Inhoudsopgave: Datum:

[1 Hoofdstuk 1: Basic Concepts and Sampling 2](#_Toc198644720)

[1.1 Basic Concepts and Sampling 2](#_Toc198644721)

[1.2 Sample testing 3](#_Toc198644722)

[2 Hoofdstuk 2: Univariate statistics 6](#_Toc198644723)

[2.1 Central Tendency and Dispersion 6](#_Toc198644724)

[2.2 Data Visualisation 7](#_Toc198644725)

[3 Hoofdstuk 3: Probability, The Central Limit Theorem, Confidence Intervals 8](#_Toc198644726)

[3.1 Probability 8](#_Toc198644727)

[3.1.1 Discrete random variables 8](#_Toc198644728)

[3.1.2 Expectation and variance of a random variable 11](#_Toc198644729)

[3.1.3 Continuous Random Variables 12](#_Toc198644730)

[3.1.4 The (Standard) Normal Distribution 13](#_Toc198644731)

[3.2 From Sample to Population 15](#_Toc198644732)

[3.2.1 The Central Limit Theorem 15](#_Toc198644733)

[3.2.2 Confidence Intervals Confidence Interval for a Large Sample 15](#_Toc198644734)

[3.2.3 Confidence Interval for a Small Sample 15](#_Toc198644735)

[4 Hoofdstuk 4: Bivariate analysis: qualitative variables 17](#_Toc198644736)

[4.1 Bivariate analysis 17](#_Toc198644737)

[4.2 Goodness-of-fit test 17](#_Toc198644738)

# Hoofdstuk 1: Basic Concepts and Sampling

## Basic Concepts and Sampling

**Variabelen en waarden:**

* **Variabele**
  + Algemene eigenschap van een object, maakt het mogelijk om objecten te onderscheiden
* **Waarde**
  + Specifieke eigenschap, interpretatie voor die variabele

Measurement Levels:

* Typen variabelen
* Meest geschikte methode voor analyse bepalen
  + visualisatiemethoden
  + centrale tendens en spreiding
  + relatie tussen variabelen onderzoeken

Qualitative vs quantitative:

* **Kwalitatief** 
  + Niet noodzakelijkerwijs numeriek
  + Beperkt aantal waarden
* **Kwantitatief**
  + Getal + meeteenheid
  + Veel waarden, vaak uniek

Qualitative scales:

* **Nominaal** 
  + Dit zijn de categorieën. Dit zijn ook kwalitatieve eigenschappen. Geen rangschikking
    - geslacht
    - ras,
    - land
    - vorm, ...
* **Ordinaal** 
  + Rangorde. Dit zijn de kwantitatieve eigenschappen. Wel een rangschikking
    - militaire rang,
    - opleidingsniveau, ...

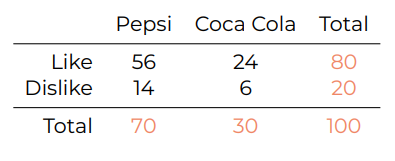
Quantitative scales:

* **Interval** 
  + Geen vast nulpunt ⇒ geen verhoudingen
  + Voorbeeld: °C, °F
* Ratio
  + Een absoluut nulpunt => proporties
  + Voorbeeld: afstand (m), energy (j), gewicht (kg)
* **Verhouding** 
  + Absoluut nulpunt ⇒ verhoudingen
  + Voorbeeld afstand (m), energie (J), gewicht (kg) ...
* **Verhoudingen**:
  + 20 m is 1/3e of ongeveer 33% langer dan 15 m
  + 20 °C is NIET 1/3e warmer dan 15 °C (omrekenen naar °F)

Afbeelding met lijn, diagram, Perceel, tekst

Door AI gegenereerde inhoud is mogelijk onjuist.Relations between variables:

Variabelen zijn gerelateerd als hun waarden **systematisch** veranderen.

Relations between variables example:

* Is er een **verband** tussen het type cola en de smaakwaardering?
  + Ja: de **verhoudingen hier zijn hetzelfde**: 14/56 = 25% en 6/24 is ook 25%

Causal Relationships:

* Onderzoekers zijn vaak op zoek naar **causale verbanden**, bijv.
  + Frustratie leidt tot agressie
  + Alcohol leidt tot verminderde alertheid
* **Oorzaak** 
  + Onafhankelijke variabele
* **Gevolg** 
  + Afhankelijke variabele

Fake correlations or “Spurious correlations”:

* **Waarschuwing**
  + Een verband tussen variabelen duidt niet noodzakelijk op een oorzakelijk verband!
* **Voorbeelden**:
  + Gewelddadige videospelletjes leiden tot gewelddadig gedrag
  + Vaccins kunnen autisme veroorzaken
  + Relatie tussen cola light drinken en obesitas

## Sample testing

Afbeelding met tekst, cirkel, Lettertype, logo

Door AI gegenereerde inhoud is mogelijk onjuist.Suppose you want to analyze a group of friends:

* Vragen die je kunt stellen:
  + Hoe lang zijn mijn vrienden?
  + Wat zijn hun gewichten?
  + Hoe veilig is hun leefomgeving?
  + Hebben ze familie? ...

Sample and Population:

* **Populatie** de verzameling van alle objecten/mensen/... die je wilt onderzoeken
* Afbeelding met Fictief personage, Superheld, Held, tekenfilm

  Door AI gegenereerde inhoud is mogelijk onjuist.**Steekproef een deelverzameling** van de populatie waaruit metingen worden gedaan
* Onder bepaalde omstandigheden zijn de resultaten van een steekproef representatief voor de populatie
* Een steekproef is gemakkelijker te analyseren dan de hele populatie

Afbeelding met tekst, schermopname, Lettertype, lijn

Door AI gegenereerde inhoud is mogelijk onjuist.Sampling method (aan te vullen in de les):

How to pick elements for a sample?

* **Random steekproef**: elk element uit de populatie heeft een gelijke kans om in de steekproef opgenomen te worden.
* **Niet-random steekproef**: de elementen voor de steekproef zijn niet willekeurig gekozen. Objecten die gemakkelijk kunnen worden verzameld, hebben meer kans om te worden opgenomen (gemakssteekproef).

Afbeelding met tekst, schermopname, Lettertype, nummer

Door AI gegenereerde inhoud is mogelijk onjuist.Stratified to variables (aan te vullen in de les):

Possible Errors:

* Metingen in een steekproef zullen meestal afwijken van de waarde in de hele populatie ⇒ **Errors**!
* Toevallig ↔ Systematisch
* Steekproeffout ↔ Niet-steekproeffout

Sampling Errors :

* **Toevallige steekproeffouten** 
  + Puur toeval – bv. je neemt een aantal studenten, maar neemt per ongeluk allemaal jongens
* **Systematische steekproeffouten** 
  + **Online enquête**: mensen zonder internet zijn uitgesloten
  + **Straatonderzoek**: alleen wie er op dat moment loopt
  + **Vrijwillige enquête**: alleen geïnteresseerden nemen deel

Non-sampling Errors:

* **Perongelijke niet-steekproeffouten** 
  + Verkeerd aangekruiste antwoorden
* **Systematische niet-steekproeffouten** 
  + **Slechte** of niet-gekalibreerde **meetapparatuur**
  + **Waarde** kan beïnvloed worden door het feit dat je meet
  + **Respondenten liegen** (aantal sigaretten per dag)

# Hoofdstuk 2: Univariate statistics

## Central Tendency and Dispersion

Mean or avarage:

* Afbeelding met Lettertype, lijn, nummer, ontwerp

  Door AI gegenereerde inhoud is mogelijk onjuist.**Rekenkundig gemiddelde**
  + Het rekenkundig gemiddelde (notatie: 𝑥) is de som van alle waarden gedeeld door het aantal waarden
* Q1 Wat gebeurt er als Ant-Man krimpt tot een grootte van 10 cm?
  + Het gemiddelde past zich redelijk wat aan
* Q2 Het rekenkundig gemiddelde van 15 getallen is 12. Welk getal moet erbij opgeteld worden om een gemiddelde van 13 te krijgen?
  + We moeten hier onze 15 \* 12 getallen doen. Daarna 15 \* 13 en dat verschil aftrekken van elkaar.

Mediaan:

* Om de mediaan te vinden, sorteer je alle waarden en kies je het middelste getal.
  + Oneven aantal waarden: geen probleem
  + Even aantal waarden: gemiddelde van de middelste twee
* Q1 Wat gebeurt er als Ant-Man krimpt tot een grootte van 10 cm?
* Q2 Wat is de mediaan van het aantal mensen dat gered is door Batman in de afgelopen acht jaar?

Mode:

Het getal dat het meeste voorkomt in een dataset

Measures of Dispersion:

Hoe ver alle getallen uit elkaar liggen

Bereik:

* Het vershil tussen het grootste verminderen van het kleinste getal
* 141 – 198 – 143 – 201 -184
* Hier is het bereik 201 – 141 = 60

Kwartiel:

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

Kwartielen berekenen:

* 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.

Interkwartiele afstand:

The interquartile range is the difference between the third and first quartile |𝑄3 − 𝑄1 |.

Variantie:

The variance (𝑠 2 or 𝜎 2 ) is the mean squared difference between the values of a data set and the arithmetic mean.

Standaardafwijking:

The standard deviation (𝑠 or 𝜎) is the square root of the variance

Vragen:

* Kan de standaardafwijking negatief zijn?
  + Nee, want het je voert een vierkantsworteloperatie hier op uit
* Wat is de kleinste waarde
* Wat is de eenheid van de standaardafwijking?
  + Centimeter
* Hoe moet je de standaardafwijking intepreteren
  + Hoe dicht liggen al die waarden rond het gemiddelde?

## Data Visualisation

# Hoofdstuk 3: Probability, The Central Limit Theorem, Confidence Intervals

## Probability

### Discrete random variables

Probality:

* Informally, probabilities represent beliefs of how likely it is that a certain event will happen when performing a certain experiment.
* In the frequentist view, probability represents the relative frequency of the occurrence of the event at hand (when performing a large number of independent experiments).

Probability: (Simple) Example:

* A fair six-sided die is thrown once.
* What is the probability of getting a “1”?
  + 1/6
* What is the probability of throwing an even number?
  + 1/2
* What is the probability of getting a number in the set {1, 2, 3, 4, 5, 6}?
  + 1

Probability: Example:

* Two games of chance:
* 1. Bet on: at least one six when throwing a fair die 4 consecutive times.
  + You will win in the long run when playing this game.
* 2. Bet on: at least one “double six” when throwing two fair dice 24 times.
  + You will lose in the long run when playing this game, even though in some sense the two bets seem equivalent. (Probability divided by six but number of throws multiplied by six.)

Probability: Observations:

* 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.

Example:

* For the first game: at least one six when throwing a fair die 4 consecutive times.
* What is the universe? Answer:
  + 𝛺 = {(1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 1, 3), … , (6, 6, 6, 4), (6, 6, 6, 5), (6, 6, 6, 6)}
* What set corresponds to the event: “you win” or “at least one 6”?
  + Answer 𝐴 = {(6, 1, 1, 1), (6, 1, 1, 2), … , (5, 5, 5, 6), (6, 6, 1, 1), … (6, 6, 5, 5), (1, 6, 6, 1), … , (6, 6, 6, 6)}

Axioms (Rules) of Probability:

* Axioms of Probability 1. Probabilities are non-negative: 𝑃(𝐴) ≥ 0 for each 𝐴.
* The universe has probability 1: 𝑃(𝛺) = 1
* Sum rule: When 𝐴 and 𝐵 are disjoint events (i.e. 𝐴 ∩ 𝐵 = ∅) then it holds that 𝑃(𝐴 ∪ 𝐵) = 𝑃(𝐴) + 𝑃(𝐵)

Properties of Probabilities:

* 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(𝐴) = 1 − 𝑃(𝐴)
* where 𝐴 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)

Examples:

* A coin is tampered with such that 𝑃(head) = 0.6. What is 𝑃(tail)?
  + 1 – 0.6 = 0.4
* From the **250 students** enrolled in this course, **200 own a PlayStation game console**, while **100 own a Nintendo Switch**. There are **75 students who own both game consoles**. What is the probability, when you pick a random student, that
  + this student does not own a PlayStation
    - 1 − 200/250 = 1/5
  + this student owns a PlayStation or a Nintendo Switch (or both)
    - (200 + 100 − 75)/250 = 225/250 = 9/10,
  + this student owns neither a PlayStation nor a Nintendo Switch.
    - 1 − 9/10 = 1/10

Independent events:

* Mathematically, two events 𝐴 and 𝐵 are **independent** when
* 𝑃(𝐴 ∩ 𝐵) = 𝑃(𝐴) x 𝑃(𝐵)

Independent Events: Example:

* When events are **independent** it means that the occurrence of one event (or knowing that the event occurred) does **not change** **the probability of the other event occurring**.
* As an example: **draw a card from a standard deck of cards and consider the following events:**
  + 𝐴 = a ♣ was drawn
* and
  + 𝐵 = an ace was drawn
* When I tell you that 𝐵 occurred, this will not change your opinion about the probability of 𝐴 occurring, this will still be 13/52 = 1/4. (Note: this also works the other way around!)
* Dit is zo omdat van de 52 kaarten (met 4 verschillende symbolen) van elk symbool exact evenveel kaarten zijn (13 per symbool)

Dependent Events: Example:

* Consider the events from before where students own either a PlayStation, a Nintendo Switch or both. Let **𝐴** be the event that a student **owns a PlayStation**, while **𝐵** represents the fact that a student **owns a Nintendo Switch**. From the numbers above we see that
  + **𝑃(𝐴) = 200 / 250 = 0.8,** and **𝑃(𝐵) = 100 / 250 = 0.4**
* when we multiply these two probabilities we get
  + **𝑃(𝐴)𝑃(𝐵) = 0.8 × 0.4 = 0.32**
* However, the probability that a student owns both a PlayStation and a Nintendo Switch, i.e. 𝐴 ∩ 𝐵 is given by
  + 𝑃(𝐴 ∩ 𝐵) = 75 / 250 = 0.3 ≠ 0.32
* Hence,
  + 𝑃(𝐴 ∩ 𝐵) ≠ 𝑃(𝐴)𝑃(𝐵)
* and the two events are **dependent**.
* In practical terms this means that e.g. knowing that a student owns a PlayStation gives you additional information about his or her ownership of a Nintendo Switch.
* Before you know that the student owns a PlayStation the probability that this student owns a Nintendo Switch is
  + 𝑃(𝐵) = 100 250 = 0.4
* However, once you know that that the student owns a PlayStation you know that this students belongs to the group of 200 PlayStation owners, of which 75 also own a Nintendo Switch. With fancy notation we write this as
  + 𝑃(𝐵 | 𝐴) = 75 200 = 0.375
* So the probability now is slightly lower.
* On the other hand, when you know that a student owns a Nintendo Switch, what is the probability of this student owning a PlayStation?
  + 𝑃(𝐴 | 𝐵) = 75 / 100 = 0.75

Discrete Random Variable:

* A player draws a card from a standard deck of cards. She gets 1 € when drawing a jack, 2 € when drawing a queen and 3 € when drawing a king. In all other cases she receives nothing. We can formalise this be defining a mapping 𝑋 from 𝛺 to ℝ:
* 𝑋 ∶ 𝛺 → ℝ ∶ 𝜔 ↦ 𝑋(𝜔) = {
  + - 1 when drawing a jack
    - 2 when drawing a queen
    - 3 when drawing a king
    - 0 in all other cases.
  + }
* We have done nothing else but associate a number with different events.

Probability Mass Function:

* Using the rules of probability we can assign probabilities to the different numerical outcomes
  + 𝑓𝑋 (3) = 𝑃(𝑋 = 3) = 𝑃(king) = 4 / 52 = 1 / 13,
  + 𝑓𝑋 (2) = 𝑃(𝑋 = 2) = 𝑃(queen) = 4 / 52 = 1 / 13,
  + 𝑓𝑋 (1) = 𝑃(𝑋 = 1) = 𝑃(jack) = 4 / 52 = 1 / 13,
  + 𝑓𝑋 (0) = 𝑃(𝑋 = 0) = 𝑃(any other card) = 40 / 52 = 10 / 13.
* It holds that 𝑓𝑋 (0) + 𝑓𝑋 (1) + 𝑓𝑋 (2) + 𝑓𝑋 (3) = 1.
* And also: 0 ≤ 𝑓𝑋 (𝑥𝑖 ) ≤ 1.

Probability mass function:

* To make things more explicit we can put the probability mass function in a table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | 0 | 1 | 2 | 3 |
| F x (x) = P(X = x) | 10 / 13 | 1 / 13 | 1 / 13 | 1 / 13 |

* We can also represent this probability mass function in a graph:

Probability Distribution: 1 die:

What is the probability of each outcome when rolling a single die?

Probability Distribution: 2 dice:

What is the probability of each outcome when rolling a 2 dice?

Using Probability Mass Func:

* Once you have a probability mass function, you can compute other probabilities.
* With the card game, what is the probability that you gain at most 2 EUR?
  + 𝑃(𝑋 ≤ 2) = 𝑃(𝑋 = 0) + 𝑃(𝑋 = 1) + 𝑃(𝑋 = 2) = 10 / 13 + 1 / 13 + 1 / 13 = 12 / 13
* We could also have used the complement rule in this case:
  + 𝑃(𝑋 ≤ 2) = 1 − 𝑃(𝑋 > 2) = 1 − 𝑃(𝑋 = 3) = 1 − 1 / 13 = 12 / 13.
* When throwing two dice, what is the probability that the sum of the results is between 4 and 7, boundaries included Answer:
  + 𝑃(4 ≤ 𝑋 ≤ 7) = 𝑃(𝑋 = 4) + 𝑃(𝑋 = 5) + 𝑃(𝑋 = 6) + 𝑃(𝑋 = 7) = 18 / 36 = 1 / 2
* We could also compute this using left-tail probabilities only:
  + 𝑃(4 ≤ 𝑋 ≤ 7) = 𝑃(𝑋 ≤ 7) − 𝑃(𝑋 < 4) = 𝑃(𝑋 ≤ 7) − 𝑃(𝑋 ≤ 3) = 21 / 36 − 3 / 36 = 1 / 2

Expectation of a R.V:

* Consider the random variable associated with playing the card game with the following probability distribution function
  + 𝑓𝑋 (3) = 𝑃(𝑋 = 3) = 1 13, 𝑓𝑋 (2) = 𝑃(𝑋 = 2) = 1 13
  + 𝑓𝑋 (1) = 𝑃(𝑋 = 1) = 1 13, 𝑓𝑋 (0) = 𝑃(𝑋 = 0) = 10 13.
* If somebody played this game a very large number of times, what would their average earnings be? I.e. what is the **expected value** of this random variable?

### Expectation and variance of a random variable

Expectation of a random variable:

* The expectation of a random variable is denoted by 𝜇𝑋 or E(𝑋) and is given by
  + 𝜇𝑋 = ∑ 𝑖 𝑥𝑖 𝑃(𝑋 = 𝑥𝑖 ) = ∑ 𝑖 𝑥𝑖 𝑓𝑋 (𝑥𝑖 ).
* 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.

Expectation of a R.V:

* The expected value of the random variable associated with the card game is given by:
  + 𝜇𝑋 = ∑ 𝑖 𝑥𝑖 𝑓𝑋 (𝑥𝑖 )
  + = 0 ⋅ 𝑓𝑋 (0) + 1 ⋅ 𝑓𝑋 (1) + 2 ⋅ 𝑓𝑋 (2) + 3 ⋅ 𝑓𝑋 (3)
  + = 0 ⋅ 10 / 13 + 1 ⋅ 1 / 13 + 2 ⋅ 1 / 13 + 3 ⋅ 1 / 13
  + = 6 / 13

Variance of a R.V:

* The variance of a R.V. is a measure of dispersion that resembles the sample variance very closely.
* Variance of a random variable
* The variance of a random variable is defined by:
  + 𝜎 2 𝑋 = ∑ 𝑖 (𝑥𝑖 − 𝜇𝑋 ) 2 𝑃(𝑋 = 𝑥𝑖 ) = ∑ 𝑖 (𝑥𝑖 − 𝜇𝑋 ) 2 𝑓𝑋 (𝑥𝑖 ).
* Note: the standard deviation is the positive square root of the variance. 𝜎𝑋 = quare root(𝜎𝑋 \*\* 2)

Variance: Intuition:

* Consider two random variables 𝑋 and 𝑌.
* The random variable 𝑋 represents the number of received emails per day for a first office worker.
* This person receives either 48 or 52 emails per day, each with a 50 percent probability.
* The random variable 𝑌 represents the number of received emails of a second office worker.
* This second person receives no emails at all with 50 percent probability and with 50 percent chance receives 100 emails.
  + Determine the probability mass function of 𝑋 and of 𝑌.
  + Compute the expected value of 𝑋 and 𝑌. What do you notice?
  + Who is experiencing a larger variability w.r.t. the number of emails received?
  + Compute the variance of 𝑋 and 𝑌.

### Continuous Random Variables

Continuous Random Variable:

* 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
* Measurements of length, time, … are examples of continuous random variables.
* The leftmost histogram illustrates that most of the 5000 men were between 175cm and 180cm tall.
* From left to right the number of intervals is increased, but the total area of all the bars always equals 1.
* If the number of intervals is increased even more, the intervals would eventually get so small that we no longer have a histogram, but rather a curve (by connecting the ’dots’ at the tops of the very tiny rectangles).
* The heights of people often follow an approximate Normal Distribution. The **Normal Distribution** is a type of **continuous probability distribution**.
* There are also other continuous probability distributions, e.g. the Exponential Distribution and the continuous Uniform Distribution.

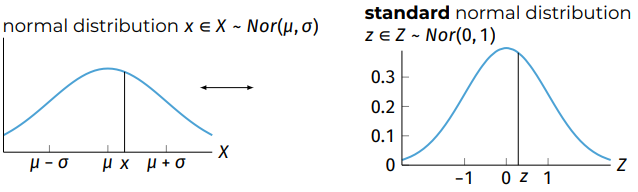
### The (Standard) Normal Distribution

Continuous Probability Distribution:

The reaction speed 𝑥 of Superman (in ms) can also be represented as a normal distribution:

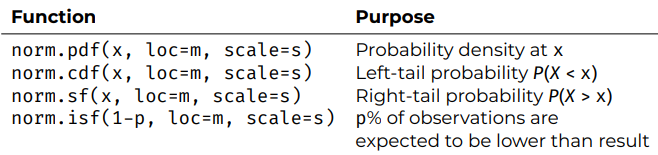
Expectation and Variance:

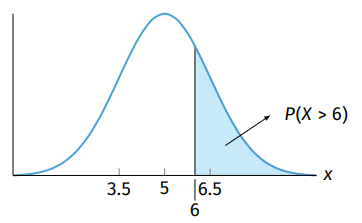
* Also for continuous R.V.’s there are formulas for expectation and variance. The expectation is given by
  + 𝜇𝑋 = ∫ +∞ −∞ 𝑥 𝑓𝑋 (𝑥) d𝑥
* while the variance 𝜎 2 𝑋 is given by 𝜎
  + 2 𝑋 = ∫ +∞ −∞ (𝑥 − 𝜇𝑋 ) 2 𝑓𝑋 (𝑥) d𝑥.
* Note: these are essentially the same formulas as for discrete R.V.’s but with integrals instead of summations.
* Note: you are not expected to know how to calculate integrals!

  
Standard normal distribution:

* 𝑥 and 𝑧 have a similar position on the Gaussian bell curve.
* What is the mathematical relationship between 𝑥 and 𝑧?
  + 𝑥 = 𝜇 + 𝑧.𝜎 and 𝑧 = (𝑥−u) / 𝜎

Python functions:



Calculating Probabilities:

* What is the probability that Superman’s observed reaction speed is over 6 ms?
* Mathematical Notation:
* 𝑃(𝑋 > 6) = ? with 𝑋 ∼ 𝑁𝑜𝑟(𝜇 = 5, 𝜎 = 1, 5)

Calculating Probabilities: 𝑧-table:

* 𝑃(𝑋 > 6) = ? with 𝑋 ∼ 𝑁𝑜𝑟(𝜇 = 5, 𝜎 = 1, 5) (Old) calculation method using z-table, e.g
* Calculate the 𝑧-score 𝑧 = 6−5 1,5 = 0, 667 so 𝑃(𝑋 > 6) = 𝑃(𝑍 > 0, 667)
* 2. Convert to a left tail probability
  + 100% probability rule: 𝑃(𝑍 > 0, 667) = 1 − 𝑃(𝑍 < 0, 667)
  + or symmetry rule: 𝑃(𝑍 > 0, 667) = 𝑃(𝑍 < −0, 667)
* 3. look for the corresponding value in the 𝑧-table
* With python:
  + import scipy.stats as stats
  + stats.norm.sf(6, loc=5, scale=1.5)

Examples:

* What is the probability that Superman reacts in less than 4 ms?
* What is the probability that he reacts in less than 7 ms?
* What is the probability that he reacts in less than 3 ms?
* What is the probability that he reacts between 2 en 6.5 ms?
* What interval contains 80% of his reaction speed?

Exponentional 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
* Another example is the amount of time (in minutes) a postal clerk spends with his or her customers. In the image below, the time is known to have an exponential distribution with an average amount of time equal to four minutes

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. T
* his means the elevator arrival time is uniformly distributed between 10 and 20 seconds once you hit the button.
* 𝑃(13 < 𝑋 < 18) = area under 𝑓𝑋 (𝑥) between 13 & 18= 5 × 0.1 = 0.5
* Note: 𝑃(13 < 𝑋 < 18) = 𝑃(13 < 𝑋 ≤ 18) = 𝑃(13 ≤ 𝑋 < 18) = 𝑃(13 ≤ 𝑋 ≤ 18) and it also holds that
  + 𝑃(10 ≤ 𝑋 ≤ 20) = 1
* so the total area under the probability density function is always equal to 1.

## From Sample to Population

### The Central Limit Theorem

The 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.
* Consider a random sample of 𝑛 observations drawn from a population with expected value 𝜇 and standard deviation 𝜎. If 𝑛 is sufficiently large, the probability distribution of the sample mean 𝑥 will approximate a normal distribution with mean 𝜇𝑥 = 𝜇 and standard deviation 𝜎𝑥 = 𝜎 √𝑛 .
* The larger the sample, the better the probability distribution of 𝑥 will approximate the expected value of the population, 𝜇.

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 Intervals Confidence Interval for a Large Sample

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**

Conf. Int. - Large Sample:

* Given a sample with mean 𝑥.
* We are looking for an interval [ 𝑥 − 𝑏 , 𝑥 + 𝑏 ] for which we can say with a level of confidence (1 − 𝛼) of e.g. 95% that 𝜇 is inside this interval.
  + 𝑃 (𝑥 − 𝑏 < 𝜇 < 𝑥 + 𝑏) = 1 − 𝛼 = 0, 95
* Because of the central limit theorem we know that: 𝑥 ∈ 𝑋 ∼ 𝑁𝑜𝑟(𝜇, 𝜎 √𝑛 )
* And because of the symmetry we can say: 𝑃 (𝑥 − 𝑏 < 𝜇 < 𝑥 + 𝑏) = 𝑃 (𝜇 − 𝑏 < 𝑥 < 𝜇 + 𝑏)

### Confidence Interval for a Small Sample

Conf. Int. - Small Sample:

* For a small sample the central limit theorem is no longer valid.
* Instead we can say:
* If a population 𝑋 has a normal distribution (𝑋 ∼ 𝑁𝑜𝑟(𝜇, 𝜎)) and you take a small sample with mean 𝑥 and standard deviation 𝑠, then 𝑡 = 𝑥 − 𝜇 𝑠 √𝑛
* will behave as a t-distribution with 𝑛 − 1 degrees of freedom

# Hoofdstuk 4: Bivariate analysis: qualitative variables

## Bivariate analysis

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!

Afbeelding met tekst, schermopname, Lettertype, nummer

Door AI gegenereerde inhoud is mogelijk onjuist.Margin Totals:

* We hebben een onderzoek en willen vergelijken of er een groot verband is tussen het antwoord van mannen en vrouwen
* Afbeelding met tekst, schermopname, Lettertype, nummer

  Door AI gegenereerde inhoud is mogelijk onjuist.Om dit verband uit te werken moeten we volgende formule uitwerken en krijgen we volgend resultaat: **(row total × column total) / n**
* Voor de laatste tabel te bereken gebruiken we volgende formule: **((o – e)² / e)**
  + O = observerd value **(rijtotaal)**
  + E = excepcted value **(cel)**
  + Afbeelding met tekst, Lettertype, schermopname, nummer

    Door AI gegenereerde inhoud is mogelijk onjuist.Voor ‘strongly disagree’ is dit: ((4 – 3.168)² / 3.168)

## Goodness-of-fit test

Afbeelding met tekst, schermopname, Lettertype, nummer

Door AI gegenereerde inhoud is mogelijk onjuist.Goodness-of-fit test:

* ‘o’ is de sample van het totaal aantal helden
* ‘pi’ is hier het percenteel aantal van ‘o’ van de volledige populatie
* ‘e’ is hier de volledige populatie \* ‘pi’
* Het resultaat (laatste kolom) is dan wederom de volgende formule ((127 - 140)² / 140)