EFR summary

Applied Statistics 1, FEB11005X

2023-2024

Lectures 1 to 7

Weeks 1 to 7

Details

Subject: Applied Statistics 1 IBEB 2023-2024

Teacher: Michel van de Velden

Date of publication: 23.02.2024

Applied Statistics 1 - IBEB - Lecture 1 - Week 1

Appling Statistics

Introduction

Statistics is the art and science of learning from data. We can conduct statistical analysis to produce a useful summary of data. Statistics helps making decisions under uncertainty, but it does not remove uncertainty.

Data

Data consists of:
- objects/individuals → found in rows
- variables → found in columns
- outcomes → found in cells

Variables

Two types of variables:
1. Categorical variables: places an individual in one of several groups (e.g. job types, gender, *etc)*
2. Quantitative variables: takes on numerical values for which typical arithmetic operations make sense, or in other words, contains numbers that can be calculated with.
- Interval data: Difference can be meaningfully interpreted but relative numbers not (e.g. 10°C is 5° warmer than 5°C, but not double the heat)
- Ratio data: As interval but now also relative numbers can be interpreted (e.g. €10 is €5 larger AND double as much as €5)

Graphs

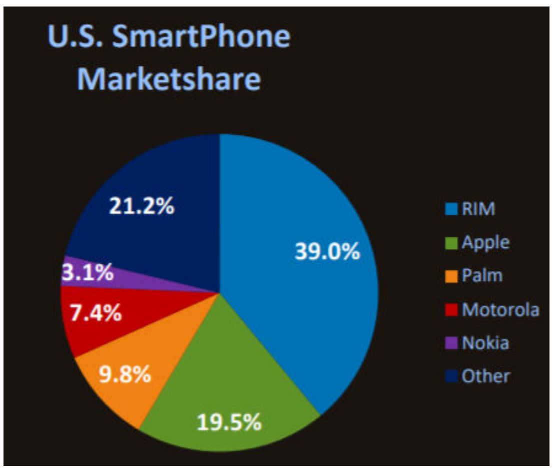
- Graphs are tools to summarize and represent data through visualization.

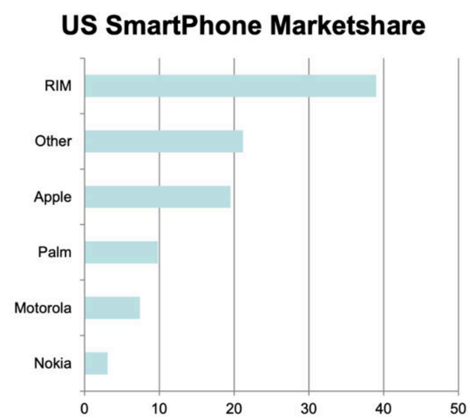
●	To choose which type of graph is used, it is important to consider the data's Categorical vs. Quantitative aspects.

Graphing the distribution of categorical variables

Graphs used are (for example):
- Pie Chart
- Bar Chart

In which can depict relative sizes





Pie chart **Bar chart**

Source: Lecture 1.1 Applied Statistics 1, slides 21 and 23 (van de Velden, 2022)

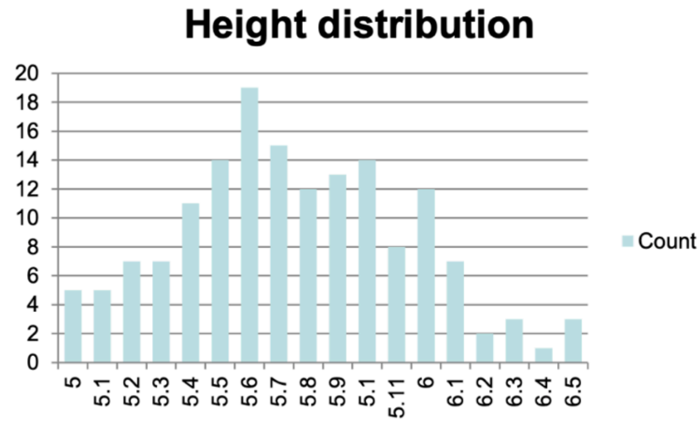
Graphing the distribution of quantitative variables

To depict relative sizes, the graphs used are:
- Histogram
- Bar Chart

Histogram

- Visual summary of the distribution of values
- Displays the distribution of a quantitative variable:
  - Horizontal axis: classes of the quantitative variables
  - Vertical axis: (relative) frequencies of the classes
- It always consists of numerical variables

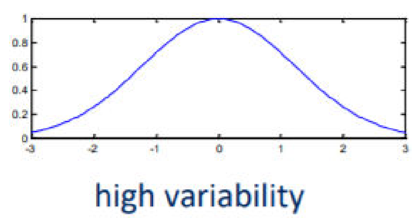
Example: *Height distribution*

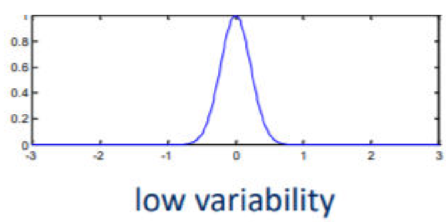


Source: Lecture 1.1 Applied Statistics 1, slide 28 (van de Velden, 2020)

What we focus on:

- Central tendency: 'middle' or midpoint of the observed values
- Spread:
  - Distribution of data around the 'middle'
  - What range of values do the observations tend to fall
  - The variability (high or low)
  -

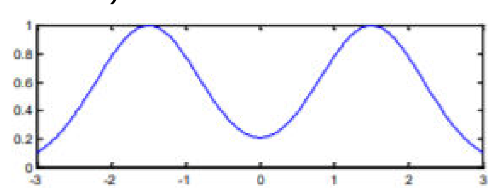




Source: Lecture 1.1 Applied Statistics 1, slide 33 (van de Velden, 2022)

- Shape: The striking patterns in distributions

Example: Two modes (bimodal)

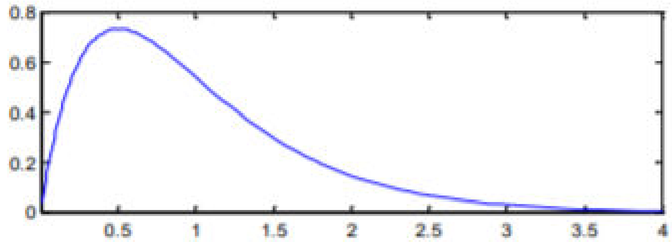


Source: Lecture 1.1 Applied Statistics 1, slide 33 (van de Velden, 2022)

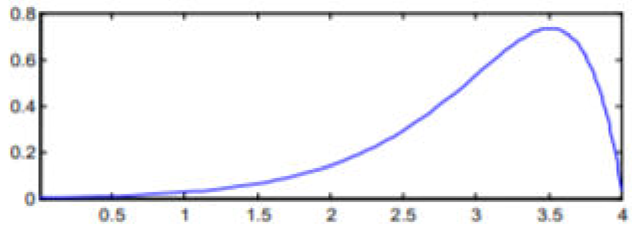
Symmetry vs Skewness in Histogram

A multitude of points in a histogram is:

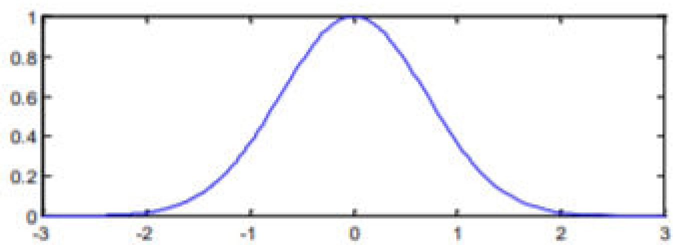
- Skewed to the right if most of the points are distributed to the right



- Skewed to the left if most of the points are distributed to the left



- Symmetrical if the points are distributed in a symmetrical way (symmetrical distribution)



Source: Lecture 1.1 Applied Statistics 1, slide 34 (van de Velden, 2022)

Stemplot

Stemplots are a fast and detailed way to graph small sets of data. Steps to make a stemplot:

- Separate each observation into a stem consisting of all but the last digit and a leaf, which is the final digit
- Write the stems in a vertical column with the smallest stem standing at the top and draw a line to the right, parallel to the column of stems
- Place each leaf to the right of its stem, in increasing order out from the stem

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

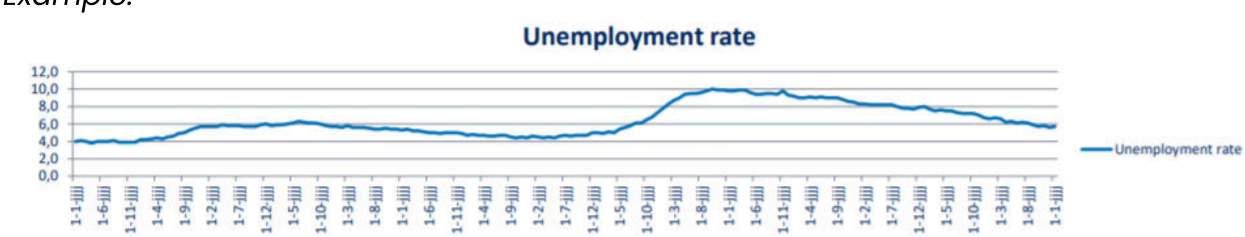
Source: Lecture 1.1 Applied Statistics 1, slide 36 (van de Velden, 2022)

Line graph: time plot

Time plot

- Time plot of a variable plots each observation against the time at which it was measured
  - o Horizontal scale of the plot: time of observation
  - o Vertical scale of the plot: variable measured or variable of interest

Example:



Source: Lecture 1.1 Applied Statistics 1, slide 40 (van de Velden, 2022)

Note:

- Be careful when analysing the data from the graphs due to the visualization of the graph maker that might make slight changes to the shape of the graph.

Statistical description of data

Central Tendency

We consider three measures of central tendency: mean, median, mode

a) Mean

$$x_{mean} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{sum}{size}$$

b) Median

To find the median:

- Rank order the observations from small to large
- For an odd number of observations: the median is the middle observation
- For an even number of observations: the median is the average of the two middle observations
- The median is the observation such that 50% of your data is smaller and 50% of your data is larger

<!> Mean and Median are equivalent if the distribution is symmetric. However, if the distribution is skewed (or if we have an outlier on one side) their positions become separated.

c) Mode

In a group of observations, the mode is the observation (class) that occurs most often.

d) Average

The average is often used to measure the central tendency. However, it is very sensitive to observations that differ from the rest: outliers.

The average is the middle value in the sense that the sum of all differences equals exactly zero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

This means that if we know (n - 1) differences, the last difference is also known. In other words: We only have (n - 1) independent terms.

Spead

In addition to measuring the central tendency, we also have to find the *distribution of values.*

- More spread: more uncertainty.
- Less spread: less uncertainty.

a) Range

Range is a common way of measuring the spread or variation of observations.

$range = value - value$

The spread, however, is not very informative if there are only a few extreme points.

b) Standard Deviation

If the *mean* is an appropriate measure for central tendency, the natural measure for spread is the standard deviation:

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

c) Variance

The variance is also used to measure the spread. It can be found by computing s^2. Thus:

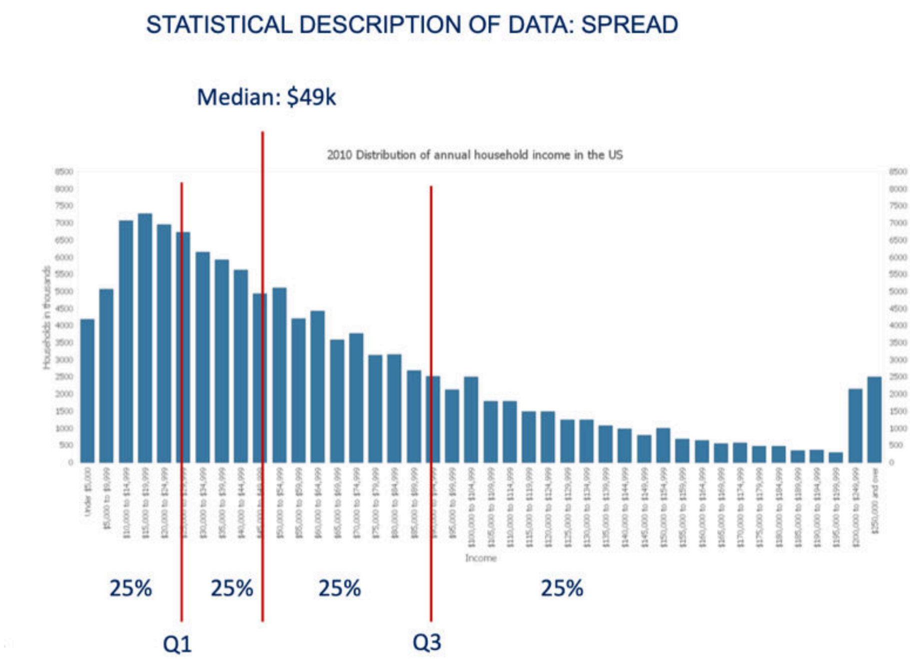
$$v = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

d) Interquartile range

If the *median* is a more accurate measure, then we should consider the percentiles of the distribution that lie around the median. Thus, we can find the interquartile range (IQR):

$IQR = 75\, percentile - 25\, percentile$

Example of considering the percentiles:



Source: Lecture 1.1 Applied Statistics 1, slide 57 (van de Velden, 2022)

A numerical summary of the data that provides information on central tendency and spread, consists of:

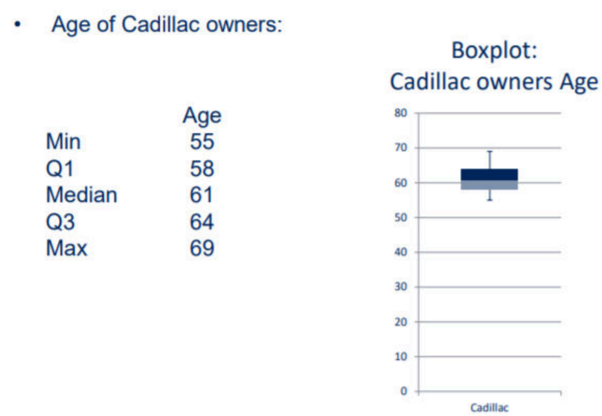
- Range
- Mean
- Standard deviation

And/or:

- Min
- Q1
- Median
- Q3
- Max

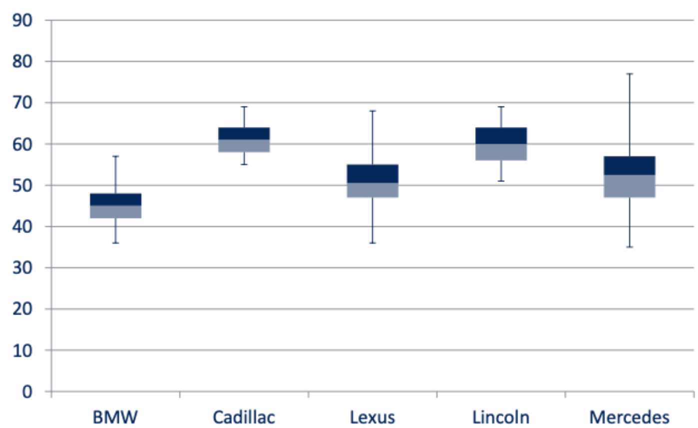
Boxplot

An alternative of having a list/table with the 5 numbers (min, Q1, Median, Q3, Max), a boxplot can be used.



Source: Lecture 1.1 Applied Statistics 1, slide 61 (van de Velden, 2022)

Advantage of using a boxplot: We can immediately compare distributions



Source: Lecture 1.1 Applied Statistics 1, slide 63 (van de Velden, 2022)

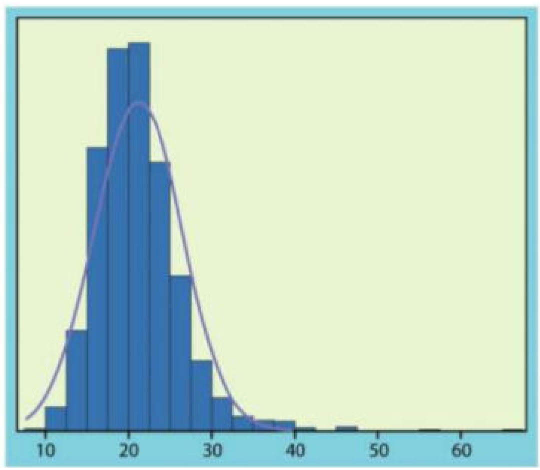
Applied Statistics 1 - IBEB - Lecture 1.2 - Week 1

Density Curves

Definition

The density curve is drawn based on the histogram which provides a mathematical approximation to the data.

A density curve is a curve that:
- Is always on or above the horizontal axis
- Has exactly area 1 underneath it
- Describes the idealized descriptions of the data (compact picture of the overall pattern of the data but ignore minor irregularities and outliers)
- The area that lies under the curve and above any range of values is the proportion of all observations that fall within that range.
- Mean, standard deviation, median, IQR apply to density curves:
- Median: Divides surface below the curve two equal parts.
- Mean: Point of gravity/balance



Source: Lecture 1.2 Applied Statistics 1, slide 5 (van de Velden, 2022)

Numerical Measures

- The mean and standard deviation on the density curve are only approximately equal to the *observed* average and standard deviation.
- Therefore, we use different notations – in general, Latin letters for observed characteristics, and Greek letters for idealized ones:

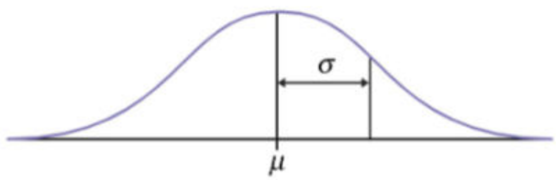
	Mean	Standard deviation
Observed distribution	$\bar{x}$	s
Idealized distribution	μ	σ

Source: Lecture 1.2 Applied Statistics 1, slide 10 (van de Velden, 2022)

Normal Distribution

A frequently used distribution f(x) is the normal distribution:

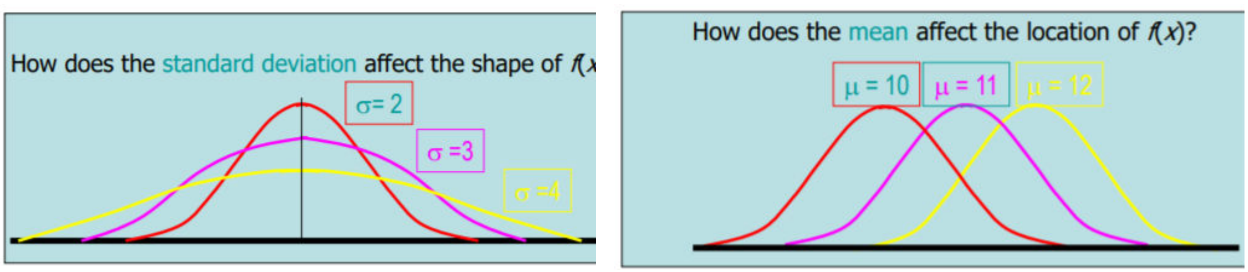
$$f(x) = \frac{1}{\sqrt{2}} e^{-\frac{(x-)^2}{2\ ^2}}$$



Properties:

Symmetric, Unimodality, Bell shape

- The curve is determined by and
- The surface below the curve is always 1. Regardless of the values for μ and σ



Source: Lecture 1.2 Applied Statistics 1, slide 15 (van de Velden, 2022)

Standardizing and Z-scores

If x is an observation from the distribution that has mean and standard deviation, then the standardized value of x (z-score) is:

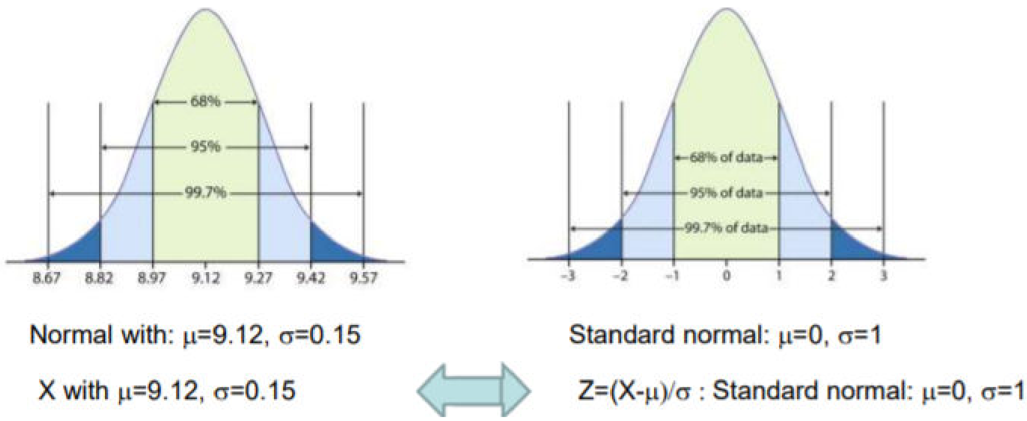
$$z = \frac{x-\mu}{\sigma}$$

Thus, a variable with a normal distribution N (μ, σ) becomes standard normal N(0,1) after standardization. This means that we can make calculations for any Normal distribution by using the standard Normal distribution.

The 68-95-99.7 rule

A normal distribution with mean and standard deviation has:

1. 68% of observations within σ of the mean μ
2. 95% of observations within 2σ of the mean μ
3. 99.7% of observations within 3σ of the mean μ



Source: Lecture 1.2 Applied Statistics 1, slide 17(van de Velden, 2022)

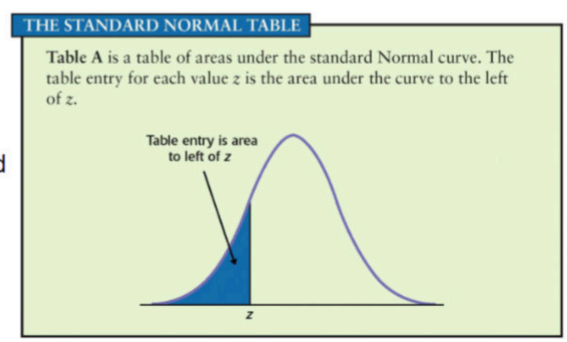
Finding normal distribution

1. State the problem in terms of observed variable x
2. Standardize x to restate the problem in terms of a standard normal variable z.

Draw a picture to show the area under the standard normal curve

3. Find the required area under the standard normal curve using Table A in your book and the fact that the total area under the curve is 1

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



Assessing the normality of data

To know whether a variable is normally distributed, we can draw a histogram. However, a histogram is not always conclusive and is troublesome for small data sets. Thus, we can instead draw a Normal quantile plot.

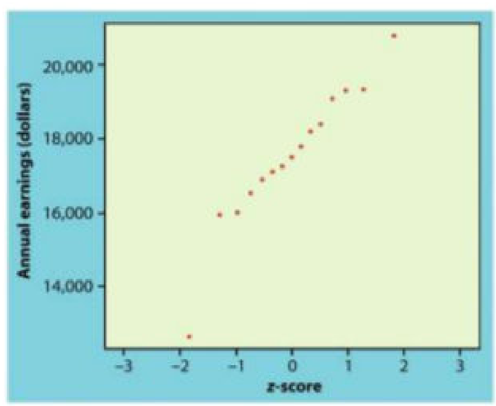
- Normal quantile plot: Display a *straight* line if the distribution is normal. Any deviations from the straight line are deviations from normality.

How to draw a Normal quantile plot?

To draw a Normal quantile plot, the following steps are followed:

1. Order the observations ascendingly
2. Find the percentile of each observation
3. Compute the z-scores from the standard normal distribution according to the percentiles
4. Make a scatterplot with the z-scores on the x-axis and the observed values on the y-axis.

Annual earnings	Observed percentile	z-score according to $N(0,1)$
12641	6.7	-1.50
15953	13.3	-1.11
16015	20.0	-0.84
16555	26.7	-0.62
16904	33.3	-0.43
17124	40.0	-0.25
17274	46.7	-0.08
17516	**53.3**	**0.08**
17813	60.0	0.25
18206	66.7	0.43
18405	73.3	0.62
19090	80.0	0.84
19312	86.7	1.11
19338	93.3	1.50
20788	100.0	inf



Source: Lecture 1.2 Applied Statistics 1, slide 41 (van de Velden, 2022)

Scatterplots

Definition

- The scatterplot represents the relationship between two quantitative variables. Each observation is depicted as one point with:

- The value of one variable is on x-axis
  - The value of the other variable is on y-axis
- We can use a scatter plot to determine the:
  - "shape" of the relationship (line, cluster etc.)
  - The direction of the relationship:

- Positive: high values of one variable correspond to high values of the other
- Negative: high values of one variable correspond to low values of the other
  - The strength of the relationship
    The closer the points are to a line, the stronger the relationship).
- In the scatter plot, look for:
  - Patterns and deviations from the pattern
  - Outliers
- Categorical variables can be added by applying different colors to points corresponding to different categories or by adding labels

Correlation

The correlation (r), measures the strength and direction of the linear relationship between two quantitative variables.

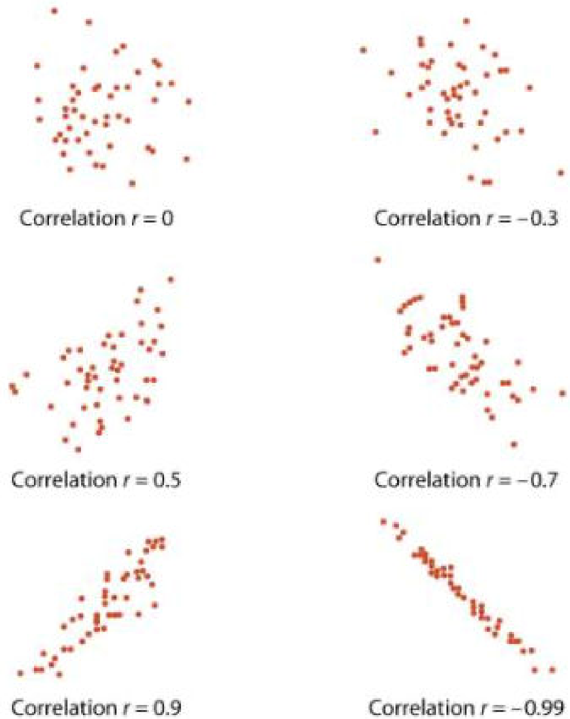
$$r = \frac{1}{n-1}\sum \frac{x_i - \overline{x}}{s_x} \times \frac{y_i - \overline{y}}{s_y}$$

- (x_i, y_i) with $i = \overline{1,n}$: values of the n individuals (for instance: (x_2, y_2) are values of the second individual)
- $\overline{x}$ and s_x are the mean and standard deviation for x- values
- $\overline{y}$ and s_y are the mean and standard deviation for y- values
- Positive relationship: high x and high y; low x and low y
- Negative relationship: high x and low y; low x and high y

In other words, the correlation is the mean product of the z-scores of x and y. Some properties include:
- Both variables have the same role
- Both variables must be quantitative
- r uses z-scores. Changes in measurement units (miles, km, cm, inches etc.) does not affect r
- r > 0 implies positive association, r < 0 implies negative association

- -1 ≤ r ≤ 1
- Only captures linear relationships
- Not robust against outliers

Correlation and scatter plot:



Source: Lecture 1.2 Applied Statistics 1, slide 54 (van de Velden, 2022)

Covariance

Correlation is a convenient measure of linear association

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

Least-squares regression

- Linear regression is used to summarize the linear relationships between two variables
- We can use regression to predict the value of one variable (y) for given values of the other value (x)
- The "best" line is one where the sum of the squared vertical distances from the points to the line is as small as possible

a) Equation of the least-squares regression line

The equation for the least-squares regression line is: $\hat{y} = a + bx$,

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

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

($\bar{x}$, $\bar{y}$: means; s_x, s_y: standard deviations)

Properties

1. The line always passes through $(\bar{x}, \bar{y})$
2. The distinction between the explanatory variable x and the response variable y is crucial
3. Interpretation slope $b = r\frac{s_y}{s_x}$ .If x changes by one stand. deviation, y changes with r stand. deviations
4. r^2 is used to give the proportion of variance in y that is explained by the variance in $\hat{y}$: $r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{variance\ in\ \hat{y}}{variance\ in\ y}$

Interpretation

– The slope: The slope gives the rate of change. The amount of change in ŷ when x increases by one unit.
– Prediction: the regression line makes it possible to predict values for y, on the basis of x values.

Applied Statistics 1 - IBEB - Lecture 2 - Week 2

Least-squares regression: SPSS procedure

- To make a scatterplot: Graphs>Scatter/dot
- Correlation: Correlate>Bivariate
- Regression: Analyze>Regression

  - Regression output:
  - Coefficients:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3,544	3,500		-1,013	,314
	AlcPerc	3,032	,722	,418	4,198	,000

a. Dependent Variable: Carbos

b. Selecting only cases for which percenta>0.5 (FILTER) = Selected

  - Constant: -3.544.
  - Slope 3.032.

Least-squares regression: Residuals and Outliers

Residual: The difference between an observed value of the response variable and the value predicted by the regression line. That is:

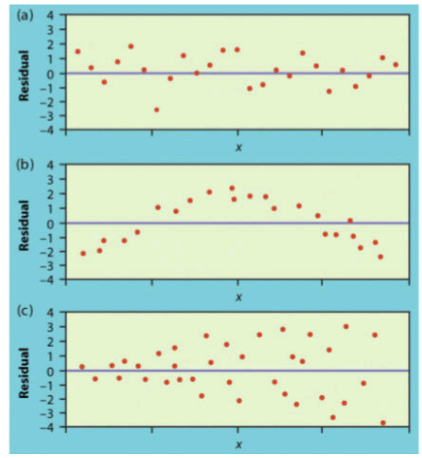
Residual = observed y - predicted y

Residual plot: The scatter plot of the regression residuals against the explanatory variable. It helps us assess the fit of a regression line.

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 are influential if their inclusion has a great influence on the determination of the line
- Especially outliers with respect to the x variable can easily become influential, even when they are not extreme outliers with respect to y

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 represents the use of a regression line for predicting values outside the range of values of the explanatory variable x. They are often not accurate.

Lurking variables

Lurking variables are variables that cannot be found among the explanatory or response variables but that may still influence the relationship/interpretation of the variables.

Association ≠ Causation

- Association is not causation: A high correlation does not mean that one variable “causes” the other to be high as well
- To determine causal relationships, we need experiments. Often this is impossible
- Sometimes we are able to establish causation without experiments. We need:
    - Strong association
    - Consistent association
    - Higher x-values have bigger effects
    - Cause precedes effect in time
    - Cause is plausible

Relations in categorical data

There is no relation between 2 variables if the conditional distributions are the same as the marginal distribution for either variable.

Simpson’s paradox

An association or comparison that holds for all of multiple groups can actually reverse direction when the data are combined to form a single group. This “reversal” is called Simpson’s paradox.

Producing data

Observation vs experiment

Observational study:

- A study in which individuals' variables of interest are measured but not influenced is called an.
- Observes individuals and measures variables of interest but does not attempt to influence the responses.

In contrast, an experiment deliberately imposes some treatment on individuals to observe their responses.

Confounding

We say that two variables, either explanatory or lurking variables, are confounded when we cannot distinguish the effects of each on a response variable.

Designing samples

Types of samples:

- **Simple random sample (SRS):** individuals are drawn at random, each individual has the same change of being selected
- **Voluntary response sample:** respondents choose to provide the data
- **Probability sample:** each respondent has a prior determined, probability of being selected
- **Stratified random sample:** The population is divided into groups of similar individuals; strata. In each stratum, a SRS is drawn and these are then combined to form the full sample

Bias

Definition

The design of study is biased if it systematically favours certain outcomes
There are three types of bias:

- Selection bias
- Information (misclassification) bias
- Confounding bias

Selection bias

Sample does not give a good representation of the population:
- Selection effects: we only observe a non-random part of the data.
- Self-selection bias: Publicity bias. People volunteer information => Leads to over- and under coverage of groups (E.x: people without internet knowledge won't participate in internet surveys)
- Nonresponse: Nonresponse of the randomly selected individuals, not everybody provides the required data.
- Texas sharp shooter bias: We formulate theories/hypotheses based on observed interesting/extreme patterns
- Confirmation bias: We see what we want to see. In particular, we see support for our theory/hypothesis and ignore other evidence

Example: conspiracy theories, football coaches/players/analysts counting chances

Information (misclassification) bias

Method of gathering the data is inappropriate and yields systematic errors in measurement:
- Response bias: respondents do not give the honest answer because of respondents' behavior, or the formulation of questions, or the presentation of a question
- Recall bias: Those exposed have a greater sensitivity for recalling exposure

Confounding bias

An observed effect is caused or influenced by one of the non-observed factors (like a lurking variable).

For example:
- Coffee drinking and cancer □ in which its confounder could be smoking
- Salary and gender □ in which its confounder could be level of education

Designing experiments

Definition

- Subjects: individuals studied in an experiment, especially if they are people
- Factors: explanatory variables
- Treatment is any specific experimental condition applied to the subjects. If an experiment has several factors, a treatment is a combination of specific value (often called a level) of each factor

In an experiment, we have at least one response variable and at least one factor that determines the treatments.

Comparative experiments

Post-test only one group:

Subject → Treatment → Response

In this case, only one group is tested:

- Problem: Placebo effect. (People tend to find results even if they actually did not receive treatment)
- In the post-test, placebo effect results are confounded with the *real effect*

Characteristics of the placebo effect:

- Quantity matters
- The 'ritual' or the means of the experiment matter
- The more expensive a placebo is the more effective it is
- Colors are important
- Placebo effects can have side effects

Overcoming the placebo effect

1. Introducing a control group that enters in the experiment but does not receive a treatment. This group is called the control group

2. However, in control group method, subjects/ experimenters know in which group they are, which influences the result

→ Solution: double blind set up: neither the experimenter nor the subject knows which treatment is received. In this case, unconscious bias is avoided

Basic principles for designing experiments

The basic principles that are taken into account when designing an experiment are:

- 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 all the above-mentioned principles are respected, it is possible that the effect of the treatment is much higher than expected (we say that there is a statistically significant effect), thus the experiments do not fully replicate the real-life situation.

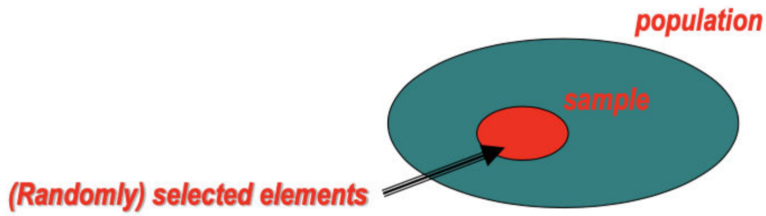
Assignment in experiments

There are three types of assignment in experiment:

- Completely randomized design
- Matched pairs designs: apply the treatment to pairs of (similar) subjects
    - Example: testing car tires → two cars run laps and measure the wear; or we use the same car twice with the same driver. Variation due to the different car and/or driver is then accounted for.
- Block design: Before the experiment, subjects are divided into groups; blocks with similar subjects. Within the blocks random assignment is carried out.

Population & samples

- The population is the entire group of individuals from which we want information.
- A sample is a part of population from which we collect information and draw conclusion about the whole.



Source: Lecture 2.2 Applied Statistics 1, slide 2 (van de Velden, 2022)

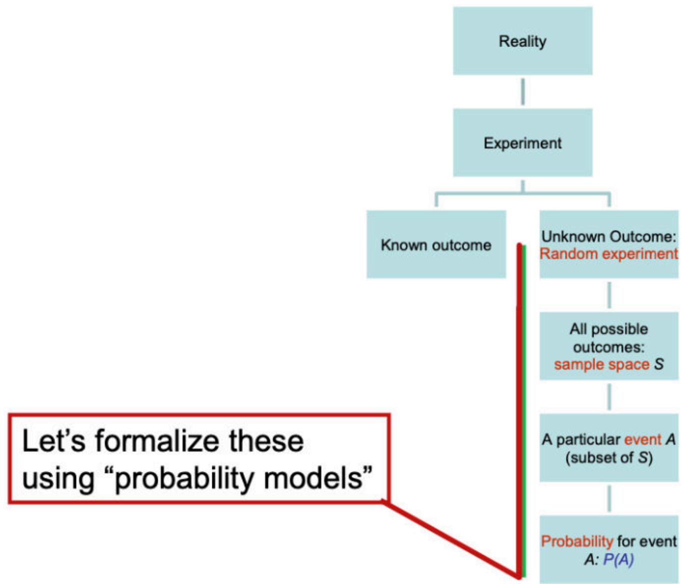
- Statistic inference: using the sample statistics (mean, standard deviation) to make statements about the population.

Applied Statistics 1 - IBEB - Lecture 3.1 - Week 3

Randomness

- Event is called random if individual outcomes are uncertain, but in the long run, a pattern can be observed in the outcomes
- The probability for a certain outcome in a random experiment is the proportion of times that this outcome occurs if we were to repeat the experiment infinitely.

Schematic way of looking at probability



Probability rules

1. $0 \leq P(A) \leq 1$ for any event A
2. $P(S) = 1$, with S being the sample space
3. Complement rules: $P(A \; does \; not \; occur) = 1 - P(A)$
4. Additional rule: Two events A and B are disjoint if they have no outcomes in common and so can never occur simultaneously: $P(A \; or \; B) = P(A) + P(B)$

Random variables

Random variable: variable whose value is a numerical outcome of a random phenomenon.
Two types:
- Discrete random variable: The outcomes are finite (countable)
- Continuous random variable: Infinite outcomes
- A probability distribution of a random variable X assigns probabilities to all values that X can take on.

Probability models

Newcomb-Benford's Law: "that the frequencies with which the leading digits of numbers occur in a large variety of data are far away from being uniform." (Formann, Anton K)

- It is an example that concerns a specific discrete distribution

Probability distributions

Probability distributions map probabilities to outcomes of random variable. Two types of random variables:

    - Discrete
    - Continuous

- Drawing the probability distribution gives a density curve
- Probability distribution is associated with mean (μ) and standard deviation (σ)

However, for discrete distributions, these can be defined as:

    - The weighted average (μ)
    - The root of the weighted average squared deviation from the mean (σ)

Continuous random variable

- The distribution is characterized by a density curve
- The probabilities are surfaces below the curve
- Probability for an event $X = a$ is always equal to 0: $P(X = a) = 0$
- The only meaningful events for a continuous random variable are intervals
    - Comparable situation: a line segment has a positive length, while no single point on the line segment does

Mean and variance of a discrete random variable

Mean

- The mean or expected value of a random variable is 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 m, 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:

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

*The standard deviation is the square root of the variance

Linear combinations

Let

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

(Note: a and b are known constants)

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

- To determine variance, we need to consider dependencies between the random variables
- Independence: If 2 variables are independent, the correlation coefficient $\rho = 0$
- Variance of the sum of 2 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	$\sigma^2_{X-Y} = \sigma^2_X - \sigma^2_Y$	$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$

Source: Lecture 2.2 Applied Statistics 1, slide 24 (van de Velden, 2022)

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 else 0	$(n\ k\,)p^k(1 - p)^{n-k}$ 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

- All events are 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.*

Binomial

- n independent repetitions of a Bernoulli experiment.
- The probability p for success is the same in each experiment.
- The random variable X is defined as: “the number of successes k out of n trials (repetitions of the experiment)”
- The sum of Binomial random variables is also Binomial distributed:
    - 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

Example: An exam consists of 10 multiple-choice questions. Each with 5 options. You pass with at least 6 correct answers. What is the probability of passing?

Poisson

- Determine the probability for the number of occurrences (successes) during a fixed time interval or on a fixed area in space

Examples:

- *Number of failures in a large computer system during a day.*
- *Number of replacement orders in each month.*
- *Number of defects in a large roll of sheet metal used to manufacture filters.*

$$\frac{e^{-\mu}\mu^{k}}{k!}$$

Key assumptions

1. The number of successes that occur in a unit of measure is independent of the number of successes that occur in any non-overlapping unit of measure.

2. The probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit.

3. The probability that 2 or more successes will occur in a unit approaches 0 as the size of the unit becomes smaller.

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 $\mu_S = \mu_X + \mu_Y$

Multiplication rule for independent events

- Events A and B are independent if that one occurs does not affect the probability that the other occurs
  => $P(A \ and \ B) = P(A).P(B)$

- If A and B are independent, the correlation is zero

  Note: If A and B are independent, the correlation between A and B is zero. But: The reverse may not be true! That is: If the correlation is zero, A and B are not necessarily independent. Think of non-linear associations!

Sampling distribution of a sample mean

- A statistic from a random sample will take different values if we take more samples from the same population.
- Sample statistics are random variables

Law of large numbers:
- Population mean μ must be finite
- Respondents are independent and randomly drawn

Central Limit Theorem:

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

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

*Regardless of the shape of the original distribution. If n is large enough, the distribution of the sample mean will be approximately normal.

Applied Statistics 1 - IBEB - Lecture 3.2 - Week 3

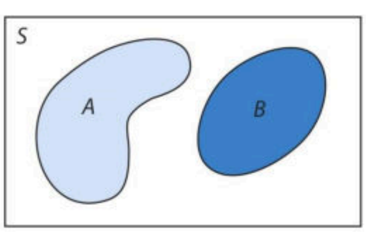
General probability rules

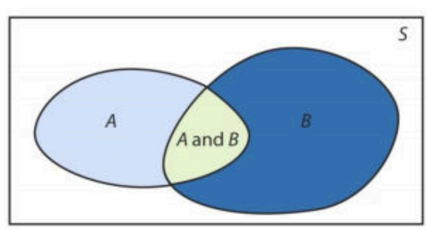
1. $0 \leq P(A) \leq 1$ for any event A
2. $P(S) = 1$, for S being the sample space
3. Complement rules: $P(A\ does\ not\ occur) = P(A^c) = 1 - P(A)$
4. Additional rule: Two events A and B are disjoint if they have no outcomes in common and so can never occur simultaneously: $P(A\ or\ B) = P(A) + P(B)$

Venn diagram

In a Venn diagram, the total area is represented by the sample space (S), and all the events are drawn in that area.

- For two disjoint events: $P(A\ or\ B) = P(A) + P(B)$
- For two events that are not disjoint: $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$





Conditional probability

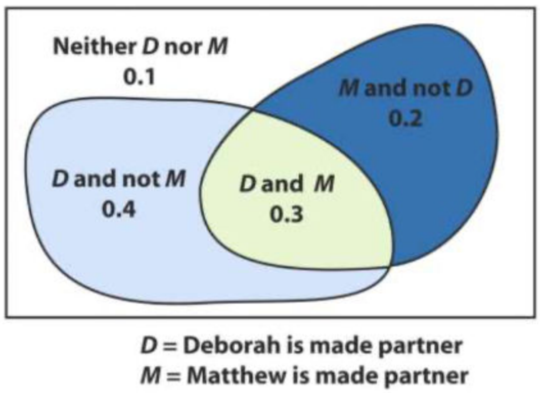
When $P(A) > 0$, the condition probability of B given A is:

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

The probability that both events A and B happen together is:

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

Example:

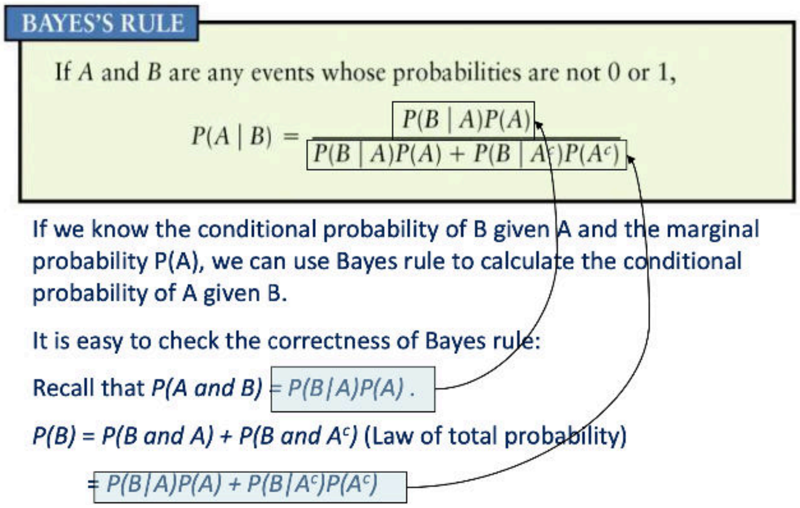


$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = 0.3/0.7 \approx 0.43$$

Applied Statistics 1 - IBEB - Lecture 4.1 - Week 4

Bayes's rule

Bayes rule makes it possible to go from one conditional probability, say P(A|B), to other the other conditional probability: P(B|A).



BAYES'S RULE

If A and B are any events whose probabilities are not 0 or 1,

$$P(A \mid B) = \frac{\boxed{P(B \mid A)P(A)}}{\boxed{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}}$$

If we know the conditional probability of B given A and the marginal probability P(A), we can use Bayes rule to calculate the conditional probability of A given B.

It is easy to check the correctness of Bayes rule:

Recall that *P(A and B)* = *P(B|A)P(A)* .

P(B) = *P(B and A)* + *P(B and A^c)* (Law of total probability)

= *P(B|A)P(A)* + *P(B|A^c)P(A^c)*

Decision theory

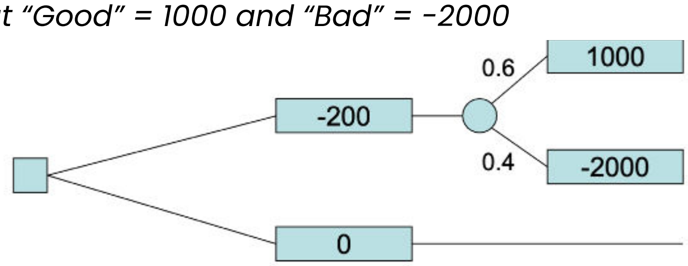
Probability: Conditional

To summarize the given probabilities, we use a tree diagram

Expected monetary value

- To make the best decision, we calculate the Expected Monetary Value (EMV) for each choice.
- EMV = the total sum of each option's money values times their associated probabilities. The choice that yields the highest EMV is considered.

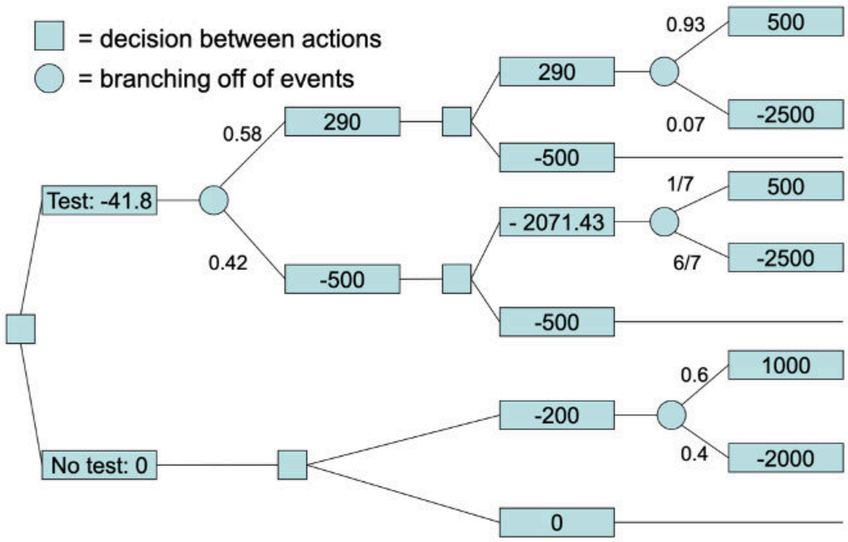
Example: Given that "Good" = 1000 and "Bad" = -2000



EMV(buy) = 0.6x1000 + 0.4x(-2000) = -200
EMV(do not buy) = 0
=> We shouldn't buy.

Decision tree

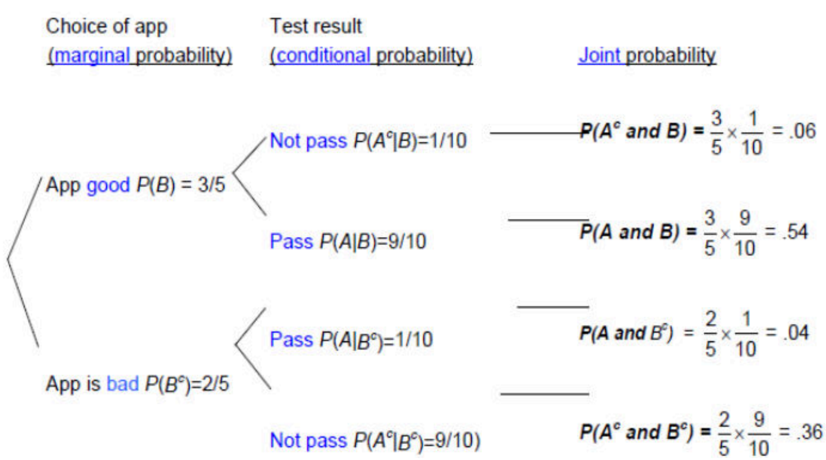
- A decision tree is useful when making a decision that involves uncertainty. The option that gives the highest EMV is usually chosen.
- A good strategy is to start filling the decision tree from right to left



Probability tree

- A probability tree is not the same as a decision tree.
- From a probability tree, we can easily infer the joint probability distribution.
- The joint probability distribution can be used to calculate both marginal as conditional probabilities.

Example:



Applied Statistics 1 - IBEB - Lecture 4.2 - Week 4

Introduction to inference

Definition

Statistical inference aims to make statements about the population based on the data obtained from a sample. It involves estimating the population parameters through sample statistics.

Confidence interval

- A confidence interval represents a range of plausible values for the population parameter (usually constructed based on a margin of error).
- Confidence interval = $[\overline{X} - margin\ of\ error, \overline{X} + margin\ of\ error]$

The sampling distribution of a sample mean

- A statistic from a random sample will take different values if we take more samples from the same population.
- Sample statistics are random variables
- The mean is an important random variable. The sample mean $\overline{X}$ will differ from sample to sample and is not equal to the population mean μ. Still, it is generally a reasonable estimate for the population mean.

Law of large numbers

Draw independent observations at random from any population with finite mean μ. As the number of observations drawn increases, $\overline{X}$ gets closer to μ.

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

When data are normally distributed
If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

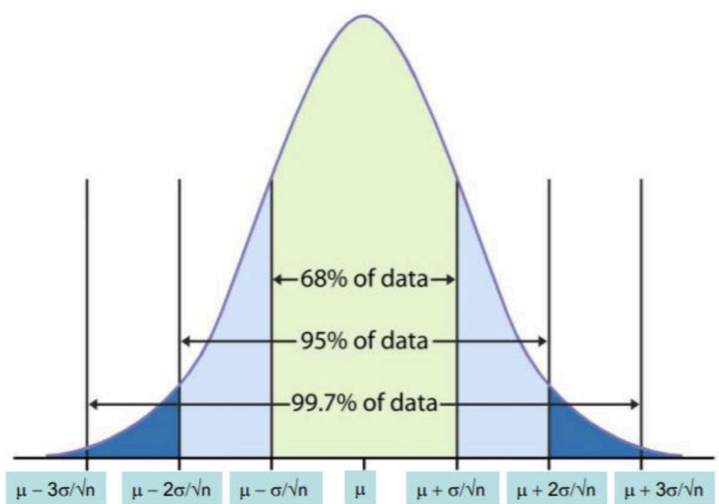
Central limit theorem

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. Regardless of the shape of the original distribution, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal:

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

Confidence interval

Margin of error: 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:

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

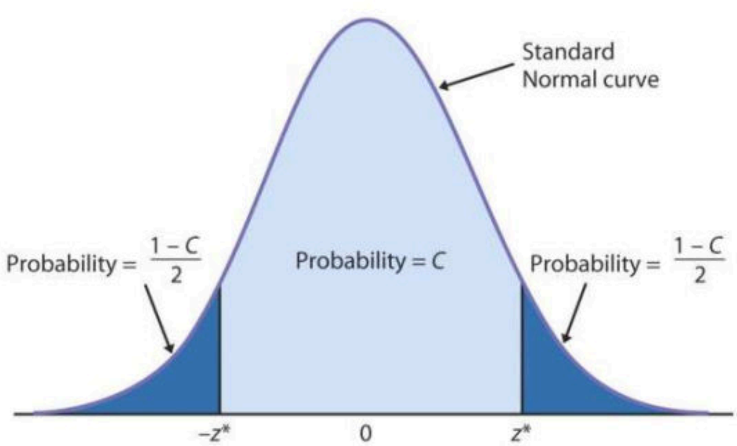
This means that there is 95% confidence that μ is in the interval $[\overline{x} - \frac{2\sigma}{\sqrt{n}}, \overline{x} + \frac{2\sigma}{\sqrt{n}}]$.

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:

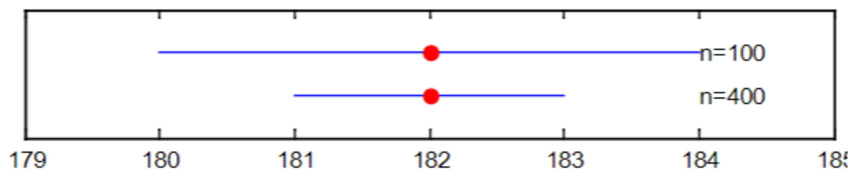
$$C = \bar{x} \pm z^{*}\frac{\sigma}{\sqrt{n}}$$

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

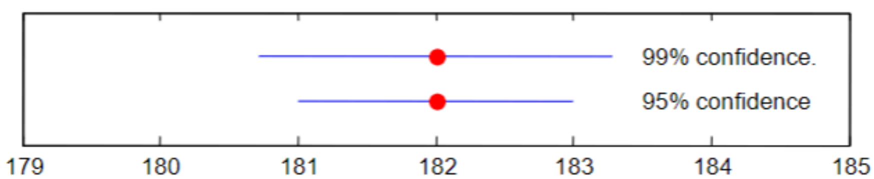
The interval is exact when the population distribution is normal and is approximately correct when n is large in other cases.

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: the greater it is, the smaller the interval becomes.
- Confidence level: the greater it is, the smaller the interval becomes.
- Critical variable z^{*}: the greater it is, the wider the interval becomes.

Choosing the sample size n

With known confidence level and margin of error, the sample size is where

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

Applied Statistics 1 - IBEB - Lecture 5.1 - Week 5

Hypothesis testing

Concepts

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

Alternative hypothesis: – Often the thing that you want to prove

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

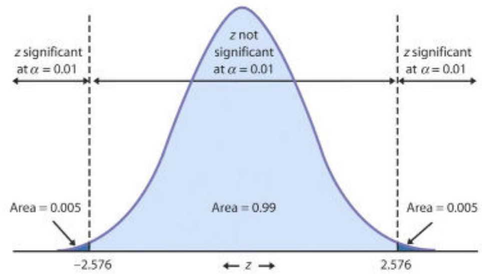
- To test a certain hypothesis, we need a test-statistic.
- A test statistic is a function from your sample for which you can evaluate how likely the null hypothesis is true.
- Formula: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

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

We reject the null hypothesis if the p-value is smaller than a certain significance level α.



Hypothesis testing advantages & disadvantages

Advantage: Clear decision (Reject, do not reject).

Disadvantage:
– Statistically significant results are not necessarily practically significant.
– Reject or do not reject completely ignores how strong the evidence against H0 is.
– If you test often, you may eventually find “statistically significant” results.

Summary

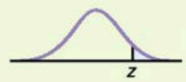
z TEST FOR A POPULATION MEAN

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**

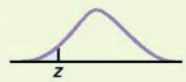
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

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

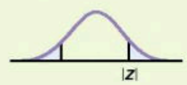
H_a: $\mu > \mu_0$ is $P(Z \geq z)$



H_a: $\mu < \mu_0$ is $P(Z \leq z)$



H_a: $\mu \neq \mu_0$ is $2P(Z \geq |z|)$



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 a hypothesis
2. Calculate test statistic (z-value)
3. Calculate P-value
4. Draw conclusions (given the significance level in the question)

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 a = 5%.

Solution:

1. Formulate hypotheses (start with alternative):

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

2. Calculate test statistic: $$z = \frac{\bar{X} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{50.575 - 50}{\left(\frac{1.65}{\sqrt{40}}\right)} = 2.20$$

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

Definition: The probability that a fixed level of significance α will reject a null hypothesis H_0 when a particular alternative value of the parameter is true.

Type 1 and type 2 errors

- Type I error: Rejecting H_0 when it is true. Its power equals α: $P(\bar{X}$ is in the rejection region | H_0 is true)
- Type II error: Not rejecting H_0 when H_a is true. The probability for a Type II error, β, is equal to: 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$=1-\beta$

Decision based on sample	Truth about the population: H_0 true	Truth about the population: H_a true
Reject H_0	Type I error	Correct decision
Not reject H_0	Correct decision	Type II error

Confidence intervals and hypothesis testing

A two-sided significance test of level α rejects a hypothesis H_0: $\mu = \mu_0$ just 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 $< \alpha$.
   => Reject if $|z| > z^*_{\alpha/2}$ (two-sided test) or $z > z^*_{\alpha}$

- The test statistic is:

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

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

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

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($\overline{X}$ in rejection region | $\mu = \mu_a$)

Example:

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

Inference for means

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

σ is unknown and thus replaced by estimator s: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$ with $n - 1$ degrees of freedom (how spread the distribution is compared to normal distribution).

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: $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$

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

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

<!> In the one-sample t test, to find approximate p-values, we find the critical value closest to the p-value observed and associate it with the corresponding degrees of freedom (closest number in the corresponding row in Table D).

Non-normality

The t statistic is valid only if the population is normally distributed.
If normality does not hold, a one-sample t test can still be used if:

- 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

Procedure:

1. Calculate the difference between the ratings for each individual in the panel.
2. Construct a CI for the difference, or perform a 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:

$$\boxed{{s_D}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2}$$

Sign test

Sign test: A test on the median.

- Insensitive to outliers
- Uses no distributional assumptions.

THE SIGN TEST 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

As some of the probability may not be found on the table given, you are advised to use the normal approximation for binomial distributions:

> **NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS**
>
> 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, $N(np, \sqrt{np(1-p)})$.
>
> As a rule of thumb, we will use the Normal approximation when n and p satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Summary of some important testing results

If the standard deviation is 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 there is no difference, the number of plusses follows binomial distribution with p=0.5.

Comparison of two groups: Independent samples

Procedure:

1. Calculate the mean ratings for the two groups
2. Construct a 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:

TWO-SAMPLE z STATISTIC

Suppose that $\overline{x}_1$ is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that $\overline{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the **two-sample z statistic**

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

has the standard Normal $N(0, 1)$ sampling distribution.

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

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

With a small sample and both populations normally distributed:

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

$$\text{df} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}.$$

THE TWO-SAMPLE t CONFIDENCE INTERVAL

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. The **confidence interval for $\boldsymbol{\mu_1 - \mu_2}$** given by

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

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:

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

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 $\bar{X} \sim N(\mu_x, \sqrt{\frac{\sigma_X^2}{n_x}})$ and $\bar{Y} \sim N(\mu_y, \sqrt{\frac{\sigma_Y^2}{n_y}})$: $\quad \bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}})$

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:

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

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

$$s_p^{\,2} = \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$$

THE POOLED TWO-SAMPLE *t* PROCEDURES

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**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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?

- It all depends on the variances. If they are equal, it is better to use the pooled variance.
- Before doing a t-test to test a difference in means, we first test whether variances differ:

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

To test for equality of variance we consider the following test statistic:

$$F = \frac{s_1^{\,2}}{s_2^{\,2}}$$

If F becomes too large, this may indicate that the variance of variable/group 1 is larger than that of variable/group 2. And vice versa.

THE *F* STATISTIC AND *F* DISTRIBUTIONS

When s_1^2 and s_2^2 are sample 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 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. F is always positive (variances greater than zero)

Summary of some important testing results

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

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

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,

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

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:

$$\boxed{\hat{p} = \bar{X} = X / n}$$

Source: Lecture 7.1 Applied Statistics 1, slide 20 (van de Velden, 2023)

- 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

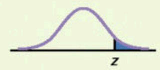
LARGE-SAMPLE TEST FOR A POPULATION PROPORTION

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**

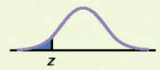
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

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

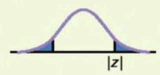
H_a: $p > p_0$ is $P(Z \geq z)$



H_a: $p < p_0$ is $P(Z \leq z)$



H_a: $p \neq p_0$ is $2P(Z \geq |z|)$



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

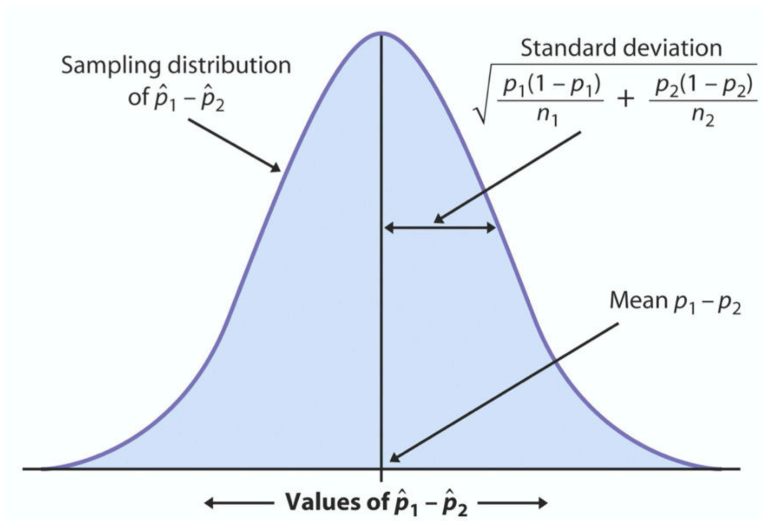
If the two are independent, it follows that the difference between the two proportions is also approximately Normally distributed.

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

SIGNIFICANCE TESTS 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

$$H_0\colon p_1 = p_2$$

compute the z **statistic**

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\mathrm{SE}_{Dp}}$$

where the **pooled standard error** is

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

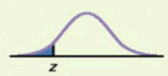
based on the **pooled estimate** of the common proportion of successes,

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

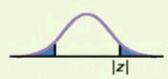
In terms of a standard Normal random variable Z, the P-value for a test of H_0 against

$H_a\colon p_1 > p_2$ is $P(Z \geq z)$

$H_a\colon p_1 < p_2$ is $P(Z \leq z)$



$H_a\colon p_1 \neq p_2$ is $2P(Z \geq |z|)$



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

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

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

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

Van de Velden, M. (2024). Applied Statistics 1 Lecture 3.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92654101

Van de Velden, M. (2024). Applied Statistics 1 Lecture 3.2 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92814464

Van de Velden, M. (2024). Applied Statistics 1 Lecture 4.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92887232

Van de Velden, M. (2024). Applied Statistics 1 Lecture 4.2 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/92970679

Van de Velden, M. (2024). Applied Statistics 1 Lecture 5.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/93093261

Van de Velden, M. (2024). Applied Statistics 1 Lecture 6.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/93272944

Van de Velden, M. (2024). Applied Statistics 1 Lecture 7.1 [Lecture Slides]. Retrieved from: https://canvas.eur.nl/courses/44026/files/93433086