­­­

Data Science & AI

Samenvatting

Giles De Praeter

Contents

[1. Sampling 3](#_Toc104807070)

[2. Univariate 5](#_Toc104807071)

[3. Central limit theorem 7](#_Toc104807072)

[3. Hypothesis testing 14](#_Toc104807073)

[4. Bivariate qual qual 17](#_Toc104807074)

[5. Bivariate qual quant 22](#_Toc104807075)

[6. Bivariate quant quant 26](#_Toc104807076)

[7. Timeseries 29](#_Toc104807077)

# Sampling

* Based on empirical research we are interested in:
  + Exploration
  + Description
  + Prediction
  + Verification
  + *Empirical Resarch is a type of research where the conclusions and decisions are based on valid data or evidence. Empirical research can be analyzed quantitatively or qualitatively.*
* The research process:

Diagram

Description automatically generated

* **Variable**: General property of an object, allows to distinguish objects
* **Value**: Specific property, interpretation for that variable
* **Measurement levels** = variable types
  + Determine most suitable method for analysis
    - visualization methods
    - central tendency and dispersion
    - examine the relationship between variables
* **Qualitative**: Not necessarily numeric limited number of values (words)
* **Quantitative**: Number + unit of measurement. Many values, often unique (numbers)
  + Often contain the result of a measurement
* **Nominal data** simply names something without assigning it to an order in relation to other numbered objects or pieces of data. (categories)
* **Ordinal data**, unlike nominal data, involves some order; ordinal numbers stand in relation to each other in a ranked fashion. (order, rank)
* **Interval**: no fixed zero point -> no proportions (you can’t say 20 degrees is 1/3 warmer than 15, because if you convert to Fahrenheit, it isn’t true anymore)
* **Ratio**: absolute zero point: 20m is 1/3 longer than 15m
* **Variables are related** if their values change systematically

Chart, scatter chart

Description automatically generated

* Researchers are often looking for causal relationships, e.g.
  + - Frustration leads to aggression
    - Alcohol leads to decreased alertness
    - ...
  + Cause Independent variable
  + Consequence Dependent variable
  + A relationship between variables does not necessarily indicate a causal relation!
* **Population**: the collection of all objects/people/… that you want to investigate
* **Sample**: a subset of the population from which measurements will be taken
  + Under certain circumstances, the results for a sample are representative for the population.
* **Random sample**: every element from the population has an equal chance of being included in the sample.
* **Non-random sample:** the elements for the sample are not randomly selected. Objects that can be collected easily are more likely to be included (convenience sampling).
* **Possible errors:**
  + Measurements in a sample will typically deviate from the value in the entire population ⇒ Errors!
  + Accidental ↔ Systematic
  + Sampling error ↔ Non-sampling error
* **Sampling Errors**:
  + Accidental sampling errors
    - Pure coincidence
  + Systematic sampling errors
    - Online survey: people without internet are excluded
    - Street survey: only who is currently walking there
    - Voluntary survey: only interested parties participate
* **Non-sampling errors**:
  + Accidental non-sampling errors
    - Incorrectly ticked answers
  + Systematic non-sampling errors
    - Poor or non-calibrated measuring equipment
    - Value can be influenced by the fact that you measure
    - Respondents lie

# Univariate

Measure of Central Tendency: What values is representative of the entire group?

* **The arithmetic mean** (notation: ) is the sum of all values divided by the number of values

A picture containing text, clock

Description automatically generated

* To find the **median**, sort all values and pick the middle number
  + Odd number of values: no problem
  + Even number of values: average of the middle two
* The **mode** is the value that appears most often in a dataset

Measures of Dispersion: How large are the differences within the group?

* **The range of a dataset** is the absolute value of the difference between the highest and the lowest value.
* **The quartiles of a sorted set** of numbers are the three values that divide the set into 4 equally large subsets. Notation: 𝑄1, 𝑄2, 𝑄3
* Different software programs have slightly different ways of calculating quartiles.
* The following method is easy to perform by hand. Start by sorting the values.
  + When 𝑛 is odd.
    - The median (𝑄2) is the middle value (as before).
    - Leave out the median. 𝑄1 is the median of the first half, 𝑄3 is the median of the second half.
  + When 𝑛 is even.
    - The median (𝑄2) is the average of the two middle values.
    - 𝑄1 is the median of the first half, 𝑄3 is the median of the second half.
* **The variance** (𝑠2 or 𝜎2) is the mean squared difference between the values of a data set and the arithmetic mean.

A picture containing text, clock, watch, gauge

Description automatically generated

* **The standard deviation** (𝑠 or 𝜎) is the square root of the variance
* Providing only a center value is never sufficient!
  + What is the dispersion?
  + How is the data distributed? Normal distribution?
  + Is the group sufficiently homogeneous?

Text

Description automatically generated

Table

Description automatically generated

* Avoid using a pie chart!
  + Disadvantages:
    - Comparing angles is harder than comparing length
    - Unusable for data with many categories

A picture containing chart

Description automatically generated

Text

Description automatically generated

# Central limit theorem

Probability distribution of a sample

* **Probability** represents the relative frequency of the occurrence of the event at hand (when performing a large number of independent experiments).
* Probabilities are numbers assigned to sets.
* These sets are part of some all-encompassing set, the “**universe**”, typically denoted 𝛺.
* The numbers (probabilities) assigned to the sets have to satisfy three simple rules (see later) in order to correspond to our intuition about how probabilities should behave.
* **Axioms of probability:**
  + Probabilities are non-negative: 𝑃(𝐴) ≥ 0 for each 𝐴.
  + The universe has probability 1: 𝑃(𝛺) = 1.
  + When 𝐴 and 𝐵 are disjoint events (i.e. 𝐴 ∩ 𝐵 = ∅) then it holds that
    - 𝑃(𝐴 ∪ 𝐵) = 𝑃(𝐴) + 𝑃(𝐵).
    - This is called **the sum rule**.
* From the three axioms of probability, we can derive all **properties of probabilities**. Some important ones are listed below:
  + **Complement rule**: for each 𝐴 it holds that
    - 𝑃(Ā) = 1 − 𝑃(𝐴)
    - where Ā represents the event “A does not occur”.
  + The impossible event has probability zero: 𝑃(∅) = 0.
  + **The general sum rule**:
    - 𝑃(𝐴 ∪ 𝐵) = 𝑃(𝐴) + 𝑃(𝐵) − 𝑃(𝐴 ∩ 𝐵).
* **Independent events**:
  + Mathematically, two events 𝐴 and 𝐵 are independent when:
    - 𝑃(𝐴 ∩ 𝐵) = 𝑃(𝐴)𝑃(𝐵)
* FOR EXAMPLES, SEE SLIDES HF 3
* **Expectation of a random variable:**
  + The expectation of a random variable is denoted by 𝜇𝑋 or E(𝑋) and is given by

Text

Description automatically generated with medium confidence

* + Note the similarity between this formula and the way you calculate the sample mean when a frequency table is given.
  + We often write 𝜇 instead of 𝜇𝑋 .
* **Variance of a random variable:**
  + The variance of a R.V. is a measure of dispersion that resembles the sample variance very closely.
  + The variance of a random variable is defined by:

Text

Description automatically generated with low confidence

* + Note: the standard deviation is the positive square root of the variance

A picture containing text

Description automatically generated

* **Probability density function:**
  + In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would be close to that sample. In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there is an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would be close to one sample compared to the other sample.
  + **A continuous random variable** takes on an uncountably infinite number of possible values.
  + In this case it doesn’t make a lot of sense to consider the probability that 𝑋 equals some number 𝑎 exactly, because this probability is always zero.
  + What does make sense is to consider the probability that 𝑋 takes on a value in some interval [𝑎, 𝑏].
  + This probability can be found be integrating (i.e. “summing up”) the probability density function of the random variable.

Graphical user interface, application, Word

Description automatically generated

Chart, line chart

Description automatically generated

Text

Description automatically generated

* **Exponential distribution**
  + Besides the Normal Distribution, there are other often used continuous distributions, e.g. the Exponentional Distribution.
  + Values for an exponential random variable occur when there are fewer large values and **more small values**. For example, the amount of money customers spend in a trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.
  + Another examples is the length, in minutes, of long distance business telephone calls.

Chart

Description automatically generated with medium confidence

* **Continuous uniform distribution:**
  + The continuous Uniform Distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.
  + It is the simplest of all continuous probability distributions. The density function is constant where every value has an equal chance of occurring. Imagine you live in a building that has an elevator that will take you to your floor. From experience, once you push the button to call the elevator, it takes between ten and twenty seconds to arrive at your floor. This means the elevator arrival time is uniformly distributed between 10 and 20 seconds once you hit the button.

Graphical user interface

Description automatically generated with medium confidence

* **Central limit theorem:**
  + If the size of the sample is sufficiently large, the probability distribution of the sample mean will approximate a normal distribution, regardless of the probability distribution of the underlying population.

Text, letter

Description automatically generated



* **Point estimate:**
  + A Point Estimate for a population parameter is a formula or equation that allows us to calculate a value to estimate that parameter.
* **Confidence interval:**
  + A confidence interval is an equation or formula that allows us to construct an interval that will contain the parameter to be estimated with a certain level of confidence.

Chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated



Chart, line chart, histogram

Description automatically generated

Text

Description automatically generated

# Hypothesis testing

* **Hypothesis**: Idea that has yet to be proven: statement regarding the numeric value of a population parameter
* **Hypothesis Test**: verification of a statement about the values of one or multiple population parameters
* **Null Hypothesis (𝐻0):** Base hypothesis, we start with assuming it is true
* **Alternative Hypothesis (𝐻1, 𝐻𝑎):** Conclusion if the null hypothesis is unlikely to be true
* **Test Statistic:** The value that is calculated from the sample
* **Region of Acceptance:** The region of values supporting the null hypothesis
* **Critical Region / Region of Rejection**: The region of values rejecting the null hypothesis
* **Significance Level:** The probability of rejecting a true null hypothesis 𝐻0
* **Testing procedure:**
  + Formulate both hypotheses (𝐻0 and 𝐻1)
  + Determine the significance level (𝛼)
  + Calculate the test statistic
  + Determine the critical region or the probability value
  + Draw conclusions
* **Probability value:**
  + The 𝑝-value is the probability, if the null hypothesis is true, to obtain a value for the test statistic that is at least as extreme as the observed value.
  + 𝑝-value < 𝛼 ⇒ reject 𝐻0: the discovered value of  is too extreme;
  + 𝑝-value ≥ 𝛼 ⇒ do not reject 𝐻0: the discovered value of can still be explained by coincidence.

Chart, line chart, histogram

Description automatically generated

* **The critical region** is the collection of all values of the test statistic for which we can reject the null hypothesis.

Text, letter

Description automatically generated

* + Left of 𝑔: region of acceptance (do not reject 𝐻0)
  + Right of 𝑔: critical region (reject 𝐻0)

Chart, line chart, histogram

Description automatically generated

* **Left-tailed testing:**
  + What would you have to change in the equation in order to calculate the correct critical value? Answer:

Text, letter

Description automatically generated



Text, letter

Description automatically generated

* **Two-tailed testing:**
  + Sometimes it can be necessary to perform a two-tailed test. In this case, two critical values need to be calculated, namely the left and right critical value.

A picture containing text

Description automatically generated



* Summary:

Text, table

Description automatically generated

* **Requirements for z-test:**
  + The sample needs to be random
  + The sample size needs to be sufficiently large (𝑛 ≥ 30)
  + The test statistic needs to have a normal distribution
  + The standard deviation of the population, 𝜎, is known
  + Sometimes these assumptions will not hold and in this case we can not use the 𝑍-test!
* What if the requirements for a 𝑧-test are not met? E.g.
  + Sample size too small (n < 30)
  + Population stdev (𝜎) unknown
  + If the variable is normally distributed, you can use the 𝑡-test
* **T-test:**
  + Determine critical value:

Chart

Description automatically generated with medium confidence

* + 𝑡-value is derived from the Student’s 𝑡-distribution, based on the number of degrees of freedom, 𝑛 − 1
  + Look for value using the function t.isf() in Python
  + Apart from this, the procedure is identical to the procedure of the 𝑧-test

Graphical user interface, text, application

Description automatically generated

# Bivariate qual qual

* Overview:

Table

Description automatically generated with medium confidence

* **Bivariate analysis:**
  + …is determining whether there is an association between two stochastic variables (𝑋 and 𝑌 ).
  + **Association** = you can predict (to some extent) the value of 𝑌 from the value of 𝑋
  + **𝑋** — Independent variable
  + **𝑌** — Dependent variable
  + Important! Finding an association does NOT imply a causal relation!

Graphical user interface, text

Description automatically generated

* Contingency tables and margin totals:

Table

Description automatically generated

Table

Description automatically generated

* **The chi-squared statistic:**
  + Sum of all these values:

Graphical user interface, text, application

Description automatically generated

* + **When is chi large enough?**
    - 2 × 2-table with 𝜒2 = 10
      * Relatively large difference
      * Indicates association
    - 5 × 5-table with 𝜒2 = 10
      * Relatively small difference
      * Does NOT indicate association
    - We need a metric independent of table size!
* **Cramér’s V:**

Table

Description automatically generated

* **Chi-squared test for independence:**
  + = Alternative to Cramér’s V to investigate association between qualitative variables.
  + Value of 𝜒2 distributed according to the 𝜒2 distribution

A picture containing shape

Description automatically generated

Text, table

Description automatically generated

* + Test procedure:

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

* **Goodness-of-fit-test**
  + The 𝜒2 test can also be used to determine whether a sample is representative for the population.
  + Goodness-of-fit-test: This test indicates to what degree a sample corresponds to a null hypothesis regarding the distribution of a qualitative variable over mutually exclusive classes.
  + LOOK AT PPT DIA 25 FOR EXAMPLE!

Text, letter

Description automatically generated

* **Standardized residuals:**
  + The standardized residuals indicate which classes make the greatest contribution to the value of 𝜒2.

Text, letter

Description automatically generated

* **Cochran’s rules:**
  + In order to apply the 𝜒2-test, the following conditions must be met (Rule of Cochran)
    - For all categories, the expected frequency 𝑒 must be greater than 1.
    - In a maximum of 20 % of the categories, the expected frequency 𝑒 may be less than 5.

# Bivariate qual quant

* Independent qualitative, dependent quantitative

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* **Two-sample t-test**
  + Comparing two population means
  + We use two samples to perform an appropriate statistical test
    - Independent samples
    - Paired samples
* **Independent samples:**

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

* **Paired samples:**

Text, table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

* **Effect size:**
  + The effect size is a metric which expresses how great the difference between two groups is
  + Control group vs. intervention group
  + Can be used in addition to hypothesis test
  + Often used in educational sciences
  + There are several definitions, here: Cohen’s 𝑑

Graphical user interface, application, table

Description automatically generated

# Bivariate quant quant

* Independent quantitative, dependent quantitative

Chart, scatter chart

Description automatically generated

* **Regression:**
  + A measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).
  + With regression we will try to find a consistent and systematic relationship between two quantitative variables.
    - Monotonic: consistent direction of the relationship between the two variables: increasing or decreasing
    - Non-monotonic: value of dependent variable changes systematically with value of independent variable, but the direction is not consistent
* **Linear regression:**
  + A linear relationship between an independent and dependent variable.
  + Characteristics:
    - Presence: is there a relationship?
    - Direction: increasing or decreasing?
    - Strength of the relationship: strong, moderate, weak, nonexistent, …
* **Method of least squares:**

Table

Description automatically generated

* **Covariance:**
  + Covariance is a measure that indicates whether a relationship between two variables is increasing or decreasing.

Text

Description automatically generated

Chart

Description automatically generated

* **Pearson correlation coefficient**
  + Pearson’s product-moment correlation coefficient 𝑅 is a measure for the strength of a linear correlation between 𝑥 and 𝑦

Text, letter

Description automatically generated

Diagram, engineering drawing, scatter chart

Description automatically generated

* **Coefficient of determination:**
  + The coefficient of determination 𝑅2 explains the percentage of the variance of the observed values relative to the regression line.
  + 𝑅2: percentage variance observations explained by the regression line
  + 1 − 𝑅2: percentage variance observations not explained by regression
  + Interpretation:

**Table

Description automatically generated**

* **Considerations:**
  + The correlation coefficient only looks at the relationship between two variables. Interactions with other variables are not considered.
  + The correlation coefficient explicitly does not assume a causal relationship.
  + Pearson’s correlation coefficient only expresses linear relationships.

# Timeseries

* Time series:
  + A time series is a sequence of observations of some variable over time.
    - Monthly demand for milk
    - Annual intake of students at HOGENT
    - Price of a share or bond on the stock exchange (hourly, daily, …)
    - Number of HTTP requests per second for a website
    - Evolution of disk usage on a backup server
  + Time series are a statistical problem: observations vary with time

Chart, scatter chart

Description automatically generated

* Time series components:
  + Level
  + Trend
  + Seasonal fluctuations
  + Cyclic patterns
  + Random noise (residuals)

Chart, line chart

Description automatically generated

* Time series decomposition:

Background pattern

Description automatically generated

* Mathematical model time series:
  + The simples model:

Text, letter

Description automatically generated

A picture containing text, screenshot, bird

Description automatically generated

* + General expression time series:

Text, letter

Description automatically generated

* Estimating the parameters:
  + Make predictions based on the time series model:
    - 1. select the most suitable model
      2. estimation for parameters 𝑏𝑖(𝑖 ∶ 1, … , 𝑛) based on observations
  + The estimations 𝑏𝑖 are selected so that they approximate the observed values as close as possible.
  + Example (next page):

Graphical user interface, application

Description automatically generated

* Moving average:
  + The moving average is a series of averages (means) of the last 𝑚 observations

Text

Description automatically generated with medium confidence

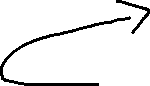
* + Example:

Chart, histogram

Description automatically generated

* Weighted moving average:
  + For 𝑆𝑀𝐴, the weights of the observations are equal
  + For a weighted moving average (𝑊𝑀𝐴), more recent observations gain relatively more weight
  + A specific form of this is single exponential smoothing or the exponential moving average (EMA):





* + - with 𝛼 the smoothing constant (0 < 𝛼 < 1), and 𝑡 ≥ 3
* exponential smoothing:
  + Equation 6 is only valid from 𝑡 = 3. Hence, we need to choose a suitable value for 𝑋2 ourselves. There are several options:

Text

Description automatically generated

* + Example:

Chart

Description automatically generated

* + Why exponential?
    - Older observations have an exponentially smaller weight
    - Wiskundig bewijs (zie cursus, niet te kennen)
  + The speed at which the old observations are “forgotten” depends on the value of 𝛼. For a value of 𝛼 close to 1, old observations are quickly forgotten, whereas for 𝛼 close to 0, this goes less fast.
  + Forecasting:
    - As a forecast for time 𝑡 + 𝑚 (𝑚 time units in the “future”), we always take the last estimate of the level:

Graphical user interface, chart

Description automatically generated

* Double exponential smoothing:
  + Basic exponential smoothing does not work well if there is a trend in the data
  + We add an additional term to model the trend. We use 𝑏𝑡 for the estimation of the trend at time 𝑡 > 1:

Text, letter

Description automatically generated

Text, letter

Description automatically generated

* + Forecasting:

Text, letter

Description automatically generated

Chart, line chart

Description automatically generated

* Triple exponential smoothing or Holt-winter’s method:
  + Some time series have recurring (seasonal) patterns:

Chart, line chart

Description automatically generated

* + Notation:

Text, letter

Description automatically generated

* + Prediction:

A picture containing chart

Description automatically generated

Chart, line chart

Description automatically generated

Text

Description automatically generated