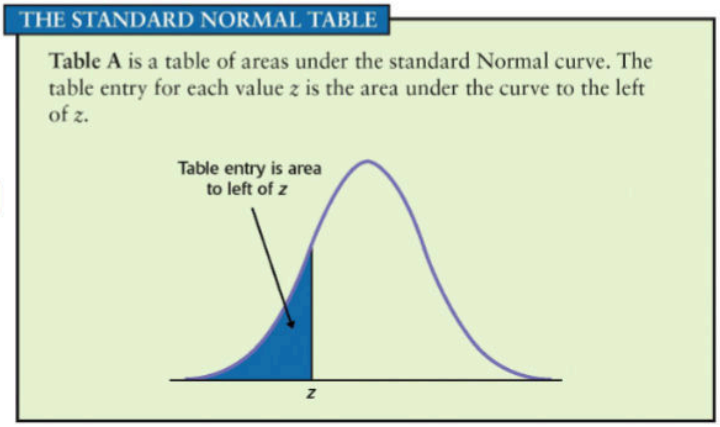
A standardized value is often referred to as a z-score. A z-score tells us how many standard deviations away from the mean a particular observation is, and in which direction. The standardized values of a standard normal distribution have a mean of 0 and a standard deviation of 1. Together, the standardized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

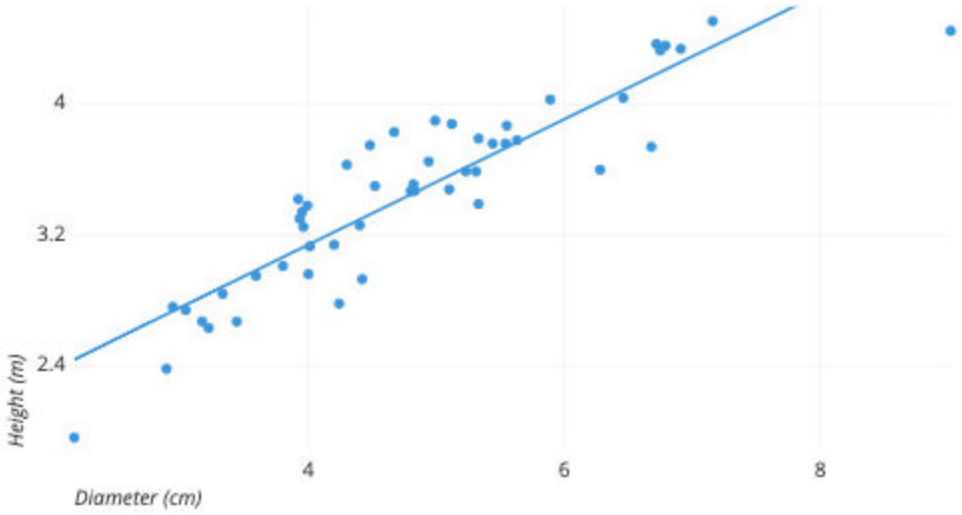
- The two variables need to be measured at the same individuals

- 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.

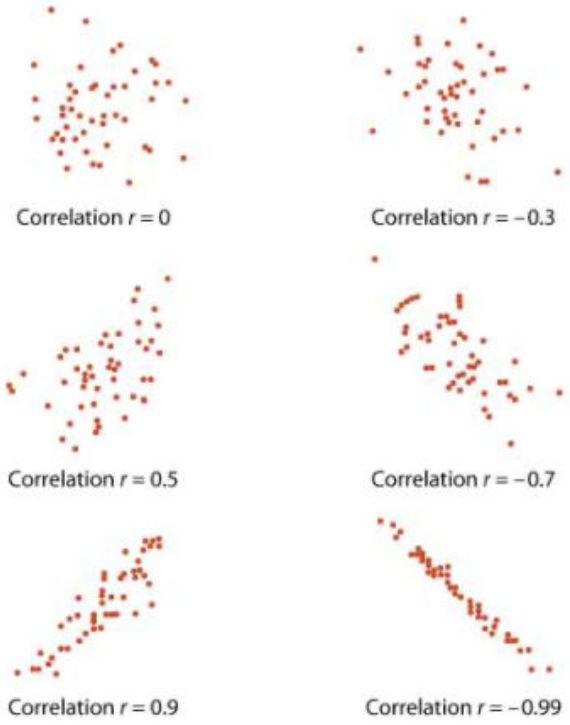
- ❖ Two variables are positively associated when high scores on one variable are associated with high scores on the other variable. An example is that a high score in height is often associated with high scores in weight.

- 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.

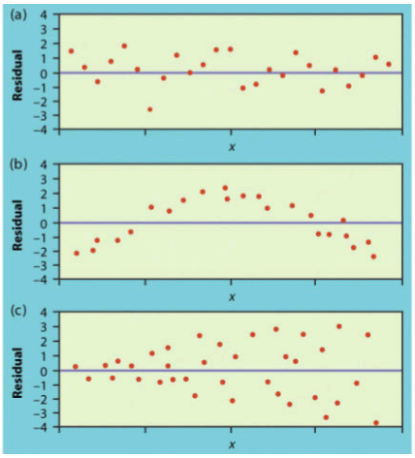
With a residual plot, it can be determined whether a regression line fits well. If the regression line fits the general pattern of the data, no patterns will be present in the residuals. An outlier is an observation that is far from the overall pattern of a residual plot.

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

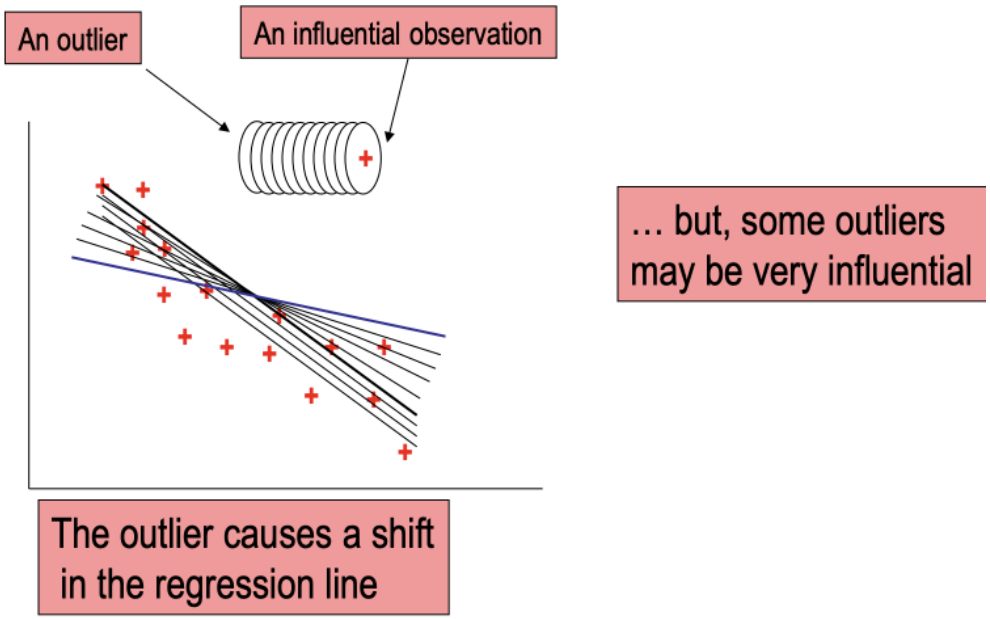
b) Curved pattern: the straight line doesn't fit well. There may be a non-linear (curved) relationship

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

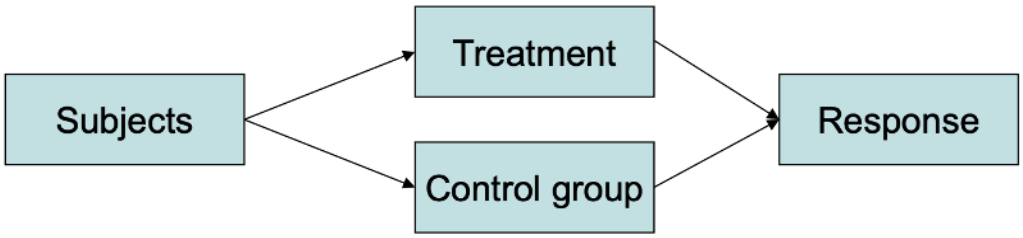
improvements or changes simply because they expect something to happen, even if they did not actually receive the treatment. This psychological response can distort the findings.

- 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.

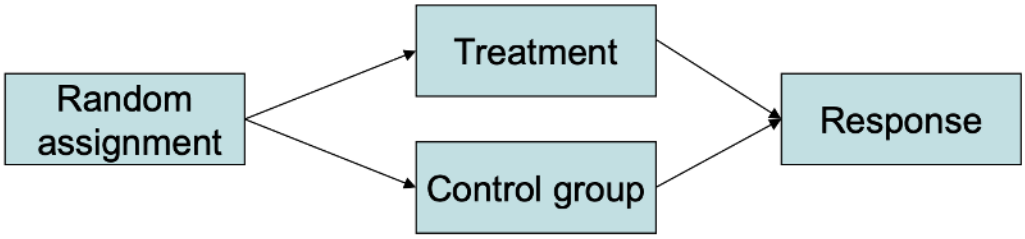
Overcoming the placebo effect

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

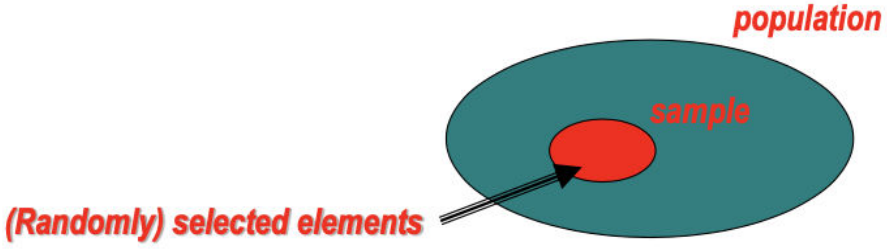
When designing an experiment, researchers need to decide how to assign participants or experimental units to different treatments.

- 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.

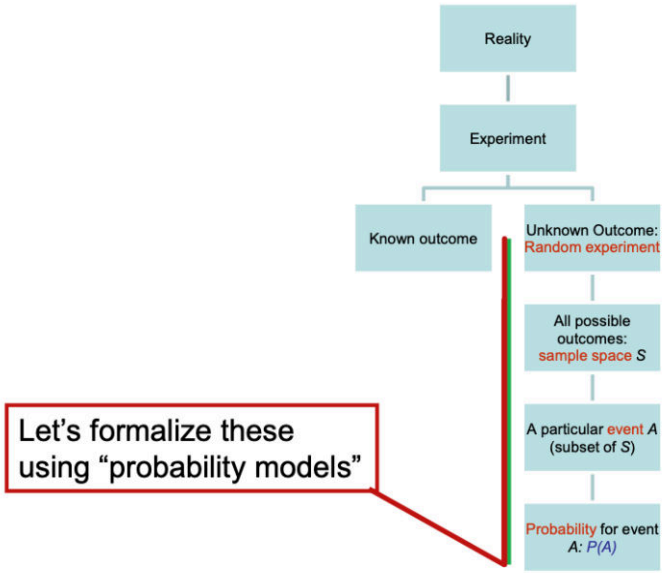
Applied Statistics 1 - IBEB
Lecture 3.1 - Week 3

Randomness

An event is considered random if the outcome of any single trial is unpredictable, yet over the long term, a pattern emerges in the results.
The probability of a specific outcome in a random experiment is defined as the proportion of times that outcome occurs if the experiment is repeated infinitely.
When studying randomness:

- **Independent Trials:** There must be a long series of independent trials, where the result of one trial does not influence the results of others.
- **Empirical Basis of Probability:** Probability is based on empirical observation. While computer simulations can model random behavior using predetermined probabilities, real-world probabilities are best estimated by observing many actual trials. Simulations are valuable because they allow for extended runs of trials, which might not be feasible in real life.
- **Short vs. Long Runs:** Short runs of trials provide only rough estimates of probabilities, while long runs give more accurate approximations.

Schematic way of looking at probability



Probability rules

1. For any event A, $0 < P(A) < 1$.
2. The probability of the entire sample space S is $P(S) = 1$.
3. Complement Rule: The probability that event A does not occur is $P(AC) = 1 - P(A)$, where AC represents the complement of A. For example, if A is the event that a randomly chosen teacher is female, then AC is the event that the teacher is male.
4. Addition Rule for Disjoint Events: Two events A and B are disjoint (mutually exclusive) if they have no outcomes in common and cannot occur simultaneously. In this case,
   $P(A \text{ or } B) = P(A) + P(B)$.

Random variables

A random variable is a variable whose value is a numerical outcome of a random phenomenon. There are two types of random variables:

1. **Discrete Random Variable:**

The variable takes on a finite (or countable) number of outcomes.
Example: If the sample space consists of {1, 2, 3, 4}, X is a discrete random variable. The possible values are separated on a number line.

2. **Continuous Random Variable:**

The variable can take on an infinite number of possible values, often within an interval.
Example: If X represents a value between 0 and 1, the sample space is {all numbers between 0 and 1}. The possible values are not distinct on a number line.

Probability models

Newcomb-Benford's Law states that in many naturally occurring datasets, the leading digits of numbers do not appear with equal frequency. Instead, smaller digits (like 1, 2, or 3) are much more likely to be the leading digit compared to larger digits (like 8 or 9). This pattern diverges significantly from what we might expect if the digits were distributed uniformly (i.e., if each digit had an equal chance of appearing as the first digit).

This law is an example of a specific type of discrete probability distribution. It applies to datasets where the numbers span several orders of magnitude, such as populations, financial figures, and even physical measurements in nature.

Probability distributions

Probability distributions assign probabilities to different outcomes or events.

- Visualizing a probability distribution often involves drawing a density curve, which represents how probabilities are distributed over the range of possible outcomes.
- Every probability distribution is associated with two key measures:

1. The **mean (μ)**, which represents the expected value or average of the outcomes.
2. The **standard deviation (σ)**, which measures the variability or spread of the outcomes around the mean.
    - For **discrete distributions**, these metrics are defined as follows:
        - The **mean (μ)** is the weighted average of all possible outcomes, where each outcome is weighted by its probability.
        - The **standard deviation (σ)** is the square root of the weighted average of the squared deviations from the mean.

Continuous random variable

1. Density Curve:

- The probability distribution of a continuous random variable is represented by a density curve, which shows how probability is distributed over the range of possible values.

2. The probabilities are surfaces below the curve:

- Unlike discrete variables, where probabilities are assigned to specific outcomes, the probability of a continuous random variable is calculated as the area under the curve over an interval.
- For example, the probability that X falls between two values a and b (i.e., $P(a < X \leq b$ is the area under the curve from a to b.

3. Point Probabilities are Zero:

- For any specific value $X = a$, the probability is always zero: $P(X = a) = 0$.

- This is because a single point has no "width" in a continuous distribution, and probability is derived from areas, not individual points.

4. Intervals are Meaningful:
- Since single points have no probability, only intervals of values have meaningful probabilities.
- For example, the probability that X lies between 5 and 10 is meaningful, but the probability that X equals exactly 5 is not.

5. Analogy - Line Segment:
- A continuous random variable is comparable to a line segment:
- The segment itself has a positive length (representing the probability of an interval).
- However, no individual point on the line segment has length (just as P(X = a) = 0 for a single point.

Mean and variance of a discrete random variable

Mean

The mean or expected value of a random variable is defined as the weighted average of the possible values of X, where the weights are the corresponding probabilities of each X_i:

$$E(X) = \mu = \sum_{all\, x_i} x_i . p(x_i)$$

Variance

The variance of a random variable is the weighted average of the squared deviations of the possible values of X from the expected value μ, where the weights are the corresponding probabilities:

$$E\left((X - \mu)^2\right) = \sigma^2 = \sum_{all\, x_i} \left(x_i - \mu\right)^2 . p(x_i)$$

Shortcut calculation:

$$\sigma^2 = E(X^2) - \mu^2 = \sum_{all\ x_i} x_i^2 . p(x_i) - \mu^2$$

The standard deviation is the square root of the variance

Linear combinations

When we create a new random variable as a linear combination of an existing random variable, we can determine how its expected value and variance are affected. Let:

- X be a random variable
- $E(X) = \mu_X$
- Variance X is: σ^2_X
- Y=aX+b: Y is a new random variable, constructed from X

After the transformation the expected value of a random variable and the variance becomes:

$E(Y) = \mu_Y = \mu_{aX+b} = a\mu_X + b$

$V(Y) = \sigma^2_Y = \sigma^2_{aX+b} = a^2\sigma^2_X$

Rules for means

1. If X is a random variable and a and b are fixed numbers:

$$\mu_{a+bX} = a + b\mu_X$$

2. If X and Y are random variables:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

Rules for variances

- Two random variables, X and Y, are said to be **independent** if knowing the outcome of one variable provides no information about the outcome of the other. In other words, the occurrence of X does not influence Y, and vice versa
- If two random variables are independent, their correlation coefficient (ρ) is zero, indicating no linear relationship between the variables. However, the reverse is not always true: a correlation coefficient $\rho = 0$ does not necessarily imply independence.

- Variance of the sum of two random variables X and Y:

R	X and Y independent	X and Y dependent
$X+Y$	$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$	$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$
$X-Y$	$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$	$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$

Discrete probability distributions

	Uniform	Bernoulli	Binomial	Poisson
Sample Space S	1, 2, …, N	0,1	1, 2, …, N	1, 2, …, N
P(X=*k*)	1/N for *k* in *S*, else 0	p, for *k*=1 (1-p) for *k*=0 0 else	$\binom{n}{k} p^k (1-p)^{n-k}$ for *k* in *S*, else 0	$\frac{e^{-\mu}\mu^k}{k!}$ for k in *S*, else 0
Mean μ	(N+1)/2	p	np	μ
Stand. deviation σ	$\sqrt{(N^2-1)/12}$	$\sqrt{p(1-p)}$	$\sqrt{np(1-p)}$	$\sqrt{\mu}$

Discrete uniform distribution

- A discrete uniform distribution is one in which all possible outcomes of a random variable have the same probability of occurring. In other words, each event is equally likely.
- Example: X is the number of eyes showing after a throw of a die.

Bernoulli distribution

- Two possible events: Success or Failure
- Probability for success is p, failure is $1 - p$
- *Example: You do one multiple choice question with 5 options for each question.*

Binomial

- A binomial experiment consists of n independent repetitions of a Bernoulli trial (an experiment with two outcomes: success or failure).
- The probability of success, p, remains the same in each trial.
- The random variable X represents the number of successes (k) out of n trials.

	Repetitions	μ	σ^2
Bernoulli	1	p	$p(1-p)$
Binomial	n	np	$np(1-p)$

  - X Binomial with n_x and p
  - Y binomial with n_y and p
  - X and Y are independent
  - $R = X + Y$ is Binomial distributed with parameters $n = n_x + n_y$ and p

This property holds because The independence of X and Y ensures that the combined trials can be treated as a single larger set of independent trials with the same probability of success p. This allows the Binomial distribution to extend to their sum.

Poisson

Not all count data follows a Binomial distribution. The Binomial distribution requires a fixed number of trials (n), while the Poisson distribution models open-ended counts where the number of observations or trials is not fixed.

The Poisson distribution is appropriate for modeling counts of events in scenarios with the following assumptions:

1. The number of successes (events) occurring in a unit of measure is independent of the number of successes in other non-overlapping units.
2. The probability of a success occurring in a unit of measure is constant for all units of equal size and is proportional to the size of the unit.
3. The probability of two or more successes occurring in an extremely small unit approaches zero as the unit size decreases.

If X is Poisson distributed with μ_X, Y is Poisson distributed with μ_Y, and X and Y are independent. Then: $S = X + Y$ is Poisson distributed with $\mu_S = \mu_X + \mu_Y$

Key Quantity

The Poisson distribution is defined by the mean number of successes per unit of measure (denoted as μ).

For a Poisson random variable X, the probability of observing k successes is:

$$P(X = K) = \frac{e^{-\mu}\mu^{k}}{k!}$$

Multiplication rule for independent events

- Events A and B are independent if the occurrence of one does not affect the probability of the other.

=> $P(A \text{ and } B) = P(A).P(B)$

- If A and B are independent, the correlation between them is zero.
- However, zero correlation does not necessarily imply independence, as non-linear associations might exist.

Sampling distribution of a sample mean

- A statistic calculated from a random sample will vary across samples drawn from the same population.
- These sample statistics are treated as **random variables**.

Law of large numbers

The sample mean converges to the population mean as the sample size increases.

- Population mean μ must be finite.
- Respondents are independent and randomly drawn.

Central Limit Theorem

For a sufficiently large sample size n, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal, regardless of the shape of the original population distribution:

$$\bar{x} \text{ is approximately } N(\mu, \frac{\sigma}{\sqrt{n}})$$

As n increases, the approximation improves.

Applied Statistics 1 - IBEB - Lecture 3.2 - Week 3

General probability rules

1. $0 \leq P(A) \leq 1$ for any event A

A probability of 0 means the event is impossible, and a probability of 1 means the event is certain.

2. The probability of the sample space S, which includes all possible outcomes, is always: $P(S) = 1$
3. Complement rules: the probability that an event A does not occur (its complement, denoted as A^c)

$$P(A^c) = 1 - P(A)$$

The sum of probabilities for an event and its complement is always 1.

4. Additional rule: For two disjoint events A and B (events with no outcomes in common, meaning they cannot occur simultaneously) :

$$P(A \text{ or } B) = P(A) + P(B)$$

Venn diagram

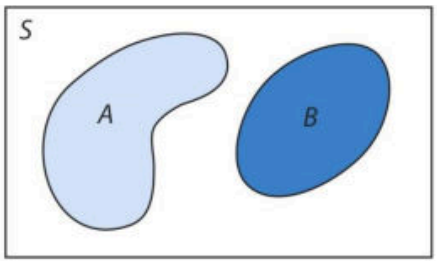
A Venn diagram is a visual tool to represent probabilities, the total area of the diagram represents the sample space (S), and individual regions within the diagram represent different events.

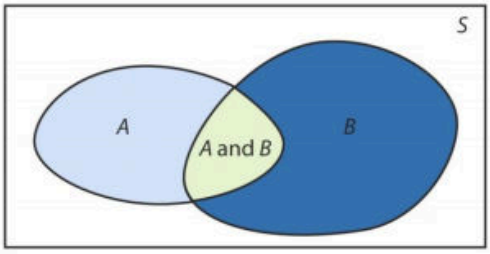
- For two disjoint events: disjoint events are represented by non-overlapping areas.

$$P(A \text{ or } B) = P(A) + P(B)$$

- For two events that are not disjoint: non-disjoint events overlap in the Venn diagram.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$





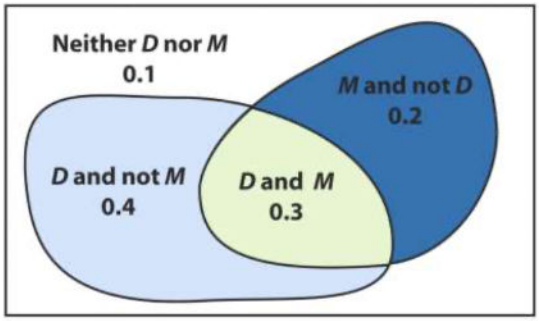
Conditional probability

Conditional probability describes the probability of event B occurring, given that event A has already occurred (assuming $P(A) > 0$). It is denoted as:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$P(B|A)$ adjusts the probability of B based on the knowledge that A has occurred.

Example:



D = Deborah is made partner
M = Matthew is made partner

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad = 0.3/0.7 \approx 0.43$$

Multiplication Rule:

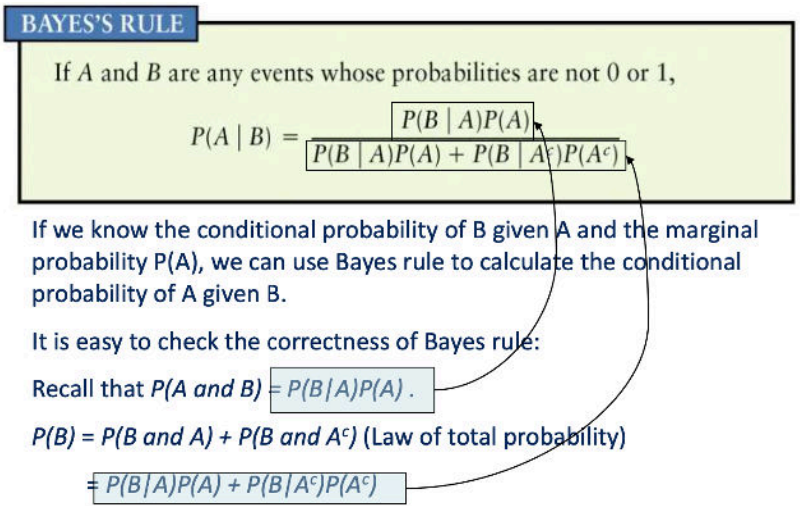
The joint probability of events A and B (i.e., both events occurring together) can be expressed in terms of conditional probability:

$$P(A \text{ and } B) = P(A).P(B|A)$$

Applied Statistics 1 - IBEB - Lecture 4.1 - Week 4

Bayes's rule

Bayes' rule allows us to update probabilities based on new evidence. It provides a method to calculate one conditional probability, such as P(B | A), when the reverse conditional probabilityP(A | B) is known, along with the individual probabilities of A and B. This makes it a powerful tool for refining predictions as new information becomes available.



Decision theory

Decision theory helps in making rational choices when faced with uncertainty. It often relies on probability concepts and visual tools such as tree diagrams to structure and analyze different possible outcomes.

Probability: Conditional

To represent probabilities in a structured way, we use tree diagrams. These diagrams help visualize different possible events, their probabilities, and how they relate to one another.

Expected monetary value

- When making decisions involving financial outcomes, we use the Expected Monetary Value (EMV) to determine the most beneficial choice.
- EMV is calculated as:

$$\Sigma(monetary\ value\ of\ each\ outcome \times associated\ probability)$$

The decision that yields the highest EMV is typically considered the best choice.

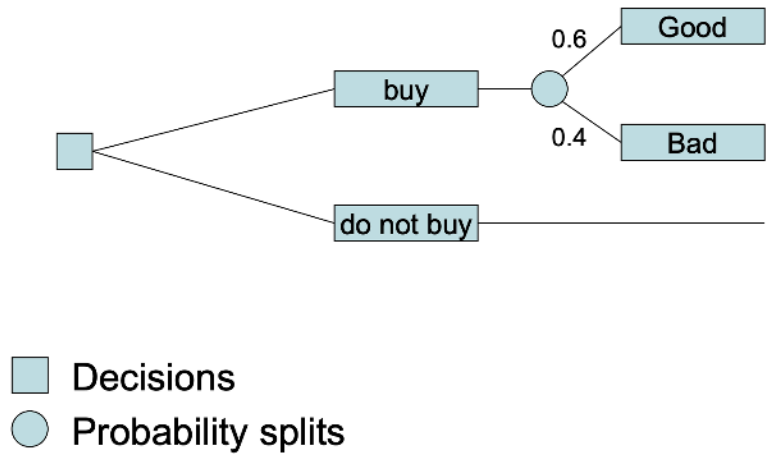
Example:

Suppose we are deciding whether to buy an asset. The possible outcomes are:

- "Good" scenario: Profit of $1,000
- "Bad" scenario: Loss of $2,000

The probabilities of these outcomes are:

- P(Good) = 0.6
- P(Bad) = 0.4



Option 1: Buy the Asset

$$EMV(buy) = (0.6 \times 1000) + (0.4 \times -2000)$$
$$= 600 - 800 = -200$$

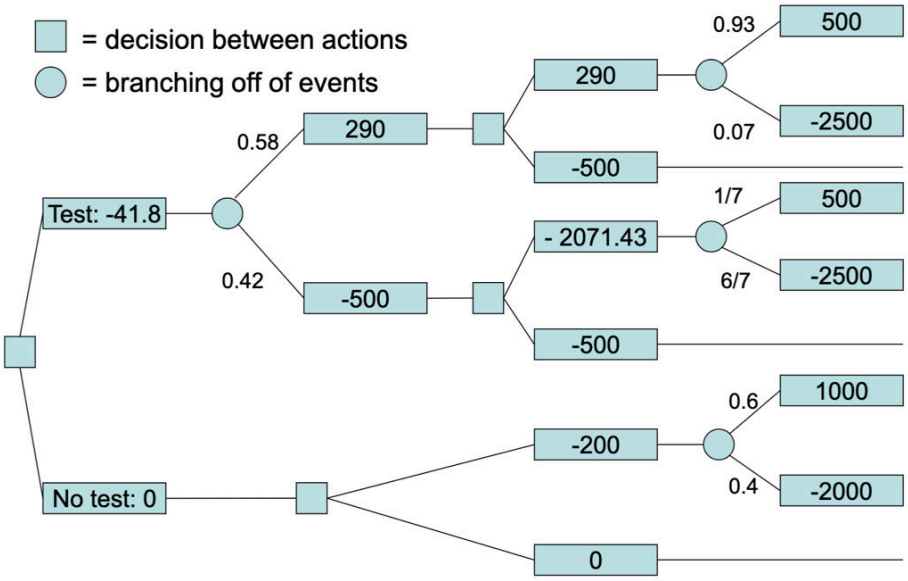
Option 2: Do Not Buy the Asset

$$EMV(do\ not\ buy) = 0$$

=>Since the EMV of buying (-200) is lower than not buying (0), the rational decision is not to buy the asset.

Decision tree

- A decision tree is a visual tool that helps evaluate different choices under uncertainty. It consists of decision nodes, chance nodes, and outcome nodes. The standard approach to solving a decision tree is to work backward (right to left), evaluating expected values at each step to determine the best course of action. The option with the highest EMV is usually selected.

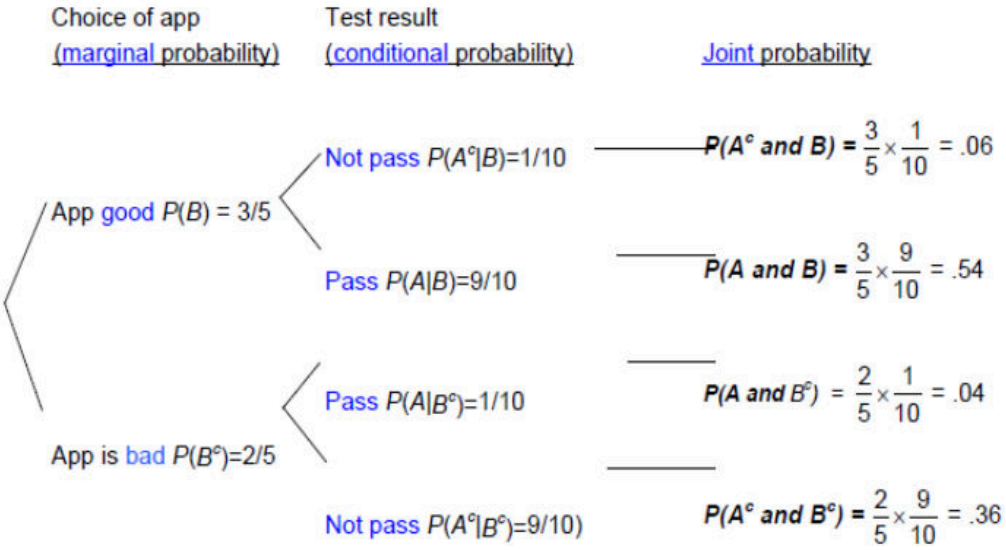


Probability tree

A probability tree is different from a decision tree. While a decision tree is used to determine the best choice among alternatives, a probability tree helps in computing probabilities systematically.

- Probability trees illustrate the sequence of events along with their conditional probabilities.
- They allow us to calculate the joint probability distribution, which represents the likelihood of multiple events occurring together.
- From the joint probability distribution, we can derive both marginal probabilities (probabilities of individual events) and conditional probabilities (probabilities given that another event has occurred).

Example:



Applied Statistics 1 - IBEB - Lecture 4.2 - Week 4

Introduction to inference

Definition

Statistical inference is the process of making conclusions about a population based on data gathered from a sample. The primary goal is to extend observations from a limited set of data points to a broader context, while accounting for inherent uncertainty and variability. This often involves examining the frequency or occurrence of certain events or successes, particularly when the data is collected from non-overlapping regions in space or time.

Statistical inference uses probability theory to quantify and describe the natural variation that occurs within the data. It allows researchers to draw meaningful conclusions while recognizing that any data collected is subject to random fluctuations. In other words, formal inference employs probabilistic methods to

assess the likelihood of various outcomes, ensuring that the conclusions are based on sound statistical reasoning and not simply random chance.

Confidence interval

- A confidence interval (CI) is a range of values that provides an estimate of where a population parameter is likely to fall, based on sample data. It reflects the uncertainty or variability inherent in the sample while offering a plausible range for the true value of the parameter. The interval is typically defined by adding and subtracting a margin of error from the sample estimate (denoted as X).
- Mathematically, the confidence interval is expressed as:

$$[\overline{X} - margin\ of\ error, \overline{X} + margin\ of\ error]$$

- The level of confidence associated with a confidence interval indicates the degree of certainty that the true population parameter lies within this range.
- For instance, a 95% confidence interval suggests that if the same sampling method were used repeatedly, approximately 95% of the resulting intervals would contain the true population parameter.

The sampling distribution of a sample mean

- A statistic derived from a random sample can vary each time we take a new sample from the same population.
- Sample statistics are considered random variables
- The sample mean ($\overline{X}$) is a particularly important random variable. Although it will fluctuate from sample to sample and generally differs from the population mean ($\overline{\mu}$), it tends to provide a reasonable estimate of the true population mean.

Law of large numbers

When independent observations are drawn randomly from a population with a finite mean, as the sample size increases, the sample mean ($\overline{X}$) becomes closer to the population mean ($\overline{\mu}$).

Conditions:

1. The population mean $(\bar{\mu})$ must be finite.
2. The observations should be independent and randomly drawn.

When data are normally distributed

If the population follows a normal distribution $N(\mu, \sigma)$, then the sample mean $\bar{x}$ of n independent observations will also follow a normal distribution $N(\mu, \sigma/\sqrt{n})$, with a mean of μ and a standard deviation of $\sigma/\sqrt{n}$.

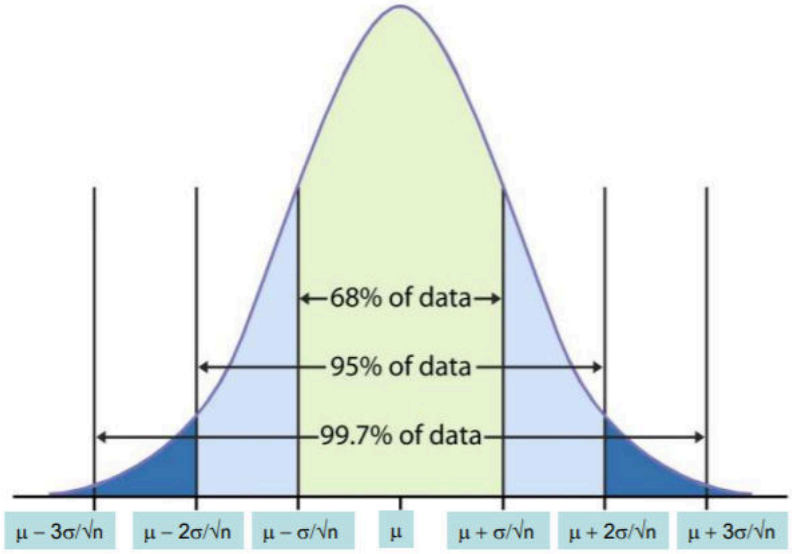
Central limit theorem

The Central Limit Theorem (CLT) states that when an Simple Random Sample (SRS) of size n is taken from any population with a finite mean and finite standard deviation, the sampling distribution of the sample mean $\bar{x}$ will approach a normal distribution as n increases. Specifically, regardless of the original population's distribution shape, the sample mean will be approximately normally distributed as:

$$\bar{x} \text{ is approximately } N(\mu, \frac{\sigma}{\sqrt{n}})$$

Confidence interval

- Confidence interval provides a range of values within which we expect the population parameter (such as the population mean) to lie, based on the sample data.
- To choose a margin of error, we use the approximate distribution of the sample mean.



The most used margin of error is 5%. This gives a 95% confidence level.

Particularly, if we take the interval $[\mu - \frac{2\sigma}{\sqrt{n}}, \mu + \frac{2\sigma}{\sqrt{n}}]$, there is a 95% probability that the sample mean considered is in that interval. In other words:

$$P\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Rearranging yields:

$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

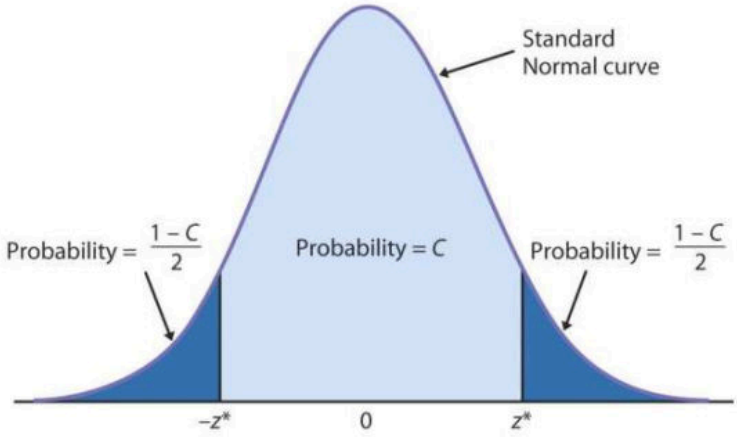
This means that there is 95% confidence that μ is in the interval $[\overline{x} - \frac{2\sigma}{\sqrt{n}}, \overline{x} + \frac{2\sigma}{\sqrt{n}}]$.

Example exercise: Standard deviation is 10 cm. The sample size is n= 400, and the observed sample mean is 182 cm.

- Thus, $x \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(\mu, \frac{10}{400})$
- An approximate 95% confidence interval for μ is [182 – 2*0.5, 182 + 2*0.5] = [181, 183].
- If we were to take 100 samples and construct a confidence interval from each sample. Then, approximately 95 of the confidence intervals capture the true value of μ

General way of obtaining the confidence intervals for the population mean

1. Establish the confidence level C:



2. Pick a SRS of size n with an unknown mean μ and known standard deviation σ. A level C confidence interval for μ is:

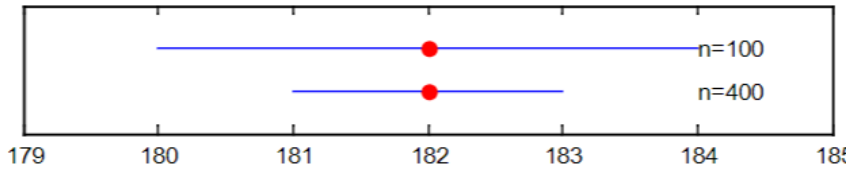
$$C = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- z^* is the critical value with area C between $-z^*$ and z^* under standard Normal curve.
- Margin of error is $m = z^* \frac{\sigma}{\sqrt{n}}$

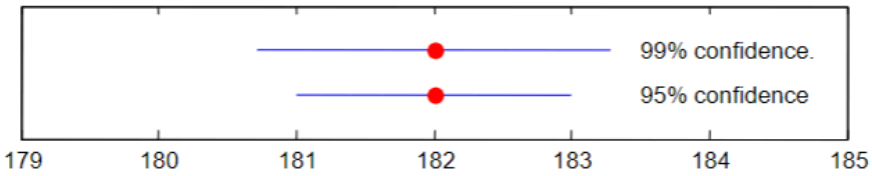
The interval is exact when the population distribution is normal and is approximately correct when n is large in other cases.

Properties of a confidence interval

The width of the interval is affected by the sample size n



The width of the interval is affected by the confidence level C



A confidence interval is usually affected by the following variables:

- Sample Size (n): A larger sample size results in a narrower confidence interval, as more data provides more precise estimates of the population parameter.
- Confidence Level (C): A higher confidence level means a wider interval because we are more certain that the interval will contain the true population parameter.
- Critical Value (z*): The critical value represents the number of standard deviations away from the mean we need to account for a given confidence level. A higher critical value results in a wider interval.

Choosing the sample size n

When determining the appropriate sample size for a study, we consider both the desired confidence level and margin of error. The sample size n can be calculated using the following formula:

$$n \geq \left(\frac{z^* \sigma}{m}\right)^2$$

Where:

- n is the sample size.
- z* is the critical value that corresponds to the desired confidence level (e.g., for a 95% confidence level, z* is approximately 1.96).
- m is the standard deviation (or estimate of standard deviation if unknown) of the population.

Applied Statistics 1 - IBEB - Lecture 5 - Week 5

Hypothesis testing

Concepts

Null hypothesis:
– Typically conservative
– Often a statement you aim to to disprove

Alternative hypothesis:
– Often the thing that you want to prove, backed by expected or observed evidence.
– The key question is whether the evidence is strong enough to establish statistical significance.

Hypotheses always relate to population parameters or a statistical model.

Example: Seeing whether profit in the banking sector changed with respect to previous years.

– *Null hypothesis:* $H_0: \mu = 0$

– *Alternative hypothesis*: $H_a: \mu \neq 0$

One-sided alternative: A parameter differs from its null value in a specific direction.
Example: $H_a: \mu > 0$

Two-sided alternative: A parameter differs from its null value in either direction.
Example: $H_a: \mu \neq 0$

Test-statistic

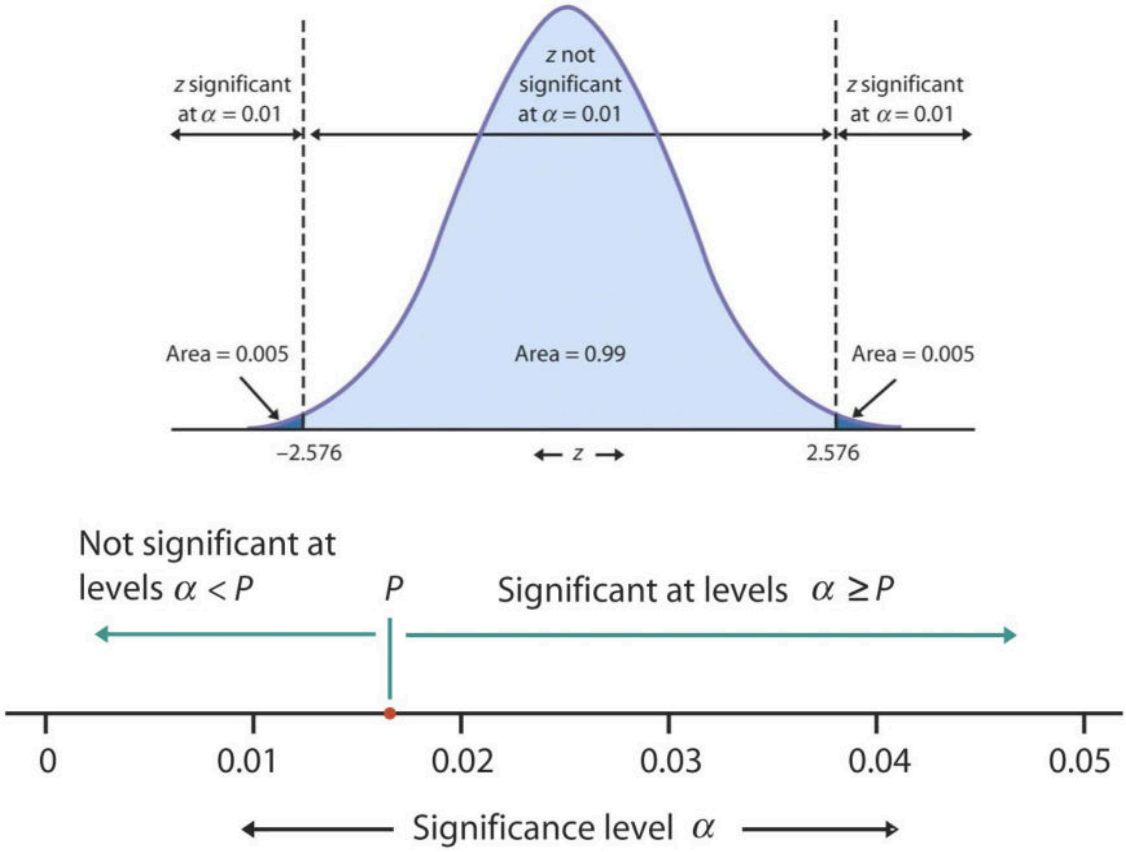
- To test a certain hypothesis, we need a test-statistic.
- A test statistic is a function of the sample that helps determine how likely the observed data is under the assumption that the null hypothesis (H_0) is true.
- Formula: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

P-value

- The probability, computed assuming that H_0 is true, that the test statistic would make a value extreme or more extreme than observed is called the P-value of the test.
- The smaller the P-value, the stronger evidence against H_0 provided by the data.

Significance level α

- The null hypothesis (H_0) is rejected if the P-value is smaller than a predefined significance level (α).
- The significance level α must be chosen before conducting the test.
- Common values for α include 0.10, 0.05, and 0.01, representing different levels of tolerance for Type I errors.



Hypothesis testing advantages & disadvantages

Advantage:

– Provides a clear decision: Reject H_0 or Do not reject H_0.

Disadvantage:
– Statistical significance does not always imply practical significance..
– A simple reject/do not reject outcome does not account for the strength of evidence against H_0.
– Repeated testing with new samples increases the likelihood of eventually finding statistically significant results, even if no true effect exists (increasing the risk of false positives).

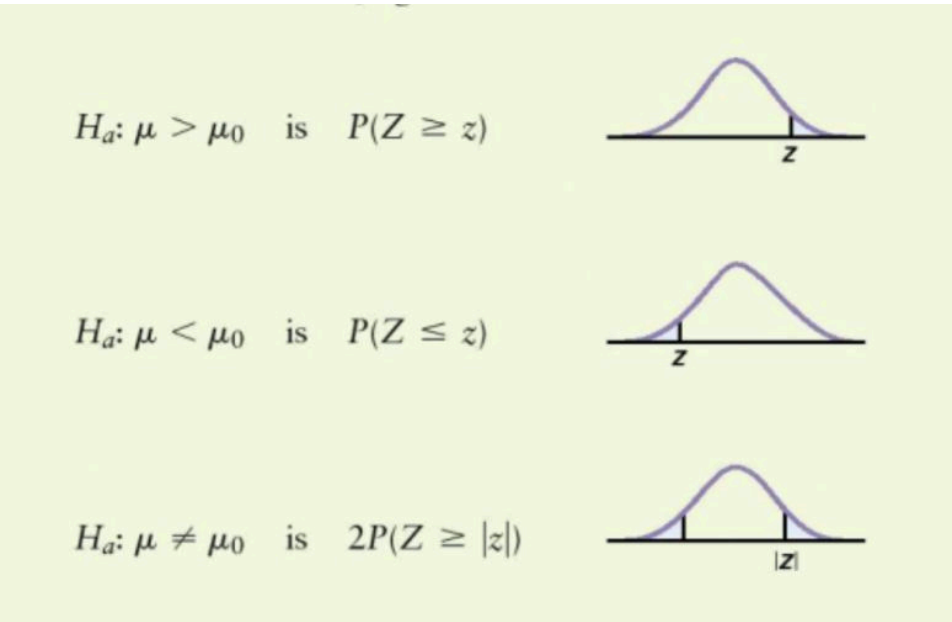
Summary

z test for a population mean

To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the one-sample z statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a variable Z having the standard Normal distribution the P-value for a test of H_0 against:



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Hypothesis testing procedure

1. Formulate hypotheses – Define the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Calculate a test statistic – Compute a value based on the sample data.
3. Evaluate the test statistic – Compare it to a threshold determined by the chosen significance level (α) to assess certainty.
4. Draw a conclusion – Based on the comparison, either reject H_0 or fail to reject H_0, interpreting the results accordingly.

Example

A trash bag producer claims that he invented a new and stronger trash bag. The old bags of the producer have a breaking point of 50 pounds.
We want to test the claim that the new bag is better. For this purpose, a sample of 40 new bags are tested.
The mean breaking weight of these 40 bags is 50.575.
The standard deviation of the breaking weight is known to be 1.65.
Perform the test using a significance level α = 5%.

Solution:

1. Formulate hypotheses (start with alternative):

$$H_0: \mu = 50$$
$$H_a: \mu > 50$$

2. Calculate test statistic: $z = \frac{\overline{X} - \mu_0}{(\frac{\sigma}{\sqrt{n}})} = \frac{50.575 - 50}{(\frac{1.65}{\sqrt{40}})} = 2.20$
3. Calculate p-value: P(Z>z)=P(Z>2.20)=0.0139
4. Give conclusion in terms of original question:
   At a 5% confidence level we reject the null hypothesis. The new bags are better.

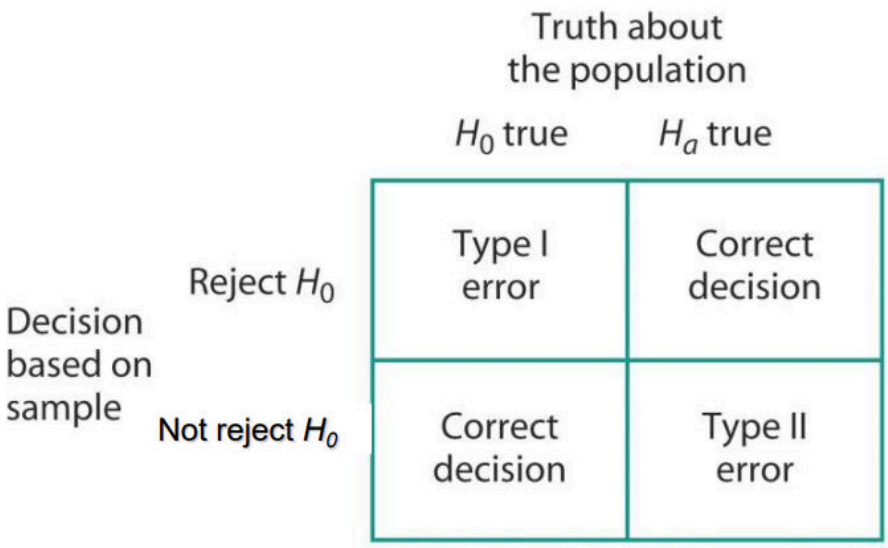
Power of a test

- A crucial aspect of hypothesis testing is a test's ability to reject the null hypothesis (H_0) when it is false. This ability is known as the power of the test.
- The power of a test is the probability that a significance test (at a fixed α level) correctly rejects H_0 when a specific alternative hypothesis is true.
- The power of a test depends on the true effect size and other factors, such as sample size and significance level.

Type 1 and type 2 errors

- Type I error: occurs when we reject H_0 when it is actually true (false positive)
  α: $P(\overline{X}$ is in the rejection region | H_0 is true)

- Type II error: occurs when we fail to reject H_0 when it is actually false (false negative).
  β: P(X is not in the rejection region| H_a is true)
- Power of a test is the complement of the Type II error β.
  Power=1 – β



Confidence intervals and hypothesis testing

A two-sided significance test of level α rejects a hypothesis H_0: $\mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ.

Power of a test continued

To calculate the power, we need three things:

1. The significance level α
2. The rejection region of the test
   => Reject when p-value < α.
   => Reject if $|z| > z^*_{\alpha/2}$ (two-sided test) or $z > z^*_{\alpha}$

- The test statistic is:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

For a one-sided test with rejection region $z > z^*_{\alpha'}$ we get:

$$\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z^* \rightarrow \bar{X} > \mu_0 + z^* . \frac{\sigma}{\sqrt{n}}$$

3. A specific value in the alternative hypothesis for which we calculate the power: If we have a specific true value μ_a that corresponds with the alternative hypothesis, we can calculate the probability of correctly rejecting null value given μ_a:

   Power = P(reject | μ_a true) = P($\bar{X}$ in rejection region | $\mu = \mu_a$)

   Example:

$$P(\bar{X} > \mu_0 + z^* . \frac{\sigma}{\sqrt{n}} | \mu = \mu_a)$$

Inference for means

If the variance is unknown, we use the t distribution: $Z = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

If σ is unknown, we replace it by its estimator s.

With $n - 1$ degrees of freedom (how spread the distribution is compared to normal distribution).

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

which gives:

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

Applied Statistics 1 - IBEB - Lecture 6 - Week 6

One-sample t test

- A one sample t test is used when there is an unknown population mean

Test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

C confidence interval: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$

Margin of error: $t^* \frac{s}{\sqrt{n}}$

<!> In a one-sample t-test, to approximate p-values, we locate the critical value closest to the observed p-value and match it with the corresponding degrees of freedom (i.e., the nearest value in the appropriate row of Table D).

Non-normality

The t-statistic is valid only if the population follows a normal distribution. However, the test can still be used under the following conditions:

- n is large enough ($n > 100$)
- n is not too small ($20 < n < 100$), but has no extreme skewness or outliers
- n is small ($n < 20$), but the population is approximately normally distributed.

Comparison of two groups

Paired sample t-test

A paired sample t-test is used when the same individuals or related subjects are measured twice under different conditions. The goal is to determine whether there is a significant difference between the two paired measurements.

Procedure:

1. Calculate the difference between the ratings for each individual in the panel.
2. Construct a confidence interval (CI) for the difference, or perform a hypothesis test, to see whether the ratings differ significantly.

Test statistic:

$$t = \frac{(\bar{D} - \mu_D)}{S_D/\sqrt{n}} \sim t_{n-1}$$

<!> The variance is usually not equal to the sum of the two variances as the two samples are not independent.

Sample variance of the difference:

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

Sign test

Sign test: A test on the median.

- Insensitive to outliers
- Uses no distributional assumptions.

For Matched pairs: Ignore pairs with difference 0; the number of trials n is the count of the remaining pairs. The test statistic is the count X of pairs with a positive difference. P-values for X are based on the Binomial B(n, 1/2) distribution.

Normal approximation for binomial distribution

When exact probabilities are not available in standard statistical tables, the normal approximation to the binomial distribution can be used to estimate probabilities more efficiently:
Suppose that a count X has the Binomial distribution with n trials and success probability p. When n is large, the distribution of X is approximately Normal, N(np, $\sqrt{np(1-0)}$).
As a rule of thumb, we will use the Normal approximation when n and p satisfy np≥ 10 and n(1 - p) ≥ 10.

Summary of some important testing results

If the population standard deviation is unknown, it should be replaced by the sample standard deviation (s), and the z-statistic is replaced by the t-statistic.

Testing for a difference in means:

- Paired samples:
    - Normality: Differences are normally distributed, use t-test.

- Non-normal: Use a sign test. Consider sign of differences. If there is no difference, the number of plusses follows binomial distribution with p=0.5.

Comparison of two groups: Independent samples

When comparing two independent groups, the goal is to determine whether there is a significant difference between their means.
Procedure:

1. Calculate the mean ratings for the two groups
2. Construct a confidence interval (CI) for the difference, or perform a test, to see whether the ratings differ significantly.
3. Construct a test statistic using the separate sample statistics of the two samples:

Suppose that $\bar{x}_1$ is the mean of an SRS of size n_1 drawn from an N(μ_1, σ_1) population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an N(μ_2, σ_2) population. Then the two-sample z statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Has the standard normal N(0,1) sampling distribution.

If n_1 and n_2 are sufficiently large, we can use:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1)$$

With a small sample and both populations normally distributed:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t(df)$$

The two-sample t confidence interval:

Draw an SRS of size n from a Normal population with unknown mean μ_1 and an independent SRS of size n_2 from another Normal population with unknown mean μ_2. The confidence interval for $\mu_1 - \mu_2$ given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least C no matter what the population standard deviations may be. The margin of error is

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t* is the value for the t(k) density curve with area C between -t* and t*. The value of the degrees of freedom k is approximated by software or we use the smaller of n_1 -1 and n_2 -1.

The two-sample t significance test:

Draw an SRS of size n_1 from a normal population with unknown mean μ_1 and an independent SRS of size n_2 from another normal population with unknown mean μ_2. To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values for the t(k) distribution where the degrees of freedom k are either approximated by software or are the smaller of n_1 -1 and n_2 -1.

Distribution of sum of normal variables

Suppose we have to random variables $\bar{X}$ and $\bar{Y}$ with:

$$E(\bar{X}) = \mu_x, V(\bar{X}) = \sigma_X^2 = \frac{\sigma_X^2}{n_x} \text{ and } E(Y) = \mu_Y, V(\bar{Y}) = \frac{\sigma_Y^2}{n_Y}$$

Then:

$$E(\bar{X} - \bar{Y}) = \mu_x - \mu_y$$

If $\bar{X}$ *and* $\bar{Y}$ are independent:

$$V(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

If $\bar{X}$~N(μ_x, $\sqrt{\frac{\sigma_X^2}{n_x}}$) and If $\bar{Y}$~N(μ_y, $\sqrt{\frac{\sigma_y^2}{n_y}}$): $\bar{X} - \bar{Y}$ ~N($\mu_x - \mu_y$, $\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_y^2}{n_y}}$)

T-test with pooled variance

Sometimes it is reasonable to assume that both populations have the same variance, that means: $\sigma 1 = \sigma 2 = \sigma$
Then:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$$

Instead of separately estimating $\sigma 1$ and $\sigma 2$, we can use one estimator based on both samples: The pooled estimate S_p^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Procedure

Draw an SRS of size n_1 from a normal population with unknown mean μ_1 and an independent SRS of size n_2 from another normal population with unknown mean μ_2. Suppose that the two populations have the same unknown standard deviation. A level C confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here t^* is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the pooled two-sample t statistic:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

And use P-values from the $t(n_1 + n_2 - 2)$ distribution.

Testing equality of variances - How to know whether it is pooled or not?

The decision to use a pooled variance depends on whether the variances of the two groups are equal. If the variances are approximately equal, pooling provides a more precise estimate of variance.
Before doing a t-test to test a difference in means, we first test whether variances differ:

$H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$ Pool if H_0 cannot be rejected.

When S_1^2 and S_2^2 are simple variances from independent SRSs of sizes n_1 and n_2 drawn from Normal populations, the F statistic:

$$F = \frac{S_1^2}{S_2^2}$$

Has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when H_0:$\sigma_1 = \sigma_2$ is true.

- If F is significantly greater than 1, it suggests that Group 1 has a larger variance than Group 2.
- If F is significantly smaller than 1, it suggests the opposite.
- If the true variances are equal: the two sample standard deviations tend to be similar and F will be close to one => deviations from 1 (in both directions) providing evidence for the alternative hypothesis.
- Table E gives right tail critical values for the F-distribution. This is enough to also do a two-sided test.
- To find the appropriate critical values be careful in assessing the degrees of freedom associated with the numerator and denominator.

Note:

1. Normality is crucial for this test
2. The F-statistic is always positive since variances cannot be negative.

Summary of some important testing results

If σ unknown, replace it by the sample statistic s. The z-statistic becomes a t statistic.
Testing for a difference in means:

Paired samples:

- Normality: Differences are normally distributed, use t-test.
- Non-normal: Use a sign test. Consider sign of differences. If no difference, number of plusses follows binomial distribution with p=0.5

Independent samples:

- First use the F-test to see if we can assume equal variances. Depending on the result of that test we choose or test:
  If we cannot reject the null hypothesis of equal variances:
- Equal variances: t-test with pooled variance

If we can reject the null hypothesis of equal variances:

- Different variances: Independent samples t-test

Applied Statistics 1 - IBEB - Lecture 7 - Week 7

Proportions

Choose an SRS of size n from a large population that contains population proportion p of "successes." Let $\hat{p}$ be the sample proportion of successes,

$$\hat{p} = \frac{count\ of\ successes\ in\ the\ sample}{n} = \frac{X}{n}$$

Then:

- As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.
- The mean of the sampling distribution is p.
- The standard deviation of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}$$

Confidence interval for proportions

- To make a confidence interval we need to know the variance.

- This depends on the unknown parameter p and the sample size n.
- As p is unknown, we approximate/estimate it:

$$\hat{p} = \overline{X} = \frac{X}{n}$$

- To estimate the variance, we can use the following estimate:

$$\hat{\sigma}_p^2 = \frac{\hat{p}(1-\hat{p})}{n}$$

The confidence interval can therefore be obtained using:

$$\hat{p} \pm z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Obtaining an interval based on a specified width:

$$M = z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\rightarrow \sqrt{n} \geq \frac{z^* \sqrt{\hat{p}(1-\hat{p})}}{M}$$

$$\rightarrow n \geq \frac{z^{*2} \hat{p}(1-\hat{p})}{M^2}$$

Typically, we do not know the proportion. So, how can we find this value?
– Use previous research
– Use worst case scenario

"Worst case scenario":
Choose n in such a way that the interval will always have the required maximum for all possible values for p :

- We need to maximize $p - p^2$
- The maximum is attained when p=0.5
- Therefore, the sample size can be chosen by using this 'worst case scenario':

$$n \geq \frac{z^{*2} \hat{p}(1-\hat{p})}{M^2} = \frac{z^{*2} 0.5(1-0.5)}{M^2}$$

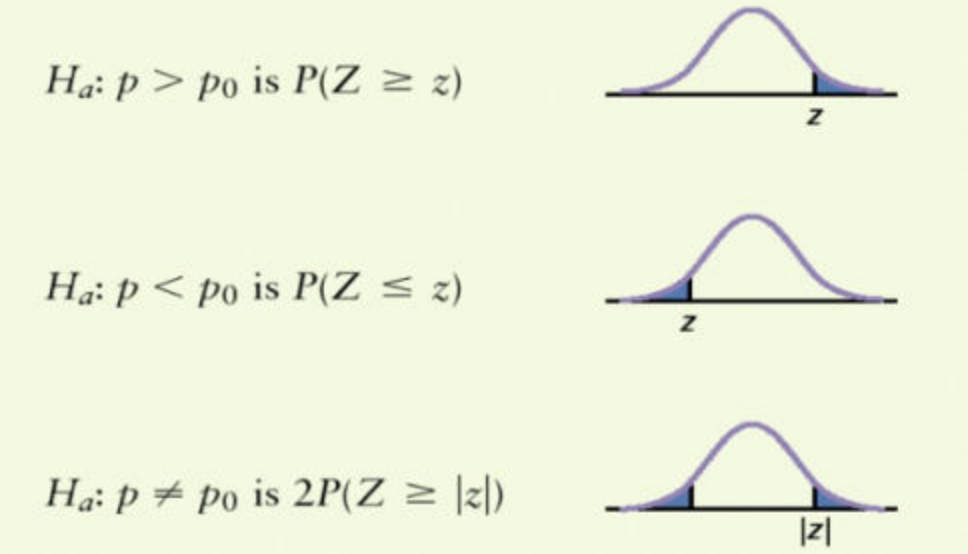
Hypothesis testing

Large-sample test

Choose an SRS of size n from a large population with unknown proportion p of successes. To test the hypothesis $H_0: p = p_0$, compute the z statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

In terms of a standard normal random variable Z, the approximate P-value for a test of H_0 against



Use this test when the expected number of successes np_0 and the expected number of failures $n(1-p_0)$ are both greater than 10.

Requirements for the proposed test and interval:
– A large sample: np and n(1-p) > 10.
– A large population: This is to ensure that the observations are independent.

Small-sample test

– For a small sample, with a large population, we can consider the binomial distribution.
– The number of successes follows a Binomial distribution Bin(n,p).

Difference in proportions

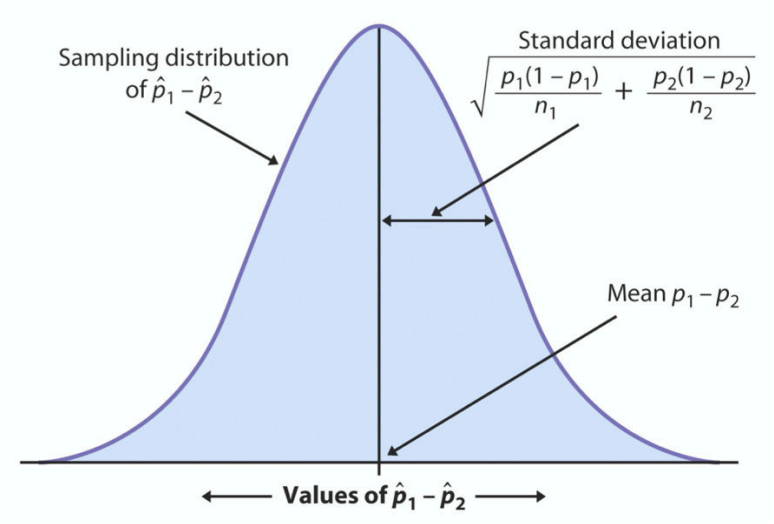
If the two are independent, it follows that the difference between the two proportions is also approximately Normally distributed.

Sampling distribution of $\hat{p}_1 - \hat{p}_2$

Choose independent SRSs of sizes n_1 and n_2 from two populations with proportions p_1 and p_2 of successes. Let D = $\hat{p}_1 - \hat{p}_2$ be the difference between the two sample proportions of successes. Then:

- As both sample sizes increase, the sampling distribution of D becomes approximately Normal.
- The mean of the sampling distribution is $p_1 - p_2$.
- The standard deviation of the sampling distribution is

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$



confidence interval for the difference between two proportions:

$$\hat{p}_1 - \hat{p}_2 \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Note: Use when the number of successes and the number of failures in each of the samples are at least 10.

Significance test for comparing two proportions

Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes. To test the hypothesis

$$H_0: p_1 = p_2$$

compute the z statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{Dp}}$$

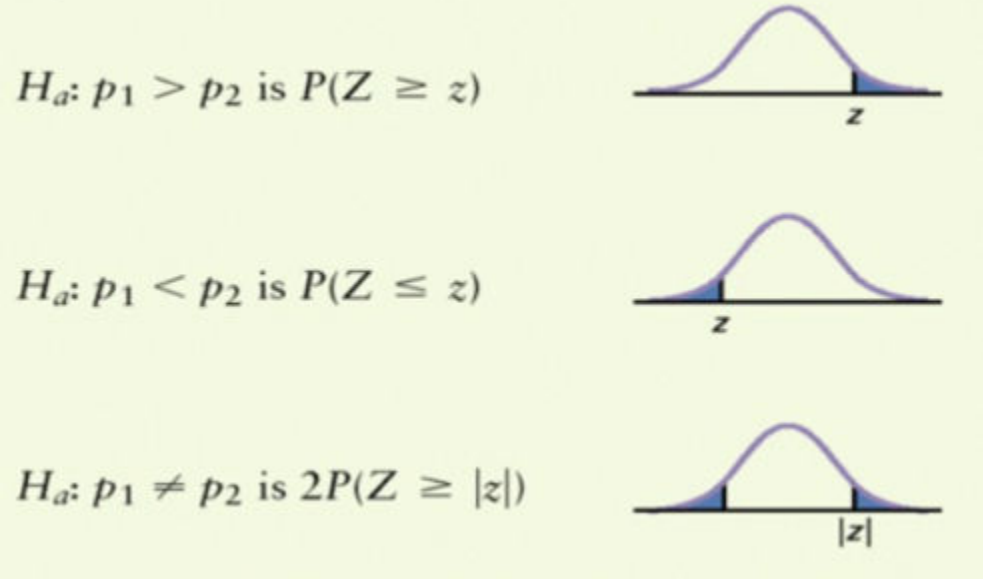
where the pooled standard error is

$$SE_{Dp} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

based on the pooled estimate of the common proportion of successes,

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In terms of a standard Normal random variable Z, the P-value for a test of H_0 against



Use this test when the number of successes and the number of failures in each of the samples is at least 5.

Reference list

Van de Velden, M. (2025). Applied Statistics 1 Lecture 1.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/99961285

Van de Velden, M. (2025). Applied Statistics 1 Lecture 1.2 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/99983220

Van de Velden, M. (2025). Applied Statistics 1 Lecture 2.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100033855

Van de Velden, M. (2025). Applied Statistics 1 Lecture 2.2 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100050816

Van de Velden, M. (2025). Applied Statistics 1 Lecture 3.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100083606

Van de Velden, M. (2025). Applied Statistics 1 Lecture 3.2 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100113880

Van de Velden, M. (2025). Applied Statistics 1 Lecture 4.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100153934

Van de Velden, M. (2025). Applied Statistics 1 Lecture 4.2 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100173922

Van de Velden, M. (2025). Applied Statistics 1 Lecture 5.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100204365

Van de Velden, M. (2025). Applied Statistics 1 Lecture 6.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100265395

Van de Velden, M. (2025). Applied Statistics 1 Lecture 7.1 [Lecture Slides]. Retrieved from:https://canvas.eur.nl/courses/47650/files/100316094