
# Statistiek 2: Project

2de bachelor Wiskunde

**Prof. dr. Martial Luyts**

UGent, Belgium

martial.luyts@kuleuven.be

# Contents

# Deel 1:

# Doel en organisatie

# 1.1 Doelstellingen

Deze cursus richt zich op zeer diverse **competenties**:

- **technische competenties**;
- **beroepsgerichte competenties**;

# 1.1.1 Technische competenties

- **Data analyse:**
  Vertalen van wetenschappelijk naar statistisch probleem
  en een gepaste oplossing aanreiken.

- **Toepassingsgericht onderzoek:**
  Analytisch en/of via computersimulatie,
  inzicht ontwikkelen in de meest gepaste analyse.

# 1.1.2 Beroepsgerichte competenties

- Efficiënt werken in groepsverband.
- Schriftelijk en mondeling rapporteren.
- Gebruik van professionele statistische software: R.
- Zelfstandig leren.

Rechtstreekse voorbereiding op bachelor- en masterproef, ...,
en de arbeidsmarkt.

# 1.2 Aanpak en organisatie

**Hoe zullen we dat realiseren?**

- Reeds opgedane statistische kennis versterken;
- Kennis data analyse verrijken;
- Statistische software leren gebruiken;
- Leren rapporteren.

# 1.2.1 Overzicht

Dit zal gebeuren via

- een beperkt aantal **hoorcolleges**
  - ▷ Rapporteren;
  - ▷ Monte-Carlo simulatie;
  - ▷ Toetsen van hypotheses;
- maar vooral via **hands-on projectwerk**.

# 1.2.2 Projecten

- **Project 1 (per 4):** Presentatie van een nieuwe analyse techniek
  (mondeling 1u, slides: $2+1$, verdediging: 2)
- **Project 2 (per 4):** Simulatiestudie of data analyse
  (protocol + schriftelijk + mondeling,
  rapport: $2$, verdediging: 2)
- **Project 3 (per 4):** Examenproject
  (schriftelijk, rapport: 2, verdediging: 3)

Verdere taken:

- Zelfstandig literatuur doornemen voorafgaand aan de les.
- 2 maal (per 2): Rapporteren van PC-lab resultaten.

# 1.2.3 Feedback

- Project 1: Na eerste versie van slides
- Project 2: Na eerste versie van slides en rapport
- Op basis van rapportering van PC-lab resultaten.
- Verdere ruimte voor feedback wordt voorzien
  tijdens/na bespreking Protocol en hulp tijdens PC-labs.
- Tot slot is er ook de mogelijkheid tot individueel contact.

# 1.2.4 Examen

**Individueel:**

- Schriftelijk openboek examen: 6
- Verdediging project: 2

# 1.2.5 Organisatie

- Groepsindeling + Ufora.
- Gebruik van LaTeX verplicht voor slides en rapporten.
- Slides worden vooraf beschikbaar gesteld via Ufora:
  - ▷ Tussentijds submitten via email.
  - ▷ Finaal submitten via email.
  - ▷ Gebruik informatieve titels bij submissie!
    bvb. Project 1 - Analysis of Variance
- Vragen i.v.m.
  - ▷ Project 1, 3: Titularis.
  - ▷ Project 2, R-code project 1 en PC-labs: Assistent.
- Actieve participatie wordt sterk aangemoedigd
  en enkel positief geëvalueerd.
- Taakverdeling wordt sterk aangemoedigd.

# Deel 2:

# Wetenschappelijk rapporteren

# 2.1 Nut

Rapportering is ontzettend belangrijk in het **beroepsleven**:

- Resultaten zijn nutteloos tenzij anderen duidelijk begrijpen wat en waarom je dit deed, en wat dit impliceert!
- De kwaliteit van de rapportering is vaak bepalend voor een wetenschappelijke doorbraak.

# 2.2 Schriftelijk rapporteren

## Hoe schriftelijk rapporteren?

- Vermeld het **doel**.
  *Welke vragen probeer je te beantwoorden?*
- Vermeld de **reden** voor de gekozen aanpak.
- Review alle **methoden** die je bestudeert of toepast.
  *Wees accuraat en voldoende gedetailleerd.*
- Beschrijf hoe je de onderzoeksvraag beantwoord hebt.
  *Geef voldoende detail opdat de resultaten*
  **exact reproduceerbaar zijn!**
  *Vermijd dat de lezer daartoe*
  *in de softwarecode moet duiken.*

## Voorbeeld:

### 6. Simulation study

#### *6.1. Design*

In each simulation scenario, we simulate 1,000 datasets, each of size $n$, following a data-generating mechanism that slightly generalizes that in Stürmer et al. [22] by evaluating the impact of model misspecification. The simulation is done in two stages. In the first stage, 6 independent confounders $C_1$–$C_6$ are simulated: $C_1$–$C_3$ are Bernoulli random variables each with mean 0.2 and $C_4$–$C_6$ are each standard normal. The so-called 'intended' treatment variable, $T$, is then simulated from a Bernoulli distribution with mean

$$P(T = 1 \,|C_1, \ldots, C_6) = \text{expit}(\alpha_0 + \alpha_1 C_1 + \cdots + \alpha_6 C_6 + \alpha_{16} C_1 C_6).$$

Then, in the second stage, two further binary covariates $C_7$ and $C_8$ are generated. $C_7$ is most likely to be present $(= 1)$ when the intended treatment $T$ is least likely; it is set to 1 with probability $\{\gamma - P(T = 1 \,|C_1, \ldots, C_6)\}$. $C_8$ is likely to be present $(= 1)$ when the intended treatment $T$ is most likely; it is set to 1 with probability $\{P(T = 1 \,|C_1, \ldots, C_6) - \delta\}$. The values of $\gamma$ and $\delta$ are chosen such that the prevalence of $C_7$ and $C_8$ is close to 0.1. The actual treatment $A$ is then simulated from a Bernoulli distribution with mean

$$P(A = 1 \,|C_1, \ldots, C_8) = \text{expit}(\alpha_0 + \alpha_1 C_1 \cdots + \alpha_8 C_8 + \alpha_{16} C_1 C_6).$$

# 2.4 Resultaten overbrengen

Resultaten moeten gebracht worden in een vorm die

- duidelijk de vragen beantwoordt;
- makkelijk toelaat de voornaamste conclusies te begrijpen.

**Basisprincipes:**

- Presenteer enkel een **subset** van de resultaten.
- Presenteer enkel **interessante** resultaten.
- De **wijze van presentatie** moet **toegankelijk** zijn.
- **Interpreteer de resultaten!!**

# 2.5 Criteria van een goede schriftelijke rapportering

1. Correct
2. Beknopt:
   - Lezers hebben beperkte tijd;
   - Beknopte rapporten zijn vaak duidelijker.
3. Duidelijk: Breng een gestructureerd verhaal
4. Inzichtelijk: Interpreteer en verklaar de resultaten.

Om dit te realiseren:

- Stel uzelf in de plaats van uw doelpubliek!
- Lees, en herlees...

# 2.5.1 Voorbeelden

## Voorbeeld 1: Omslachtige rapportering

**De nulhypothese ($H_0$)**

We stellen als nulhypothese dat laag geboortegewicht onafhankelijk is van het gewicht van de moeder (bij haar laatste menstruatieperiode).

**De alternatieve hypothese ($H_A$)**

We stellen als alternatieve hypothese dat laag geboortegewicht afhankelijk is van het gewicht van de moeder (bij haar laatste menstruatieperiode).

**Test**

We controleren of we $H_0$ kunnen verwerpen op het 5% significantieniveau aan de hand van een ongepaarde t-test (tweezijdig). We vergelijken dus gewicht van de moeder (die een baby heeft met laag geboortegewicht) met gewicht van de moeder (die een baby heeft met normaal geboortegewicht).

**De teststatistiek**

- t-verdeling met 187 vrijheidsgraden
- We bekomen de waarde -2.3537 voor de teststatistiek

**P-waarde en betrouwbaarheidsinterval**

- We bekomen als P-waarde 0.01962
- We bekomen als 95% betrouwbaarheidsinterval (afgerond op twee cijfers na de komma) [-20.52,-1.81]

**Conclusie**

We kunnen dus $H_0$ op het 5% significantieniveau verwerpen en besluiten dat het krijgen van een baby met laag geboortegewicht afhankelijk is van het gewicht van de moeder.

- Dit kan **bondiger** en **informatiever**:
  'Via de ongepaarde t-test detecteren we een verschil in gemiddeld gewicht (in pond) vóór zwangerschap tussen moeders van baby's met versus zonder laag geboortegewicht (gemiddeld verschil: -9.3, 95% BI: -20.5 tot -1.8, P = 0.02).'

## Voorbeeld 2: Omslachtige rapportering

Als eerste kenmerk willen we de verdeling van het geslacht onderzoeken. Hiervoor voerden we een chi-kwadraat test uit. We bekomen een p-waarde van 0.25.
Vervolgens bekijken we de distributie van de rassen. Er wordt onderscheid gemaakt in 'white non-hispanic', 'black non-hispanic', 'hispanic', 'Asian or Pacific Islander', 'American, Indian or Alaskan native' en anderen. Ook hier gebruiken we een chi-kwadraat test. Deze geeft ons een p-waarde van 0.65.
De derde factor is de aanwezigheid van hemofilie. We voeren opnieuw een chi-kwadraat test uit met als resultaat een p-waarde van 0.22.
Ten vierde bekijken we in welke mate symptomen zich voordoen, gebaseerd op de 'Karnofsky Performance Scale'. Er worden vier mogelijke waarden gerapporteerd: 100, 90, 80 en 70, waarbij geldt: hoe lager de score, hoe ernstiger de symptomen. De chi-kwadraat test snelt ons weer ter hulp en bezorgt ons een p-waarde van 0.86.

## Voorbeeld 3: Uit wetenschappelijke literatuur

**Statistical considerations**

The primary outcome measure was headache score at the one year follow up. Secondary outcome measures included headache score at three months, days with headache, use of medication scored with the medication quantification scale (MQS),[12 13] the SF-36, use of resources, and days off usual activities. We revised the statistical plan to employ adjusted rather than unadjusted analyses after publication of the initial protocol but before we conducted any analyses. We analysed our data on Stata 8 software (Stata Corporation, College Station, Texas) using ANCOVA for continuous end points, $\chi^2$ for binary data, and negative binomial regression for count data such as number of days of sick leave. We entered minimisation variables into regression models as covariates. We analysed data according to allocation, regardless of the treatment received. We conducted sensitivity analyses to examine the possible effect of missing data (see appendix on bmj.com).

**Results** Headache score at 12 months, the primary end point, was lower in the acupuncture group (16.2, SD 13.7, n = 161, 34% reduction from baseline) than in controls (22.3, SD 17.0, n = 140, 16% reduction from baseline). The adjusted difference between means is 4.6 (95% confidence interval 2.2 to 7.0; P = 0.0002). This result is robust to sensitivity analysis incorporating imputation for missing data. Patients in the acupuncture group experienced the equivalent of 22 fewer days of headache per year (8 to 38). SF-36 data favoured acupuncture, although differences reached significance only for physical role functioning, energy, and change in health. Compared with controls, patients randomised to acupuncture used 15% less medication (P = 0.02), made 25% fewer visits to general practitioners (P = 0.10), and took 15% fewer days off sick (P = 0.2).

# Voorbeeld 4: Niet inzichtelijke tabel

| | LOW | |
|---|---|---|
| LWT | Nee | Ja |
| 80 | 0 | 1 |
| 85 | 1 | 1 |
| 89 | 0 | 1 |
| 90 | 3 | 0 |
| 91 | 0 | 1 |
| 92 | 0 | 1 |
| 94 | 0 | 1 |
| 95 | 5 | 1 |
| 96 | 0 | 1 |
| 97 | 0 | 1 |
| 98 | 1 | 0 |
| 100 | 3 | 2 |
| 101 | 0 | 1 |
| 102 | 0 | 2 |
| 103 | 2 | 1 |
| 105 | 2 | 5 |
| 107 | 2 | 0 |
| 108 | 1 | 0 |
| 109 | 1 | 1 |
| 110 | 7 | 4 |
| 112 | 3 | 1 |
| 113 | 3 | 0 |
| 115 | 5 | 2 |
| 116 | 1 | 0 |
| 117 | 1 | 1 |

| | LOW | |
|---|---|---|
| LWT | Nee | Ja |
| 118 | 2 | 0 |
| 119 | 3 | 0 |
| 120 | 12 | 5 |
| 121 | 3 | 1 |
| 122 | 1 | 1 |
| 123 | 3 | 0 |
| 124 | 2 | 0 |
| 125 | 2 | 1 |
| 127 | 1 | 0 |
| 128 | 1 | 1 |
| 129 | 1 | 0 |
| 130 | 6 | 7 |
| 131 | 1 | 0 |
| 132 | 2 | 1 |
| 133 | 2 | 0 |
| 134 | 3 | 0 |
| 135 | 4 | 0 |
| 137 | 1 | 0 |
| 138 | 1 | 1 |
| 140 | 3 | 0 |
| 141 | 1 | 0 |
| 142 | 0 | 2 |
| 147 | 2 | 0 |
| 148 | 0 | 1 |
| 150 | 3 | 2 |

| | LOW | |
|---|---|---|
| LWT | Nee | Ja |
| 153 | 1 | 0 |
| 154 | 1 | 1 |
| 155 | 2 | 1 |
| 158 | 2 | 0 |
| 160 | 2 | 0 |
| 165 | 0 | 1 |
| 167 | 1 | 0 |
| 168 | 1 | 0 |
| 169 | 2 | 0 |
| 170 | 4 | 0 |
| 175 | 1 | 0 |
| 182 | 1 | 0 |
| 184 | 1 | 0 |
| 185 | 1 | 0 |
| 186 | 1 | 0 |
| 187 | 0 | 2 |
| 189 | 1 | 0 |
| 190 | 1 | 1 |
| 200 | 0 | 1 |
| 202 | 1 | 0 |
| 215 | 1 | 0 |
| 229 | 1 | 0 |
| 235 | 1 | 0 |
| 241 | 1 | 0 |
| 250 | 1 | 0 |

Figuur 2: kruistabel: laag geboortegewicht tegenover gewicht bij de laatste menstruatie

# Voorbeeld 5: Onduidelijke (foutieve?) tabel

| Variabele | Groep 1 (n=535) | Groep 2 (n=534) |
|---|---|---|
| **Geslacht** | | |
| Man | 448 (41.9) | 433 (8.1) |
| Vrouw | 87 (40.5) | 101 (9.5) |
| **Etnische afkomst** | | |
| White Non-Hispanic | 272 (25.4) | 279 (26.1) |
| Black Non-Hispanic | 150 (14.0) | 150 (14.0) |
| Hispanic | 101 (9.4) | 94 (8.8) |
| Asian, Pacific Islander | 9 (0.8) | 5 (0.5) |
| American Indian, Alaskan Native | 3 (0.3) | 6 (0.6) |
| **IV Druggebruik** | | |
| Nooit | 451 (42.2) | 451 (42.2) |
| Momenteel | 1 (0.1) | 3 (0.3) |
| Vroeger | 83 (7.7) | 80 (7.5) |
| **Hemofilie** | | |
| Ja | 20 (1.9) | 13 (1.2) |
| Nee | 515 (48.2) | 521 (48.7) |
| **Karnofsky score** | | |
| 100 | 185 (17.3) | 178 (16.7) |
| 90 | 245 (22.9) | 257 (24.1) |
| 80 | 90 (8.4) | 83 (7.7) |
| 70 | 15 (1.4) | 16 (1.5) |

Tabel 1: Waarden zijn aantallen (percenten) van patiënten.

# Voorbeeld 6: Beknopte en duidelijke tabel

**Table 2. Primary and Secondary End-Point Events.***

| End Point | Placebo (N=7933) | | Dalcetrapib (N=7938) | | Hazard Ratio with Dalcetrapib (95% CI) | P Value |
|---|---|---|---|---|---|---|
| | Patients with Event | Event Rate at 3 Yr | Patients with Event | Event Rate at 3 Yr | | |
| | *no. (%)* | *% (95% CI)* | *no. (%)* | *% (95% CI)* | | |
| Primary end point | 633 (8.0) | 9.1 (8.4–9.9) | 656 (8.3) | 9.2 (8.5–9.9) | 1.04 (0.93–1.16) | 0.52 |
| Death from coronary heart disease | 125 (1.6) | 1.8 (1.5–2.2) | 118 (1.5) | 1.6 (1.3–1.9) | 0.94 (0.73–1.21) | 0.66 |
| Nonfatal acute myocardial infarction | 407 (5.1) | 6.0 (5.4–6.7) | 414 (5.2) | 5.9 (5.3–6.5) | 1.02 (0.89–1.17) | 0.80 |
| Hospitalization for unstable angina with objective evidence of acute myocardial ischemia | 92 (1.2) | 1.3 (1.0–1.5) | 84 (1.1) | 1.3 (1.0–1.6) | 0.91 (0.68–1.22) | 0.54 |
| Cardiac arrest with resuscitation | 10 (0.1) | 0.1 (0.0–0.2) | 14 (0.2) | 0.2 (0.1–0.3) | 1.41 (0.63–3.18) | 0.40 |
| Stroke of presumed atherothrombotic cause | 73 (0.9) | 1.0 (0.8–1.2) | 91 (1.1) | 1.4 (1.1–1.7) | 1.25 (0.92–1.70) | 0.16 |
| Death from any cause | 229 (2.9) | 3.4 (2.9–3.9) | 226 (2.8) | 3.1 (2.7–3.6) | 0.99 (0.82–1.19) | 0.90 |
| Unanticipated coronary revascularization procedure† | 672 (8.5) | 9.6 (8.9–10.3) | 674 (8.5) | 9.5 (8.8–10.3) | 1.00 (0.90–1.11) | 0.97 |

* The primary efficacy end point was a composite of death from coronary heart disease, major nonfatal coronary events (acute myocardial infarction, hospitalization for unstable angina with objective evidence of acute myocardial ischemia, or cardiac arrest with resuscitation), or stroke of presumed atherothrombotic cause. Secondary efficacy end-point events included each component of the primary composite end point, unanticipated coronary revascularization (not including revascularization for restenosis at the previous intervention site), and death from any cause. Event rates are Kaplan–Meier estimates through 36 months.

† Data are for procedures other than those for restenosis at the previous intervention site.

- Beschrijf naast de tabel, de methoden en onderstellingen en interpreteer de belangrijkste resultaten;
- Nagaan van de onderstellingen kan in Appendix..

## Tabellen: Enkele waarschuwingen

- Structureer de tabel op een manier die vergelijking toelaat.
- Vermijd een overdaad aan getallen! Rond af!

## (Wainer, 1993) Drie redenen om af te ronden

Meer dan 2 cijfers zijn

- moeilijk op te pikken.
- nagenoeg nooit statistisch te rechtvaardigen.
- meestal onvoldoende interessant.
- **Voorbeelden:**
  - ▷ "This year's school budget is $27,329,681.32"
    of "This year's school budget is about 27 million dollars".
  - ▷ "Mean life expectancy of Australian males is 67.14 years"
    of "Mean life expectancy of Australian males is 67 years".

## Statistische nauwkeurigheid

- Beschouw het percentage mannen in een studie.
- Om 0.56 (2 cijfers) te rapporteren, mag de standaard error niet groter zijn dan $0.005$. zó kunnen we het verschil maken tussen 0.56 en 0.57 of 0.58 ($1.96 \times 0.005 \approx 0.01$).

- **Voorbeeld:**

## Slechte tabel

| | Sample mean | | Trimmed mean | | Median | |
|---|---|---|---|---|---|---|
| | Normal | $t_5$ | Normal | $t_5$ | Normal | $t_5$ |
| Mean | 0.98515 | 0.98304 | 0.98690 | 0.98499 | 0.99173 | 0.98474 |
| Bias | −0.01485 | −0.01696 | −0.01310 | -0.01501 | -0.00827 | -0.01526 |
| SD | 0.33088 | 0.33067 | 0.34800 | 0.31198 | 0.39763 | 0.35016 |
| MSE | 0.10959 | 0.10952 | 0.12116 | 0.09746 | 0.15802 | 0.12273 |
| Rel. Eff. | 1.00000 | 1.00000 | 0.90456 | 1.12370 | 0.69356 | 0.89238 |

## Goede tabel

| | Normal | | | $t_5$ | | |
|---|---|---|---|---|---|---|
| | Samp mean | Trim mean | Median | Samp mean | Trim mean | Median |
| Mean | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 0.98 |
| Bias | −0.01 | −0.01 | −0.01 | −0.02 | −0.02 | −0.02 |
| SD | 0.33 | 0.35 | 0.40 | 0.33 | 0.31 | 0.35 |
| MSE | 0.11 | 0.12 | 0.16 | 0.11 | 0.10 | 0.12 |
| Rel. Eff. | 1.00 | 0.90 | 0.69 | 1.00 | 1.12 | 0.89 |

Voor mondelinge presentatie is dit mogelijk nog teveel detail.

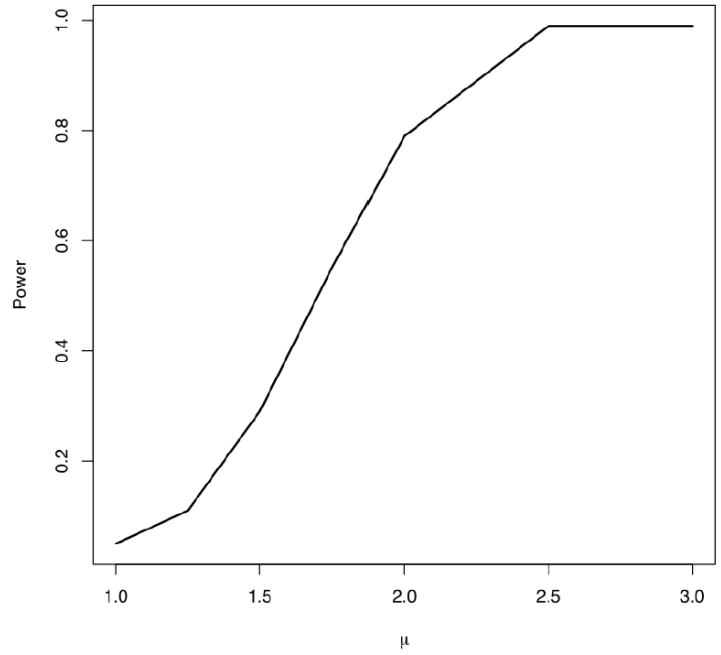
## Grafieken zijn vaak effectiever dan tabellen

**Voorbeeld:** Power van de $t$-test voor $H_0 : \mu = 1.0$ versus $H_1 : \mu \neq 1.0$ voor normaal verdeelde data $(n = 15)$.

- **Tabel**

| $\mu$ | 1.0 | 1.25 | 1.50 | 1.75 | 2.00 | 2.50 | 3.00 |
|---|---|---|---|---|---|---|---|
| power | 0.05 | 0.11 | 0.29 | 0.55 | 0.79 | 0.99 | 0.99 |

- **Grafiek**



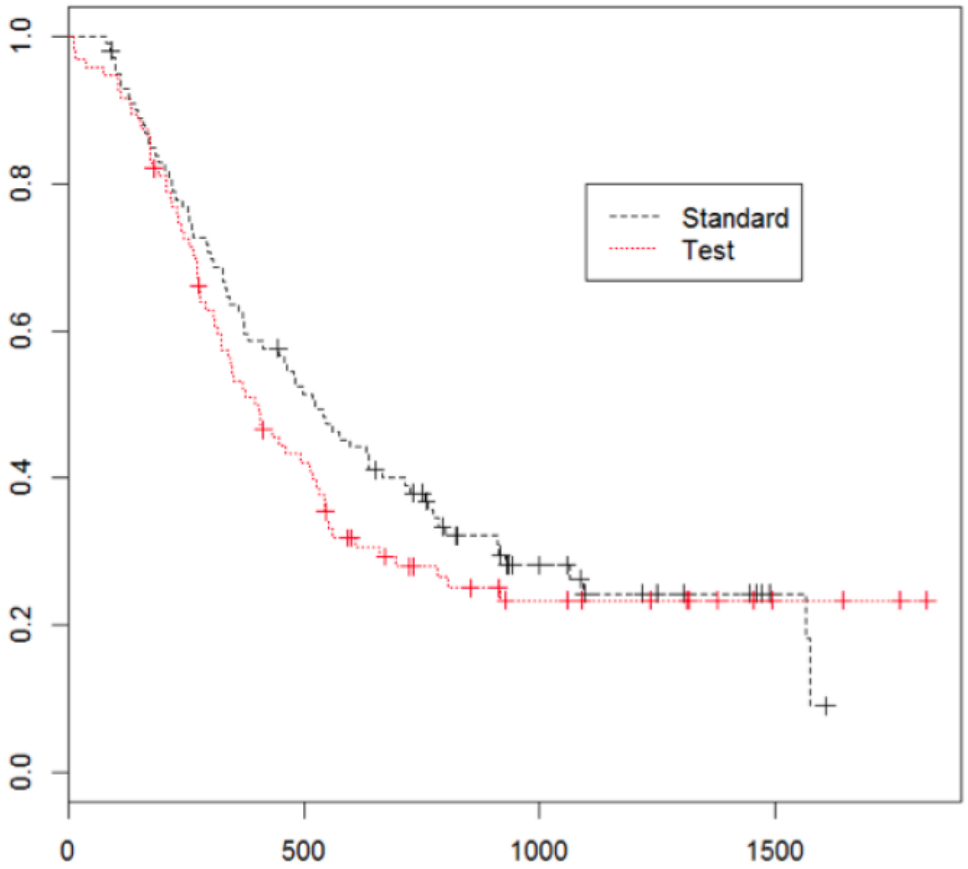
## Definiëring notatie

$H_0$: het aantal partners heeft geen invloed op de levensduur; alle $\mu_i$ gelijk.

$H_1$: het aantal partners heeft wel een invloed op de levensduur; niet alle $\mu_i$ gelijk.

Vermijd symbolen, tenzij ze informatief zijn en uitgelegd worden!

# Assen labelen



## Voorbeeld 7: Overbodige/onduidelijke rapportering

- 'Het behandelingseffect heeft een 95% BI van [1.94,3.75] en 0 is hier niet in bevat.'
- 'De p-waarde is te hoog om de nulhypothese te verwerpen. Bovendien bevat het 95% BI de waarde nul...'

## Voorbeeld 8: Tot slot, geef inzicht en interpretatie

- **Voorbeeld 1:**

  **Minder goed...**

  Treatment group was significant (t = 3.82, df = 38, p = 0.001).

  **Veel beter...**

  Mean diastolic blood pressure were significantly lower in the treatment group (n = 20) than in the controls (n = 20); the t-test comparing the treatment mean of 88.0 (s.d. 7.3) with the control mean of 98.5 (s.d. 9.8) was statistically significant (p = 0.001). A 95% confidence interval for the treatment effect is (4.9 to 16.0); even a 5 mmHg drop in diastolic blood pressure is considered to be clinically significant among this patient population.

- **Voorbeeld 2:**

  **Minder goed...**

  The data were analyzed using t-tests and Wilcoxon tests.

  **Veel beter...**

  Mean diastolic blood pressures in the treatment and control groups were compared using a t-test. Because the data contained extreme values, t-tests were supplemented by a rank-based method, the Wilcoxon test.

# 2.5.2 Referentiebronnen

**Referentiebronnen** van diverse kwaliteit:

- **Wetenschappelijke tekstboeken** en **artikels**
  - ▷ Web of Knowledge: `webofknowledge.com`
  - ▷ Google scholar: `scholar.google.be`
- **Wikipedia**
- ...

## Refereren = Respecteren

- (artikel) Berk, K. N. (1978). Comparing Subset Regression Procedures. *Technometrics*, **20**, 1-6.
- (boek) Dixon, W. J. (ed.) (1983). *BMDP Statistical Software* (Vol. 1, 3rd ed.). Berkeley, CA: University of California Press.
- (technisch rapport) Hogg, R. V., Smith, J., Jones, L., and Smith, S. (1973). A New Sample Adaptive Distribution-Free Test. Technical Report 24, University of Iowa, Dept. of Statistics.
- (hoofdstuk) McQueen, M. Y. (in press). *Kruskal's Proof Refuted*, in The Theorems and Proofs of Kruskal, eds. J. Doe and B. Doe, Chicago: Rand McNally.
- (web) Bilodeau, A. (1994). Into the Net: A Reporter's Transformation. *Computer-Mediated Communication Magazine [online]*, **1**, 8. Available at http://www.rpi.edu/decemj/cmc/mag/archive.html.

# 2.6 Criteria van een goede mondelinge rapportering

Dezelfde criteria gelden hier:

1. Correct: Maar met minder detail
2. Beknopt: Overdadige slides werken storend
3. Duidelijk: Breng structuur aan en werk met voorbeelden
4. Inzichtelijk: Interpreteer en verklaar de resultaten.

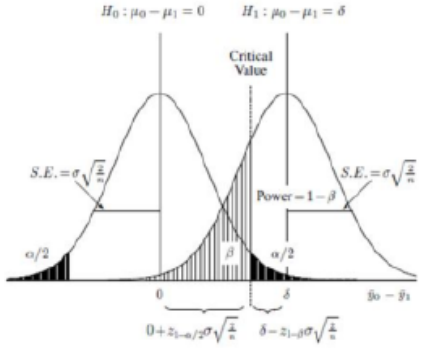
Om dit te realiseren:

- Stel uzelf in de plaats van uw doelpubliek!
- Oefen uw presentatie!

# Voorbeeld 1: Onvoldoende correct

Sample Size Calculation

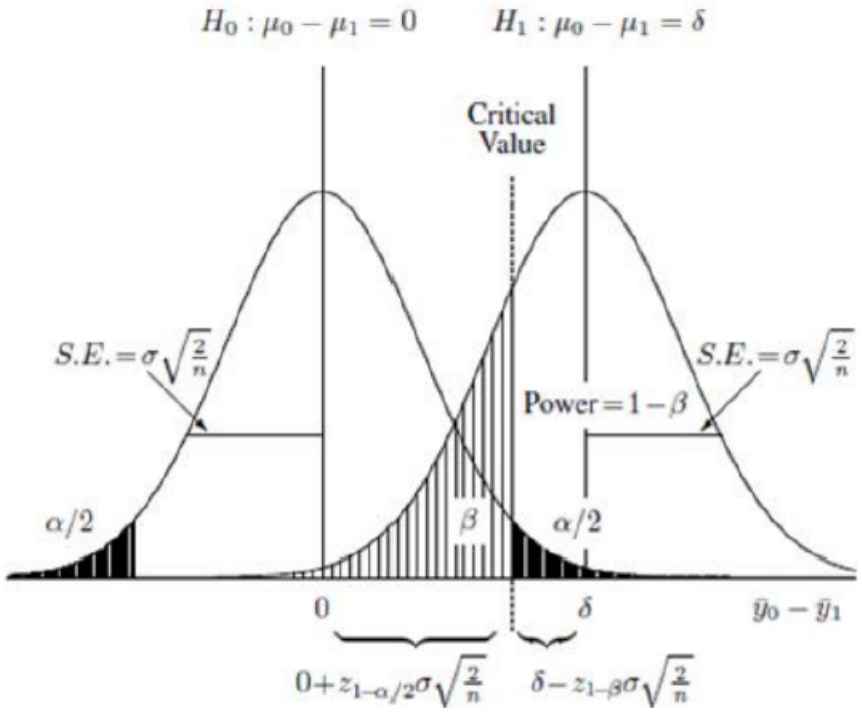
## Formule voor gepaarde t-test




$\alpha$ = kans op een typel-fout
$\beta$ = kans op een typell-fout
$\sigma$ = variantie
$\mu_0$ = gemiddelde onder de nulhypothese
$\mu_1$ = gemiddelde onder de alternatieve hypothese

## Sample Size Calculation

$$0 + z_{1-\alpha/2}\sigma\sqrt{\frac{2}{n}} = \delta - z_{1-\beta}\sigma\sqrt{\frac{2}{n}} \Leftrightarrow n = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{(\frac{\mu_0-\mu_1}{\sigma})^2}$$



## Voorbeeld 1

2 groepen met elk een verschillend dieet

- verschil in gemiddelde: 10
- standaardafwijking: 16
- power $= 1 - \beta$: 0.8
- $\alpha$: 0.05

$$n = \frac{2(z_{0.975} + z_{0.8})^2}{(\frac{10}{16})^2} \approx 41$$

## Voorbeeld 2: Overdadige slide

- Fresh fruit leads Chile's export mix - Chile emerges as major supplier of fresh fruit to world market due to ample natural resources, consumer demand for fresh fruit during winter season in U.S. and Europe, and incentives in agricultural policies of Chilean government, encouraging trend toward diversification of exports and development of nontraditional crops - U.S. Dept. of Agriculture, Economic Research Service Report
- Chile is among the developing economies taking advantage of these trends, pursuing a free market economy. This has allowed for diversification through the expansion of fruit production for export, especially to the U.S. and Western Europe. Chile has successfully diversified its agricultural sector to the extent that it is now a major fruit exporting nation. Many countries view Chile's diversification of agriculture as a model to be followed.

## Voorbeeld 3: Teveel detail, geen voorbeelden

Stochastische vectoren

- $X : \Omega \to \mathbb{R}^n$ is $\mathcal{F}$-meetbaar a.s.a. elke projectie $X_i : \Omega \to \mathbb{R}$ is $\mathcal{F}$-meetbaar
- bijgevolg
  - ▷ stochastische vector over $(\Omega, \mathcal{F}, \mathcal{P})$ is vector van stochastische veranderlijken over $(\Omega, \mathcal{F}, \mathcal{P})$
  - ▷ elk van die veranderlijken functie van dezelfde toestand $\omega \in \Omega$

## Gezamenlijke verdelingsfunctie

- $\forall B \in \mathcal{B}^n(\mathbb{R}) = \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R})$ is

$$X^{-1}(B) \in \mathcal{F}$$

- omdat het halfopen interval

$$]-\infty, x_1] \times ]-\infty, x_2] \times \ldots \times ]-\infty, x_n] \in \mathcal{B}^n(\mathbb{R})$$

is voor een stochastische vector $X = (X_1, ..., X_n)$

$$\begin{aligned} &\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_n(\omega) \leq x_n\} \\ &= \bigcap_{i=1}^{n} \{\omega : X_i(\omega) \leq x_i\} \in \mathcal{F} \end{aligned}$$

## Gezamenlijke verdelingsfunctie

- bijgevolg is

$$\mathcal{P}\left(\bigcap_{i=1}^{n}\{\omega : X_i(\omega) \leq x_i\}\right) = \mathcal{P}\left(X_1 \leq x_1, X_2 \leq X_2, \ldots, X_n \leq x_n\right)$$
$$\equiv F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

  goed gedefinieerd

- $$F_{X_1,X_2,\ldots,X_n} : \mathbb{R}^n \to [0,1]$$

  wordt verdelingsfunctie van de stochastische vector $(X_1, X_2, \ldots, X_n)$ genoemd

## Voorbeeld 4: Betere opbouw

Vectoren van s.v. zijn stochastische vectoren.

stochastische vector over $(\Omega, \mathcal{F}, \mathcal{P})$

$\mathcal{F}$-meetbare afbeelding $X : \Omega \to \mathbb{R}^n$

Voorbeelden:

- Metingen bij verschillende individuen:
  bvb. virale lading $X_i$ bij 100 HIV-patiënten.
- Verschillende metingen per individu:
  thrombocieten $X_i$, hematocriet $Y_i$, leeftijd $Z_i$, ...

## Voorbeeld 5: Breng structuur aan en maak het beknopt

### Voorbeelden:

- Metingen bij verschillende individuen: bvb. virale lading $X_i$ bij 100 HIV-patiënten $i = 1, ..., 100$ na 1 jaar op HAART.
- Verschillende metingen per individu: bvb. hoeveelheid thrombocieten $X_i$, hematocriet $Y_i$, C-reactief proteïne $Z_i$, ... bij 100 IZ-patiënten $i = 1, ..., 100$

**versus**

### Voorbeelden:

- Metingen bij verschillende individuen:
  bvb. virale lading $X_i$ bij 100 HIV-patiënten.
- Verschillende metingen per individu:
  bvb. thrombocieten $X_i$, hematocriet $Y_i$, leeftijd $Z_i$, ...

## Voorbeeld 6: Onduidelijk

- $M_0$ verdeelt populatie in 2:
  teststatistiek $N_-$ en $N_+$,
  geobserveerde waarden: $n_-$ en $n_+$
- Onder nulhypothese verwachten we dat:

$$n_+ \approx n_- \approx \frac{n}{2}$$

- Links-eenzijdige test: verwerpen indien $n_+$ te groot
  Tweezijdige test: verwerpen indien $n_+$ te groot of te klein

Vermijd overvloed aan notatie tijdens presentaties.

# 2.7 Slotwoord

- Via dit opleidingsonderdeel wensen we u beter te wapenen
  - ▷ door u voeling te geven met het statistische redeneerproces
  - ▷ en een aantal nuttige competenties mee te geven voor de grote en sterk groeiende arbeidsmarkt rond data science.
- Hopelijk wordt het een boeiende ervaring!
- Feedback wordt sterk geapprecieerd!

# Deel 3:

# Monte Carlo simulatie

# 3.1 Wat is een simulatie studie?

- **Simulatie:** Een numerieke techniek om experimenten uit te voeren op een computer.
- **Monte Carlo simulatie:** Computer experiment gebruiken makende van willekeurige steekproeven die afkomstig zijn vanuit kansverdelingen.
  - ▷ Onschatbaar in statistiek.
  - ▷ Vaak, wanneer statistici praten over 'simulaties', bedoelen ze 'Monte Carlo simulaties'.

## Grondgedachte:

- Eigenschappen van statistische methoden moeten zo worden vastgesteld zodat deze methoden met vertrouwen gebruikt kunnen worden.
- Exacte analytische afleidingen van eigenschappen zijn **zelden** mogelijk.
- Grote steekproef benaderingen van eigenschappen zijn vaak mogelijk, echter...
  - ▷ evaluatie van de relevantie van de benadering voor (eindige) steekproefomvang die in de praktijk te verwachten is, is nodig.
  - ▷ analytische resultaten vereisen mogelijk **aannames** (bijvoorbeeld normaliteit).
  - ▷ maar wat gebeurt er als deze veronderstellingen worden geschonden? Analytisch resultaten, zelfs afkomstig vanuit grote steekproeven, zijn misschien niet mogelijk

## Vaak optredende problemen:

Onder verschillende voorwaarden:

- Is een schatter **biased** in eindige steekproeven?
  Is het nog **consistent** onder afwijkingen van aannames?
  Wat is de **steekproefvariantie**?
- Hoe **vergelijkt** dit met concurrerende schatters
  op basis van van bias, precisie, enz.?
- Gaat een procedure voor het construeren van een **betrouwbaarheidsinterval** voor een parameter het vooropgestelde **nominale dekkingsniveau** bereiken?
- Behaalt een **hypothese toets** het vooropgestelde **significantie niveau** of **grootte**?
- Zo ja, welke **power** is mogelijk tegen verschillende alternatieven voor de nulhypothese?
  Leveren verschillende testprocedures een ander power?

## Monte Carlo simulatie to the rescue

- Al deze eigenschappen kunnen worden afgeleid als we de **echte steekproef verdeling** van de schatter onder een bepaalde reeks voorwaarden (eindige steekproef grootte, ware verdeling van de gegevens, enz.) kennen.
- Deze verdeling geeft weer hoe de schatter varieert
  over herhaalde steekproeftrekkingen uit dezelfde populatie
- Wanneer de afleiding van de ware steekproef verdeling niet traceerbaar is, kan men Monte Carlo simulatie gebruiken om het **te benaderen** door experimenten uit het echte leven na te bootsen.

# 3.2 Hoe worden simulatie-experimenten uitgevoerd?

Een typische Monte Carlo simulatie houdt het volgende in:

- Genereer $S$ onafhankelijke datasets
  onder de voorwaarden van interesse.
- Bereken de numerieke waarde van de schatter/teststatistiek $T(\text{data})$ voor elke dataset $\Rightarrow T_1, \ldots, T_S$.
- Wanneer $S$ groot genoeg is, zullen **samenvattingsstatistieken** over $T_1, \ldots, T_S$ goede **benaderingen** moeten zijn voor de ware steekproef eigenschappen van de schatter/teststatistiek onder de voorwaarden van belang.

## Simpel voorbeeld:

Vergelijk drie schatters voor het **gemiddelde** $\mu$ van een verdeling op basis van i.i.d. trekkingen $Y_1, \ldots, Y_n$

- Steekproef gemiddelde $T^{(1)}$
- 20% getrimd steekproef gemiddelde $T^{(2)}$
- Steekproef mediaan $T^{(3)}$

## Opmerking:

- Als de verdeling van de gegevens **symmetrisch** is, schatten alle drie schatters inderdaad het gemiddelde
- Als de verdeling scheef is, doen ze dat niet

**Procedure:**

Voor een bepaalde keuze van $\mu$, $n$, en ware onderliggende verdeling

- Genereer onafhankelijke trekkingen $Y_1, \ldots, Y_n$ uit de verdeling.
- Bereken $T^{(1)}$, $T^{(2)}$, $T^{(3)}$.
- Herhaal $S$ keren $\Rightarrow$

$$T_1^{(1)}, \ldots, T_S^{(1)}; \quad T_1^{(2)}, \ldots, T_S^{(2)}; \quad T_1^{(3)}, \ldots, T_S^{(3)}$$

- Bereken voor $k = 1, 2, 3$

$$\widehat{\text{gemiddelde}} = \frac{1}{S} \sum_{s=1}^{S} T_s^{(k)} = \overline{T}^{(k)}$$

$$
\begin{aligned}
\widehat{\mathsf{bias}} &= \overline{T}^{(k)} - \mu \\
\widehat{\mathsf{SD}} &= \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(T_s^{(k)} - \overline{T}^{(k)})^2} \\
\widehat{\mathsf{MSE}} &= \frac{1}{S}\sum_{s=1}^{S}(T_s^{(k)} - \mu)^2 \approx \widehat{\mathsf{SD}}^2 + \widehat{\mathsf{bias}}^2
\end{aligned}
$$

## Relatieve efficiëntie:

- Voor **elke** schatters waarvoor

$$E(T^{(1)}) = E(T^{(2)}) = \mu \;\Rightarrow\; RE = \frac{\mathsf{var}(T^{(1)})}{\mathsf{var}(T^{(2)})}$$

  is de relatieve efficiëntie van schatter 2 tot schatter 1

- Als de schatters **biased** zijn, berekent men standaard

$$RE = \frac{\mathsf{MSE}(T^{(1)})}{\mathsf{MSE}(T^{(2)})}$$

- In beide gevallen: $RE < 1$ betekent dat schatter 1 de voorkeur heeft (schatter 2 is in deze zin inefficiënt ten opzichte van schatter 1).

## In R:

```
set.seed(3)
S <- 1000
n <- 15
trimmean <- function(Y){mean(Y,0.2)}
mu <- 1
sigma <- sqrt(5/3)
res <- matrix(ncol=5,nrow=S)

for (i in 1:S){
   y <- rnorm(n,mu,sigma)
   # y <- mu + sigma*rt(n,df=1)
   # a <- 1/(sigma^2)
   # s <- mu/a
   # y <- rgamma(n,shape=a,scale=s)
   res[i,1] <- mean(y)
   res[i,2] <- trimmean(y)
   res[i,3] <- median(y)
   res[i,4] <- sd(y)
   res[i,5] <- t.test(y,mu=1)$p.value
}
```

## Results:

```
head(round(res,4),5)
```

```
First 5 rows

       [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 0.7539 0.7132 1.0389 1.0165 0.3644
[2,] 0.6439 0.4580 0.3746 1.1081 0.2337
[3,] 1.5553 1.6710 1.9395 0.9814 0.0458
[4,] 0.5171 0.4827 0.4119 1.3892 0.1997
[5,] 1.3603 1.4621 1.3452 0.8438 0.1204
```

```
apply(res,2,mean)
```

```
[1] 0.9851517 0.9868953 0.9917283 1.2738126 0.5042291
```

```
apply(res,2,sd)
```

```
[1] 0.3308836 0.3480049 0.3976266 0.2399838 0.2889470
```

## Performantie van schattingen van onzekerheden:

Hoe goed vertegenwoordigen geschatte standaardfouten de **ware steekproef variatie**??

- Bv., voor steekproef gemiddelde $T^{(1)}(Y_1, \ldots, Y_n) = \overline{Y}$

$$SE(\overline{Y}) = \frac{s}{\sqrt{n}}, \quad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$$

- De Monte Carlo (MC) standaard deviatie benaderd de **ware steekproef variation**

  $\Rightarrow$ Vergelijk **gemiddelde** van geschatte standaard fouten
  met MC standaarddeviatie

## Voor steekproef gemiddelde:

(MC standard deviation 0.331)

```
apply(res,2,sd)

[1] 0.3308836 0.3480049 0.3976266 0.2399838 0.2889470

apply(res,2,mean)[4]/sqrt(n)

[1] 0.328897
```

## Gebruikelijk 100(1-$\alpha$)% betrouwbaarheidsinterval voor $\mu$:

Gebaseerd op steekproefgemiddelde

$$\left[ \overline{Y} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{Y} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}} \right]$$

Bereikt het interval het nominale dekkingsniveau $1-\alpha$?

```
low <- res[,1] - qt(0.975,n-1)*res[,4]/sqrt(n)
upp <- res[,1] + qt(0.975,n-1)*res[,4]/sqrt(n)
mean(ifelse((low<mu)&(mu<upp),1,0))
```

```
[1] 0.949
```

## Simpel voorbeeld:

Grootte en power van de gebruikelijke $t$-test voor het gemiddelde

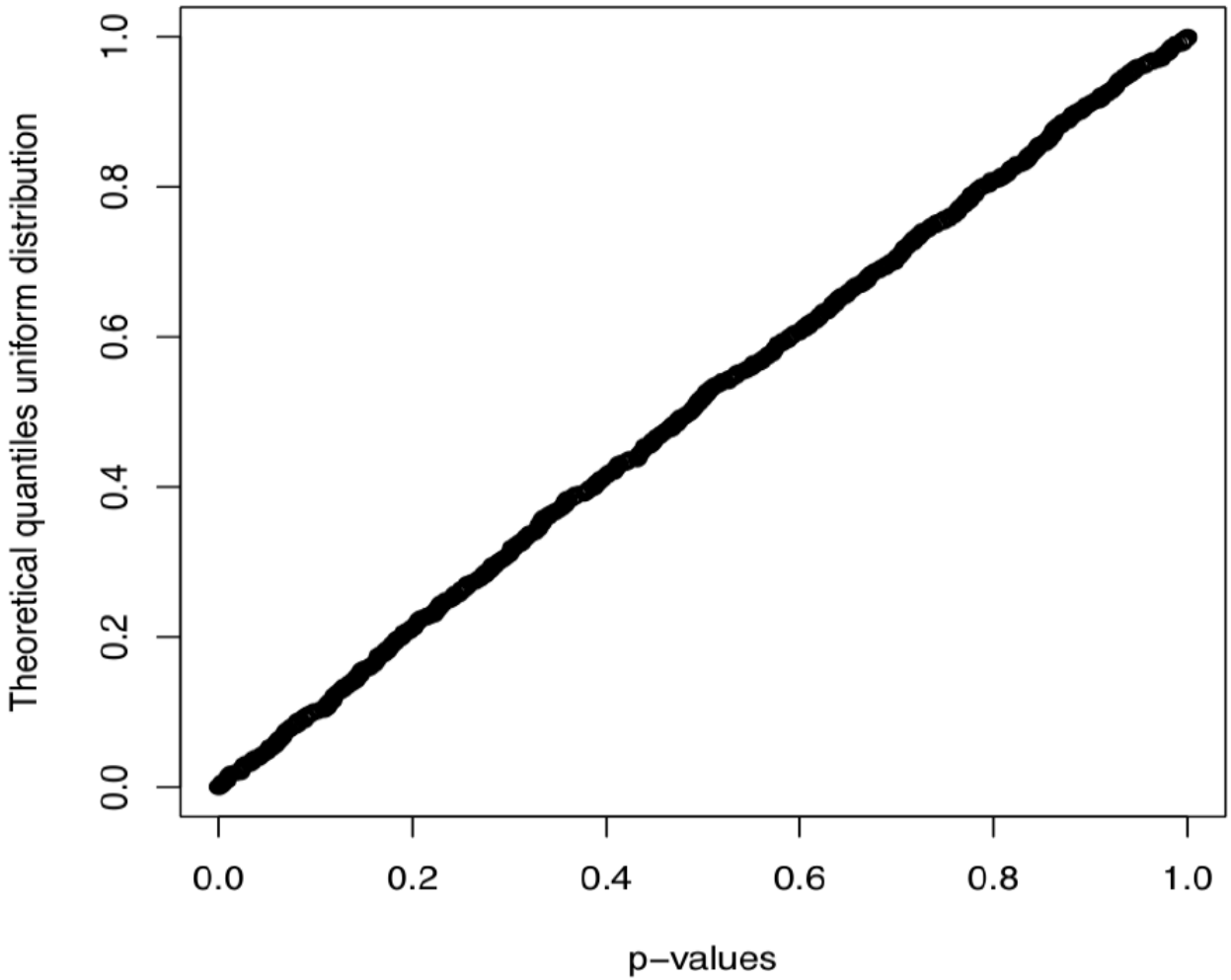
$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

- Om te evalueren of de grootte/niveau van de test de vooropgestelde $\alpha$ bereikt, genereer gegevens onder $\mu = \mu_0$ en bereken de proportie van afwijzingen van $H_0$.
- Benader de **ware** kans om $H_0$ te verwerpen als het waar is $\approx \alpha$.
- Om de power te evalueren, genereer data onder een alternatief $\mu \neq \mu_0$ en bereken de proportie van afwijzingen van $H_0$.
- Benader de **ware** kans van het verwerpen van $H_0$ wanneer het alternatief waar is (power).
- Als de werkelijke grootte $> \alpha$ is, dan heeft de evaluatie van de power fouten.

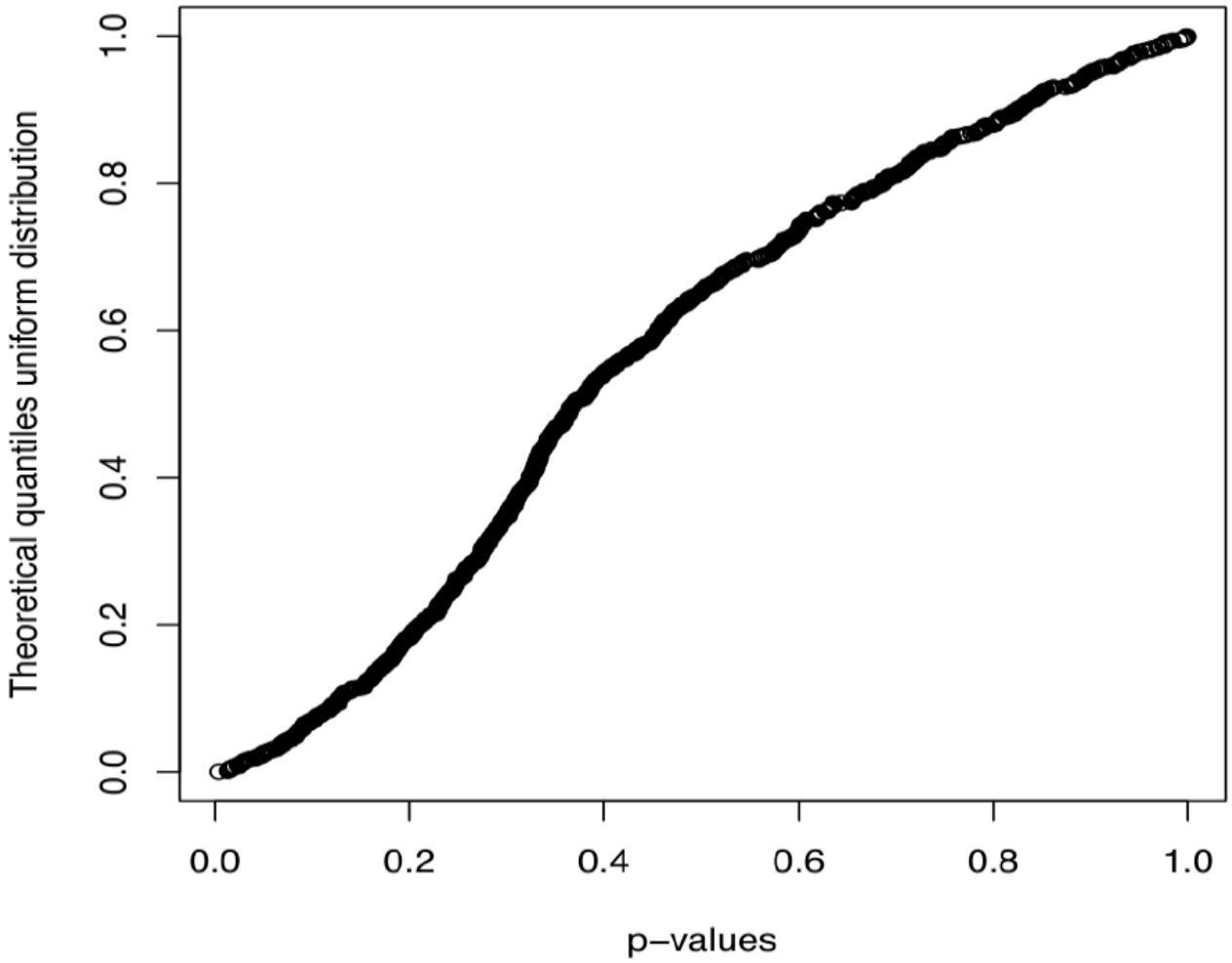
## Grote van test (gamma verdeling):

```
mean(ifelse(res[,5]<0.05,1,0))

[1] 0.126

qqplot(res[,5], ppoints(S), xlab="p-values",
    ylab="Theoretical quantiles uniform distribution")
```

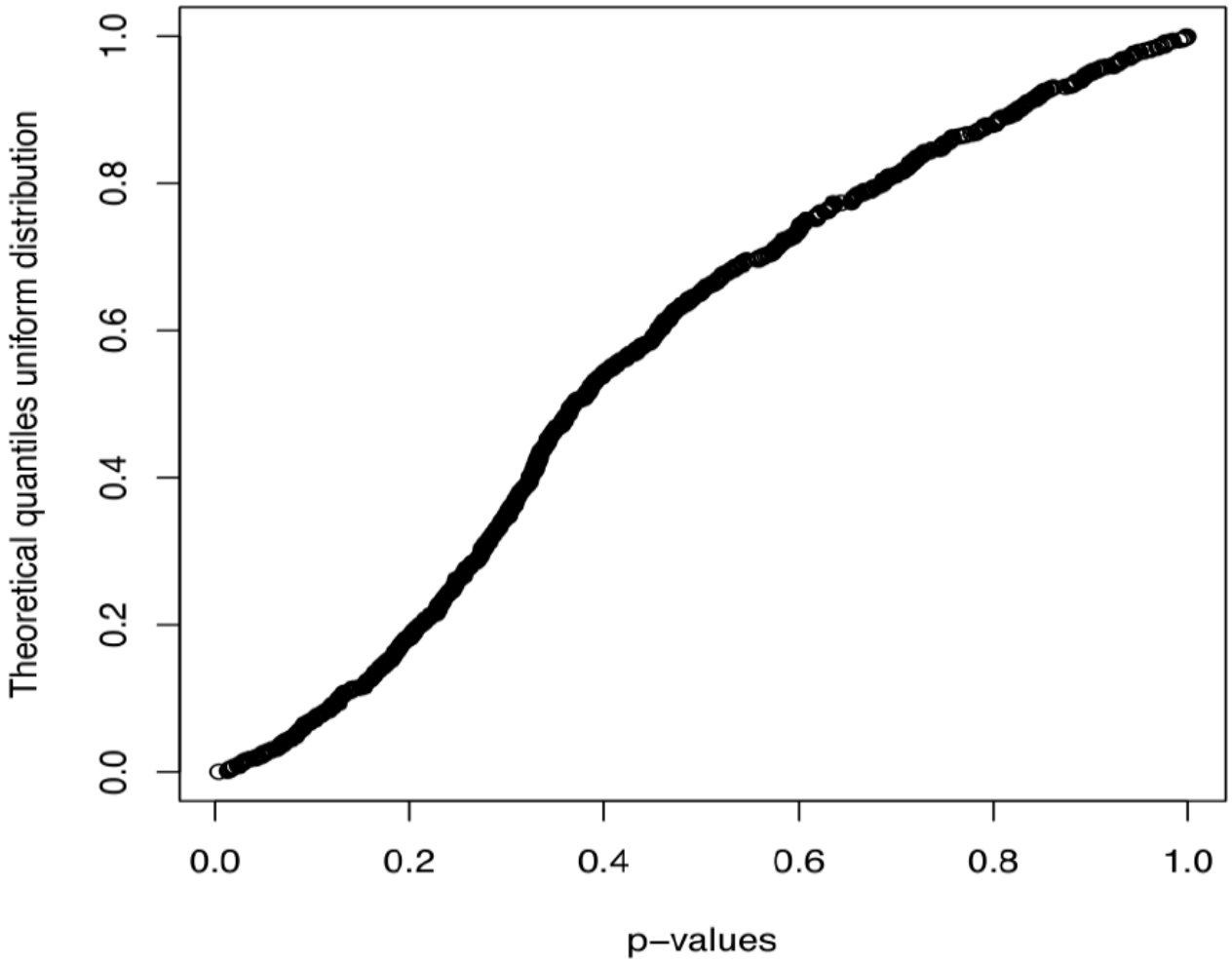
# QQ-plot p-waardes voor normale verdeling



# QQ-plot p-waardes voor t-verdeling



## QQ-plot p-waardes voor gamma verdeling



**Probleem:**

Hoe goed approximeren de **Monte Carlo metingen** de geschatte eigenschappen van de **ware steekproef verdeling** van de schatter/teststatistiek?

- Is $S = 1000$ groot genoeg om een idee te krijgen van de echte steekproef eigenschappen? Hoe "geloofwaardig" zijn de resultaten?
- Een simulatie is gewoon een experiment zoals elk ander, dus **gebruik statistische principes**!
- Elke dataset levert een trekking op van de echte steekproef verdeling, dus $S$ is de "**steekproefgrootte**" waarop schattingen van gemiddelde, bias, SD, enz. van deze verdeling zijn gebaseerd.
- Selecteer een "**steekproefgrootte**" (aantal datasets $S$) dat een acceptabele precisie van de approximatie zal bereiken!

## Kiezen van $S$:

Schatting voor $\theta$ (werkelijke waarde $\theta_0$)

- Schatting van het gemiddelde van steekproef verdeling/bias:

$$\sqrt{\mathsf{var}(\overline{T} - \theta_0)} = \sqrt{\mathsf{var}(\overline{T})} = \sqrt{\mathsf{var}\left(S^{-1}\sum_{s=1}^{S} T_s\right)} = \frac{\mathsf{SD}(T_s)}{\sqrt{S}} = d$$

  waar $d$ de acceptabele fout is, dus we verkrijgen:

$$S = \frac{\{\mathsf{SD}(T_s)\}^2}{d^2}$$

- Kan "raden" $\mathsf{SD}(T_s)$ uit asymptotische theorie, voorrondes.

## Coverage kansen, grootte, power

- Schatten van een **proportion** $p$ (coverage kansen, grootte, vermogen)
- Bv., voor een hypothese test, $Z = \#\text{verwerpingen} \sim \text{binomial}(S, p)$ zodat

$$\sqrt{\text{var}\left(\frac{Z}{S}\right)} = \sqrt{\frac{p(1-p)}{S}}$$

- Slechtste geval by $p = 1/2 \Rightarrow 1/\sqrt{4S}$
- Als $d$ een acceptabele fout is, dan $S = 1/(4d^2)$;
  bv., $d = 0.01$ yields $S = 2500$.
- Voor coverage, grootte, $p = 0.05$.

## Sommige interessante principes:

- Zorgvuldige **planning** is vereist.
- **Factoren** die van belang zijn om in het experiment te variren:
  steekproef grootte $n$, verdeling van de data, omvang van de variatie, ...
- Resultaten moeten **opgenomen en opgeslagen** worden in een systematische, verstandige manier.
- Kies niet alleen factoren **gunstig** voor een methode die je hebt ontwikkeld!
- **Steekproef grootte** $S$ (aantal datasets) moet een acceptabele precisie opleveren, maar houd het eerst klein.
- Stel een **verschillende seed** in voor elke run en **houd records bij**!!!
- **Documenteer your code**!!!

## Belangrijke principes voor rapporteren:

Simulatie is **nutteloos** tenzij anderen **ondubbelzinnig en duidelijk** begrijpen wat je deed, waarom je het deed en wat het betekent!

- Geef de **objectieven** – Waarom deze simulatie? Welke specifieke vragen probeert u te beantwoorden?
- Review alle **methoden** die worden bestudeerd – wees precies en gedetailleerd.
- Beschrijf **precies** hoe je gegevens hebt gegenereerd voor elke keuze van factoren – er moeten voldoende details gegeven worden zodat een lezer zijn/haar **eigen programma** zou kunnen schrijven om je resultaten te reproduceren!

# Deel 4:

# Hypothese toetsen

# 4.1 Voorbeeld

- Is de gemiddelde kostprijs voor huisarts (fp) anders dan 130?

De procedure om te beslissen of er voldoende bewijs is om aan te nemen dat de gemiddelde kosten voor huisarts anders zijn dan 130, wordt **hypothesetoets** genoemd!

# 4.2 Nul en alternatieve hypothese

- In de praktijk wordt de onderzoeksvraag geformuleerd in termen van een **nullhypothese $\mathbf{H_0}$** en een **alternatieve hypothese $\mathbf{H_a}$**:

$$H_0 : \mu_{fp} = 130 \qquad H_a : \mu_{fp} \neq 130$$

- Op basis van onze waargenomen gegevens zullen we onderzoeken of $H_0$ kan afgewezen ten gunste van $H_a$

- Zo niet, dan is de nulhypothese $H_0$ aanvaard en besluit men dat de gemiddelde kosten voor huisarts niet anders zijn dan 130

# 4.3 Centrale limiet stelling

- Om dit besluit te kunnen trekken, gaan we gebruik maken van een steekproef
- Gebaseerd op deze steekproef kunnen we $\mu_{fp}$ schatten a.d.h.v. het steekproefgemiddelde $\bar{x}$, welke 1 gerealiseerde waarde is van de stochastische veranderlijke $\bar{X}$
- **Vraag:**
  - Wanneer bevindt de verdeling van $\bar{X}$ zich rond $\mu$?
  - Wanneer vertoont de verdeling van $\bar{X}$ veel (weinig) variabiliteit?

- **Antwoord:** **Centrale limiet stelling (CLT)**

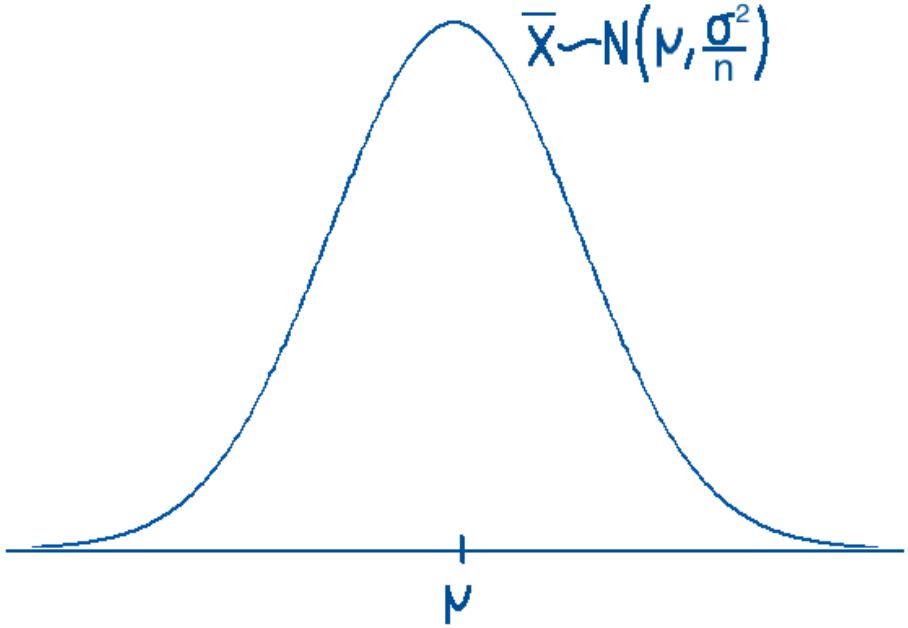
Voor elke willekeurige variabele $X$ met gemiddelde $\mu$ en variantie $\sigma^2$, en voor $n$ voldoende groot,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Dit houdt in dat voor voldoende grote steekproeven $\bar{x}$ een **unbiased schatting** is voor $\mu$, wat nauwkeuriger is naarmate de steekproef groter wordt

- Deze resultaten zijn de **belangrijkste motivatie voor het uitvoeren van grote studies**, aangezien het verzamelen van aanvullende informatie (meer waarnemingen, grotere steekproef) zal leiden tot verhoogde precisie in de schatting

- Dankzij de CLT kunnen we berekenen hoe waarschijnlijk het is dat een schatting ver van de juiste waarde of dicht bij de juiste waarde ligt




- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N\left(0,1\right)$, the standaard normale verdeling
- Maar: $\sigma^2$ is vaak ongekend!

- We kunnen de populatie standaarddeviatie $\sigma$ schatten met de steekproefstandaarddeviatie $s$

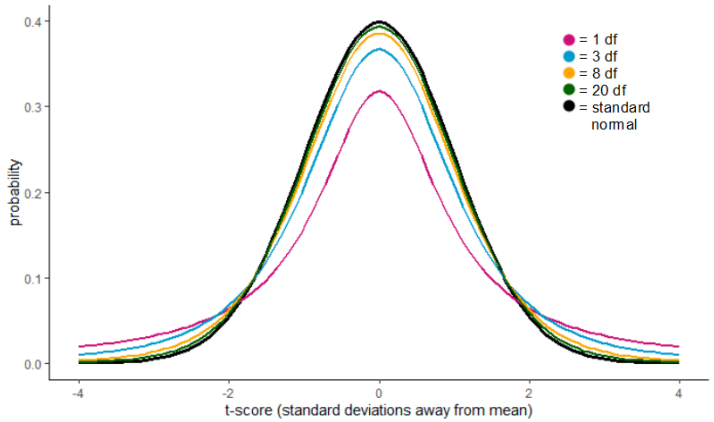
- **Vraag:** Welke verdeling zal $\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$ dan volgen?

- **Antwoord:**
  - **Voor $\mathrm{n}$ voldoende groot:**

$$\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} \sim t_{df=n-1}$$

  - **Voor $\mathrm{n}$ klein en $\mathrm{X}$ normaal verdeeld:**

$$\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} \sim t_{df=n-1}$$



- **Bewijs:**

$$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{S^2/\sigma^2}} = \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$$

Uit de CLT weten we dat $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Verder, $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$ en $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$.

Aangezien de laatste formule van de vorm $W = U + V$ is en $U$ en $V$ zijn onafhankelijk, kunnen we zeggen dat: $M_W(t) = M_U(t) + M_V(t)$.

Nu,

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow V = \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1$
  Thus: $M_V(t) = (1-2t)^{-1/2}$.

- $X_i \sim N\left(\mu, \sigma^2\right) \Rightarrow \frac{X_i-\mu}{\sigma} \sim N(0,1) \Rightarrow W = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_n$
  Thus: $M_W(t) = (1-2t)^{-n/2}$

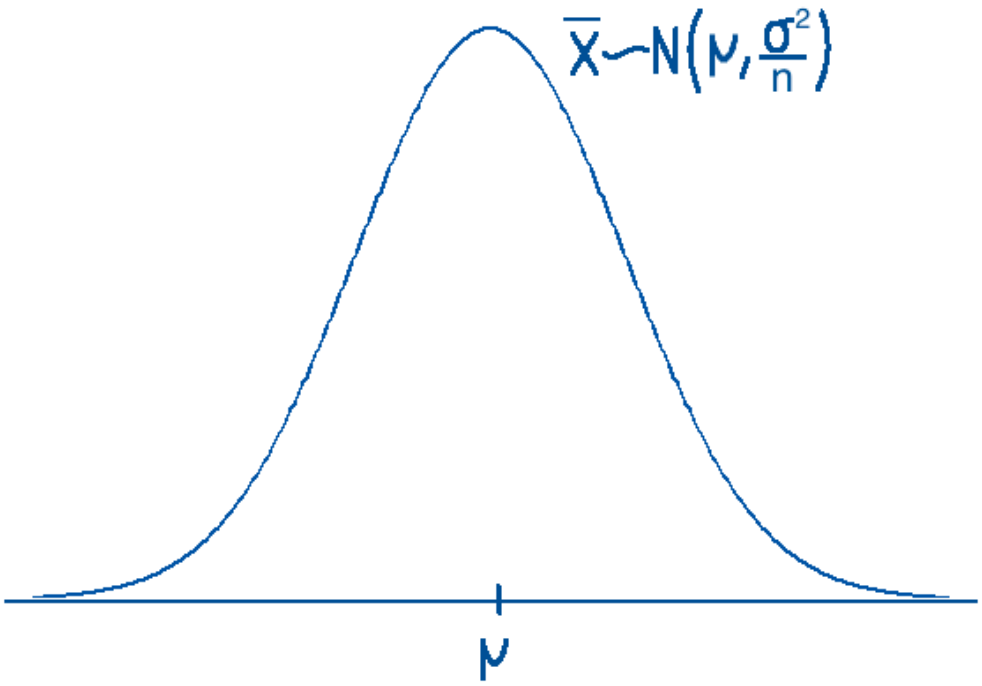
$\Rightarrow M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$, welke de moment genererende functie is van een stochastische veranderlijke met $\chi_{n-1}$ verdeling.

Aangezien $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}$ en de definitie van een t-verdeling de volgende is: Als $Z \sim N(0,1)$, $U \sim \chi_{n-1}$ en $Z$ en $U$ zijn onafhankelijk, dan $\frac{Z}{\sqrt{U/(n-1)}} \sim t_{n-1}$; hebben we het bewezen!

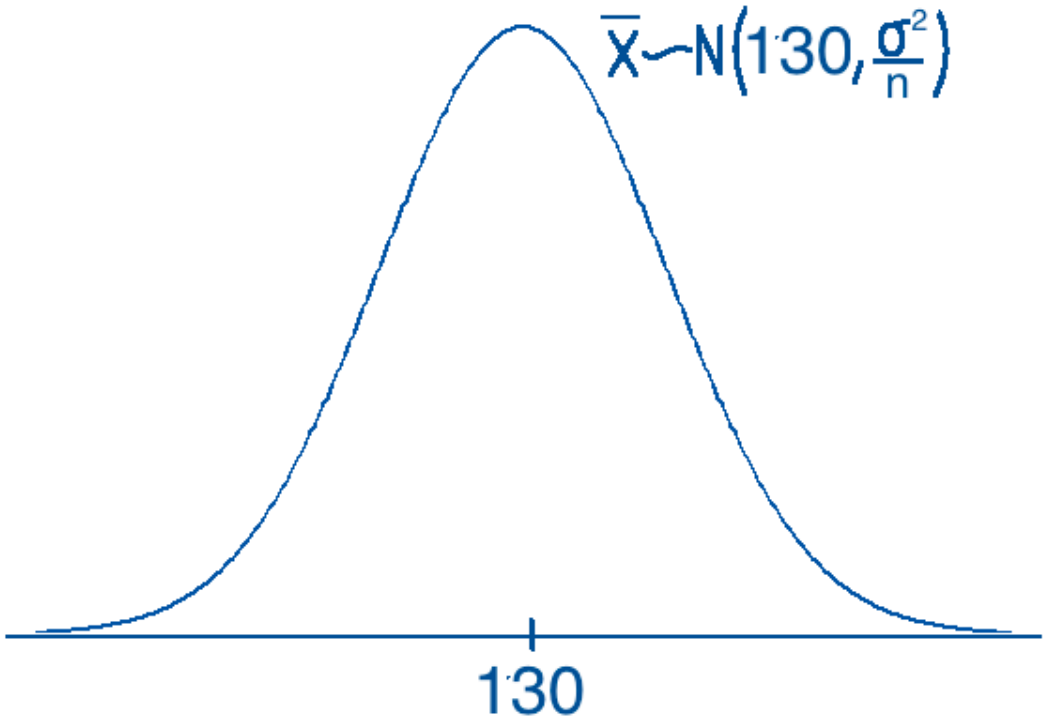
# 4.4 Test statistiek

- Intuïtief is het duidelijk dat $H_0 : \mu_{fp} = 130$ wordt verworpen als het waargenomen steekproefgemiddelde $\bar{x}$ te ver weg van 130 ligt
  - **Steekproef:** $\bar{x} = 63.74$
- **Vraag:** Maar wat is te ver weg?
- Als dit resultaat zeer onwaarschijnlijk is door puur toeval
- Anders gezegd, als dit resultaat helemaal niet is wat je verwacht te zien als $\mu_{fp} = 130$

- De CLT zal ons helpen bij het beslissen, aangezien het beschrijft welke waarden voor $\bar{x}$ verwacht kunnen worden als men herhaaldelijk nieuwe steekproeven zou trekken. Als $n$ voldoende groot is, hebben we dat:



- Hier willen we weten welke waarden voor $\bar{x}$ kunnen worden verwacht als $\mu_{fp} = 130$



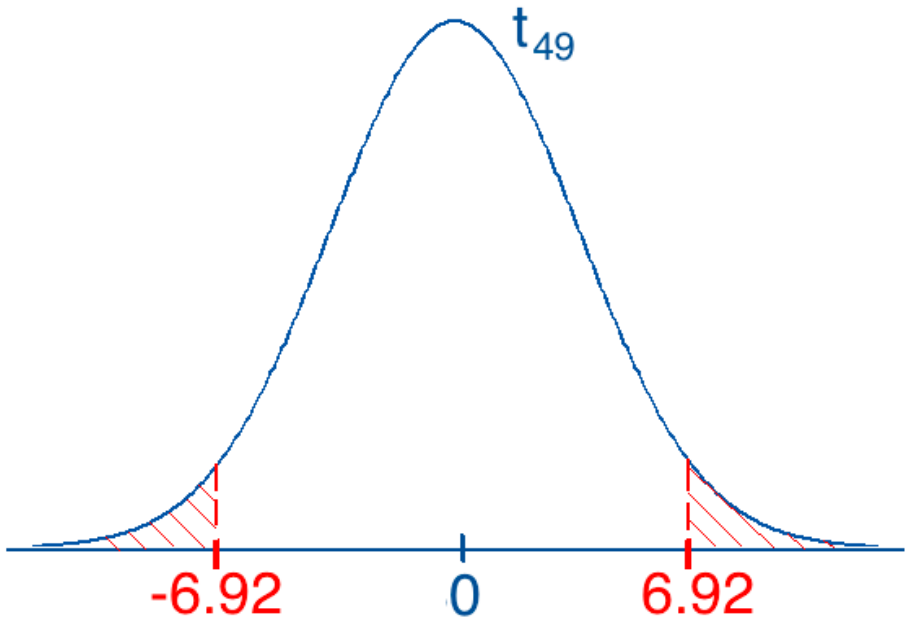
- Zoals eerder gezegd, $\sigma^2$ is vaak onbekend, en we zullen ons baseren op

$$T = \frac{\bar{X} - 130}{\sqrt{\frac{s^2}{n}}} \overset{H_0}{\sim} t_{df=n-1}$$

  - T heet een **test statistiek**
    - In statistiek bestaan er verschillende test statistieken, afhankelijk van de gedefinieerde hypothese

  - T kan worden berekend voor onze gegeven steekproef, gedefinieerd als de **steekproefwaarde (t)**
    - $t = \frac{63.74 - 130}{9.57} = -6.92$

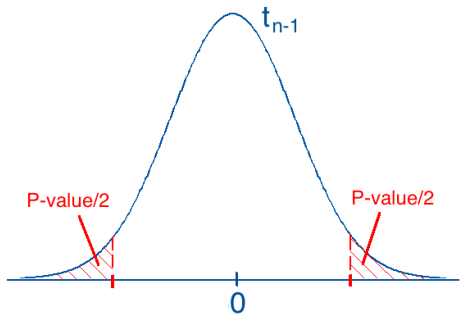
- Om $H_0$ te afwijzen, moet het **rode gebied** voldoende laag zijn



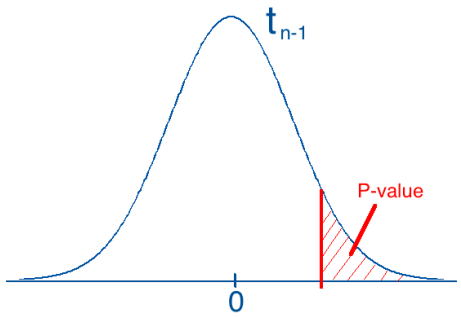

  - In statistiek wordt het **rode gebied** gedefinieerd als de **P-waarde**
  - **P-waarde** = **Kans** dat, als $H_0$ waar is, de **teststatistiek** even **extreem is als of extremer is dan** de **steekproefwaarde**

- Afhankelijk van de hypothese worden voor deze waarde verschillende locaties gegeven

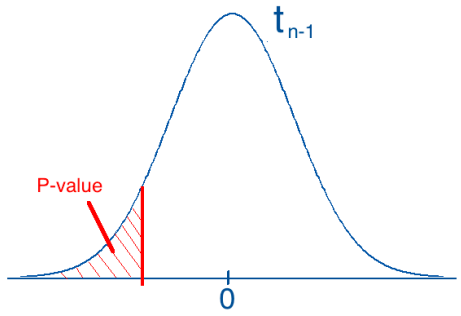
- $H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$




- $H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$




- $H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$

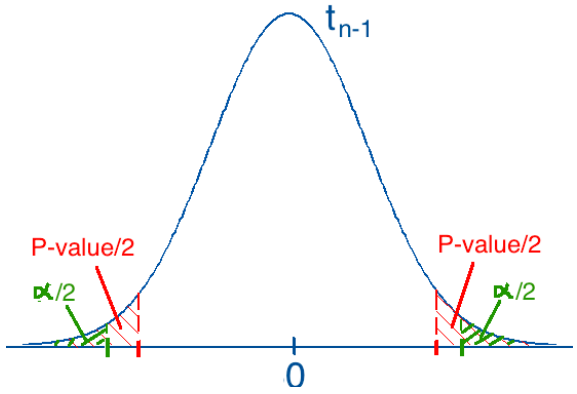


- **Vraag:** Maar wat is voldoende laag?

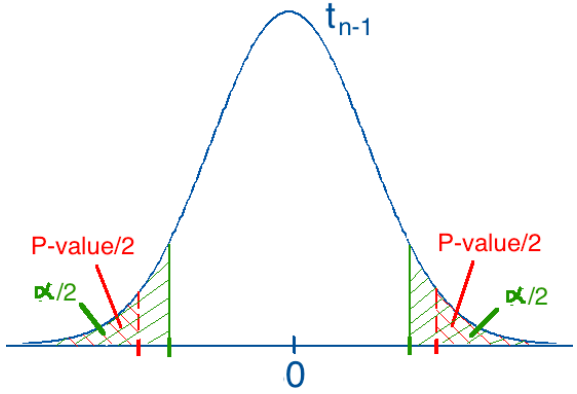
- Men specificeert daarom het **significantieniveau** $\alpha$
  - $\alpha =$ **Kans dat**, als $H_0$ waar is, de **test statistiek** dit niet kan detecteren
  - Dit betekent dat het betrouwbaarheidsniveau $(1 - \alpha)$ de kans definieert dat, als $H_0$ waar is, de test dit kan detecteren
  - In de praktijk wordt vaak $1\%$, $5\%$ of $10\%$ gebruikt voor $\alpha$

# 4.5 Besluitvorming

- **Besluit:**
  - Accepteer $H_0$ als $P \geq \alpha$




  - Verwerp $H_0$ als $P < \alpha$



# 4.6 Soorten fouten

| | | **Realiteit** | |
|---|---|---|---|
| | | $H_0$ correct | $H_0$ niet correct |
| **Test resultaat** | $H_0$ correct | ✓ | **Type II error** |
| | $H_0$ niet correct | **Type I error** | ✓ |

- **Type I fout:**
  - Treedt op als $H_0$ correct is, maar de test leidt tot een significant resultaat
  - **Vraag:** Hoe groot is de kans dat dit gebeurt?
  - Stel dat de test wordt uitgevoerd bij $\alpha = 5\%$
  - Als $H_0$ correct is, zal men in 5% van de gevallen een significant resultaat zien

- Dus, P(Type I fout) $= \alpha$

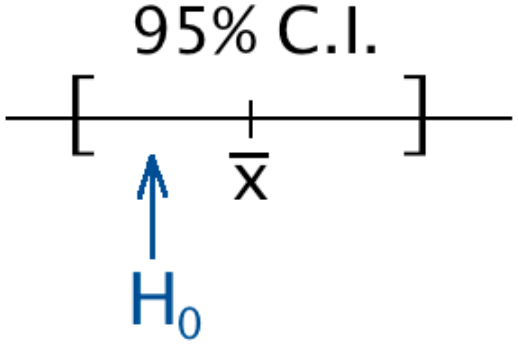
- **Type II fout:**
  - Treedt op als $H_0$ onjuist is, maar de test dit niet heeft gedetecteerd
  - **Vraag:** Hoe groot is de kans dat dit gebeurt?
  - In tegenstelling tot de type I-fout, is de kans op het maken van een type II-fout niet gemakkelijk te beheersen, en hangt af van verschillende aspecten van de steekproef(en) en populatie(s), en wordt aangegeven met $\beta$
  - De **power** van een statistische toets is $\mathbf{1 - \beta}$, de kans om $H_0$ correct te verwerpen

# 4.7 Hypothesetesten versus betrouwbaarheids intervallen

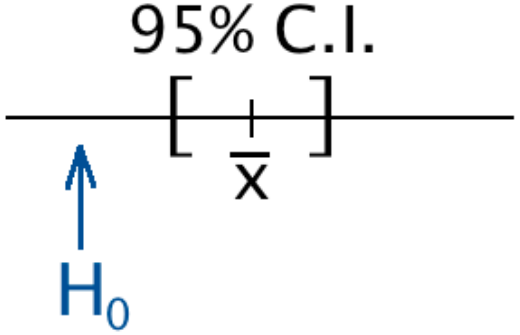
- Voor het voorbeeld kunnen er conclusies getrokken over de populatie gemiddelde kosten per patiënt per maand
  - $H_0 : \mu_{fp} = 130$ & $H_a : \mu_{fp} \neq 130 \rightarrow p < 0.00001$
  - $95\%$ C.I.: $[44.51; 82.98]$
- We weten van de C.I. dat de gemiddelde gemiddelde kosten waarschijnlijk tussen $44,51$ en $82,98$ zullen zijn, exclusief $130$
- De significantietest heeft de waarde 130 afgewezen als mogelijke waarde voor $\mu_{fp}$
- Dus **beide procedures komen overeen!**

- **Vraag:** Maar komen deze altijd overeen?

- **Antwoord:** Ja, op voorwaarde dat de niveaus van significantie en betrouwbaarheid **complementair aan elkaar zijn**:
  - Accepteer $H_0$ ($p \geq \alpha = 0.05$:)




  - Verwerp $H_0$ ($p < \alpha = 0.05$:)



# Deel 5:

# Enkele veelgebruikte parametrische testen

# 5.1 Analyse van 1 gemiddelde

- **Hypothese:**

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$$

- **Test statistiek:** One-sample t-test

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} \overset{H_0}{\sim} t_{df=n-1}$$

- **Assumptie:**
  - Grote steekproef ($\geq 30$) of kleine steekproef van een normale verdeling

- **Voorbeeld:**

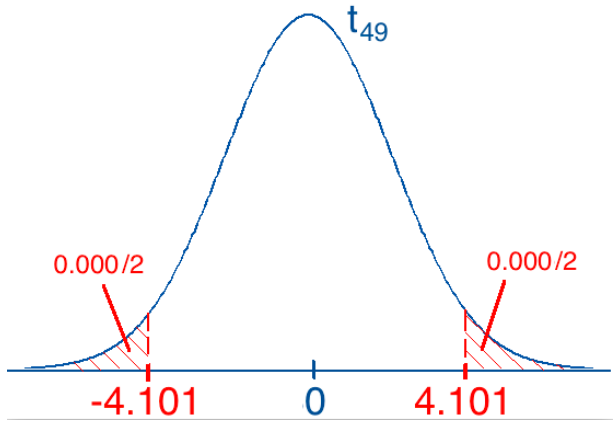
| | n | Mean | Std dev | Se | t | df | p-value | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|---|---|---|
| Average cost | 50 | 63.744 | 67.684 | 9.572 | -4.101 | 49 | 0.000 | 55.508 | 82.980 |

$\bar{x}$

$$T_t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$n - 1$

P-value (2-sided)



# 5.2 Vergelijking van 2 gemiddelden: Ongepaarde data

- Stel dat we geïnteresseerd zijn in een **kwantitatieve vergelijking van twee groepen**, waarbij er **geen relatie** is tussen de waarnemingen van de eerste groep en die van de tweede groep (ongepaard)

  **Voorbeelden:**

  - Verschillen de gemiddelde kosten per patiënt van artsen met een tweede specialisme van de gemiddelde kosten per patiënt van artsen zonder?
  - Zijn de kosten verschillend voor mannelijke en vrouwelijke artsen?
  - Is er een verschil tussen artsen die een medische opleiding hebben gevolgd in de VS of in het buitenland?

- In plaats van te testen met een referentiewaarde ($\mu_0$), worden 2 onafhankelijke groepen (aangeduid met $1$ en $2$) nu vergeleken:

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

- **Test statistiek: Two-sample unpaired t-test**

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$

  - $s_p^2$ is een schatter van de gepoolde populatievariantie van de twee groepen, als volgt berekend:

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2}{n_1+n_2-2}$$

  - $n_1$ en $n_2$ zijn de steekproefomvang van respectievelijk groep $1$ en $2$

- **Assumpties:**
  1. Homoscedasticiteit
     - Er bestaan testen om de varianties te vergelijken (bv., **Levene's test**)
  2. Grote steekproeven ($\geq 30$) of
     ten minste één kleine steekproef: Beide moeten afkomstig zijn van normale verdeling

- Als homogeniteit niet aanwezig is, kan men vertrouwen op de **Welch t-statistiek**, een aanpassing van de klassieke t-test:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \overset{H_0}{\sim} t_{df}$$

  - $s_1$ en $s_2$ zijn de geschatte standaarddeviatie van respectievelijk groep 1 en 2
  - $df = \left( \frac{s_1^2}{n_1^2} + \frac{s_2^2}{n_2^2} \right) / \left( \frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)} \right)$

- **Voorbeelden:**
  (1) Zijn de totale gemiddelde kosten per patiënt per maand verschillend voor artsen die een medische opleiding hebben gevolgd in de VS of in het buitenland?

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

| | Mean difference | Se | t | df | p-value | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|---|
| Equal variances | 73.583 | 53.750 | 1.369 | 184 | 0.173 | -32.462 | 179.628 |
| Unequal variances | 73.583 | 49.938 | 1.473 | 102.424 | 0.144 | -25.494 | 172.630 |

| | Levene's test | p-value |
|---|---|---|
| Average costs | 5.166 | 0.024 |

- **Assumpties nakijken:**
  1. Homoscedastisticiteit
     - **Hypothese:** $H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_a : \sigma_1^2 \neq \sigma_2^2$
     - **Test statistiek: Levene's test**
     - **Besluitvorming:** $P = 0.024 < 0.05 \rightarrow$ Ga niet uit van gelijke varianties!
  2. Grote steekproeven: $100$ in USA, $86$ in buitenland

- We zullen dus gebruik maken van de **Welch t-test**, toegewezen in de tweede rij (UNEQUAL VARIANCES)
  - **Besluitvorming:** $P = 0.144 > 0.05$
  - **Geen significant verschil** tussen gemiddelde kosten voor artsen die een medische opleiding hebben gevolgd in de VS of in het buitenland

(2) Is de totale gemiddelde kost per patiënt per maand verschillend voor artsen met of zonder tweede specialisme?

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

| | Mean difference | Se | t | df | p-value | lower 95% CI | upper 95% CI |
|---|---|---|---|---|---|---|---|
| Equal variances | -0.987 | 74.247 | -0.013 | 184 | 0.989 | -147.471 | 145.498 |
| Unequal variances | -0.987 | 41.558 | -0.024 | 113.081 | 0.981 | -83.968 | 81.995 |

| | Levene's test | p-value |
|---|---|---|
| Average costs | 0.106 | 0.746 |

- **Assumpties nakijken:**
  1. Homoscedastisticiteit
     - **Hypothese:** $H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_a : \sigma_1^2 \neq \sigma_2^2$
     - **Test statistiek: Levene's test**
     - **Besluitvorming:** $P = 0.746 > 0.05 \rightarrow$ Ga uit van gelijke varianties!
  2. Kleine steekproef: $157$ zonder, $29$ met $2^{de}$ specialisme
     $\rightarrow$ Normaliteit moet worden gecontroleerd (zie Chapter $6.2$)

- We zullen gebruik maken van de **klassieke t-test**, toegewezen in de eerste rij (EQUAL VARIANCES)
  - **Besluitvorming:** $P = 0.989 > 0.05$
  - **Geen significant verschil** tussen gemiddelde kosten voor artsen met of zonder tweede specialisme

# 5.3 Vergelijking van 2 gemiddelden: Gepaarde data

- Het kan voorkomen dat beide groepen aan elkaar zijn gekoppeld, bv.,
  - Meting voor & meting na
  - Linkeroog & rechteroog
  - Vraag zonder hulp & met hulp
- Vergeleken met de ongepaarde t-toets, is het belangrijkste verschil dat elke waarneming van het eerste steekproef nu op unieke wijze overeenkomt met één waarneming van het tweede steekproef, en vice versa
- Dus, we hebben **gepaarde data**

- De variabele van belang wordt het **verschil van de kwantitatieve meting tussen beide groepen**

$$D = X_1 - X_2$$

- Om beide middelen te vergelijken, wordt de volgende **hypothese** geconstrueerd:

$$H_0 : \mu_D = 0 \qquad H_a : \mu_D \neq 0$$

- **Test statistiek: One-sample t-test**

$$T = \frac{\bar{X_D}}{\sqrt{\frac{s_D^2}{n}}} \overset{H_0}{\sim} t_{n-1}$$

- De one-sample t-test voor het verschil wordt vaak de **two-sample paired t-test** genoemd

- **Assumptie:**
  - Grote steekproef ($\geq 30$) of kleine steekproef van een normale verdeling
- **Voorbeeld: Een meting voor een evenement & een meting na het evenement**
  - Data:

| Id | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| before | 93 | 77 | 98 | 107 | 84 | 99 | 103 | 75 | 76 | 96 |
| after | 69 | 81 | 75 | 84 | 65 | 85 | 73 | 68 | 72 | 96 |
| difference | 24 | -4 | 23 | 23 | 19 | 14 | 30 | 7 | 4 | 0 |

  - **Vraag:** Is er een verschil in gemiddelde voor en na het evenement?

- Uitkomst:

$$H_0 : \mu_D = 0 \qquad H_a : \mu_D \neq 0$$

| | Mean | Se | t | df | p-value | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|---|---|---|
| Before-after | 14 | 3.670 | 3.815 | 9 | 0.004 | 5.699 | 22.301 |

  - Verschil in gemiddelde: $14$
  - P-waarde $= 0.004$
  - Kleine steekproef ($n = 10$), dus de steekproef moet van een normale verdeling komen

# Deel 6:

# Assumpties nagaan

# 6.1 Introductie

- Voor alle testen die in deel 5 zijn besproken, hebben we enkele veronderstellingen gemaakt:
  - Normaliteit van gegevens (als geheel of binnen een groep)
  - Homogeniteit (voor 2 of meer groepen)
- In dit hoofdstuk worden meer details gegeven over de testen voor deze aannames

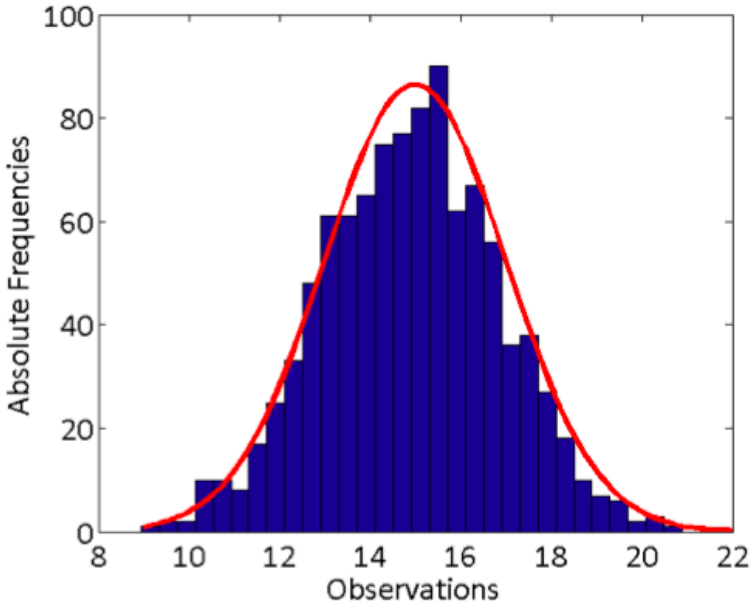
# 6.2 Normaliteit nagaan

- Normaliteit van data (als geheel of binnen een groep) kan worden gecontroleerd met
  - Grafische methodes
    - **Histogrammen**
    - **Probability-probability figuren**
    - **Quantile-quantile figuren**
  - Formele testen
    - **Shapiro-Wilk's test**
    - **Kolmogorov-Smirnov test**

# 6.2.1 Grafische methodes

**Methode 1: Histogram**

- Bereken het steekproefgemiddelde $\bar{x}$ en de steekproefvariantie $s^2$
- Bereken de kansdichtheidsfunctie van de Gauss-verdeling met parameters $\mu = \bar{x}$ en $\sigma^2 = s^2$
- Plot de respectieve dichtheidsfunctie (opnieuw geschaald naar een oppervlak van Nh met h de binbreedte) over het histogram in het midden van elke bin



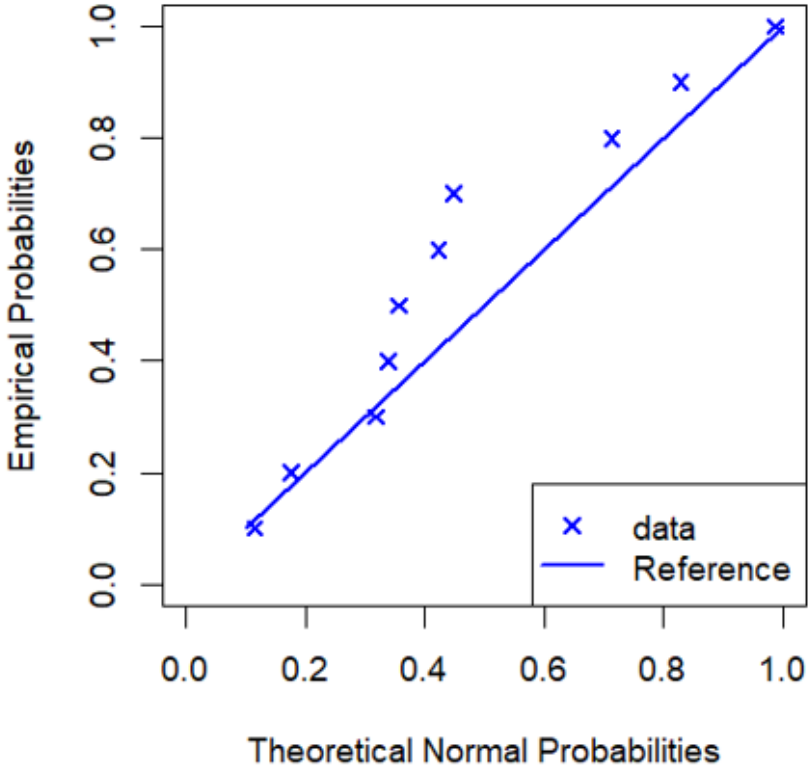
## Methode 2: Probability-Probability plot (pp-plot)

- **Voorbeeld:**

| x | Rank(x) |
|---|---|
| 9.25 | 3 |
| 12.90 | 8 |
| 7.64 | 2 |
| **18.73** | **10** |
| 9.45 | 4 |
| 10.45 | 7 |
| 14.26 | 9 |
| 10.23 | 6 |
| 9.61 | 5 |
| **6.67** | **1** |

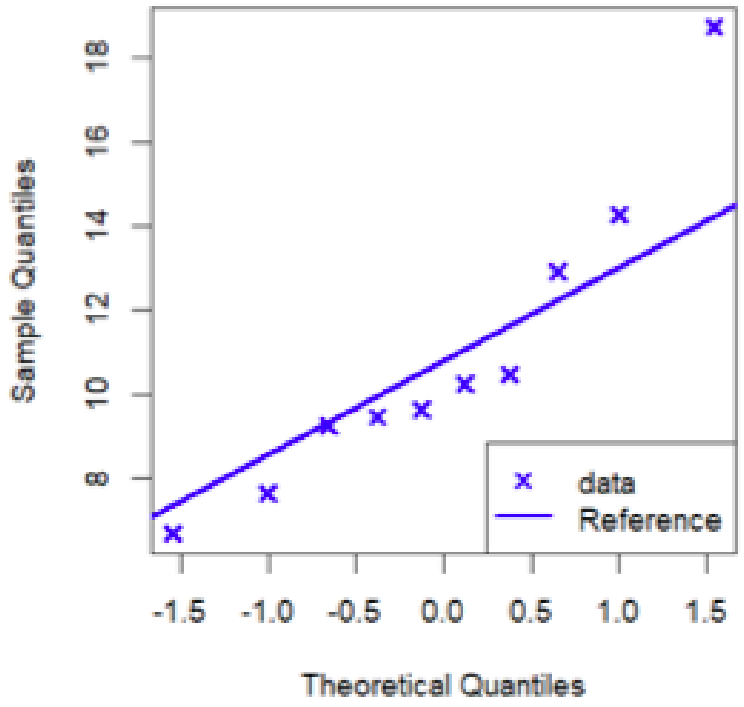
- Bereken de empirische cumulatieve kansen, d.w.z. $P_{ECP}(X \leq x)$, en theoretische cumulatieve kansen op basis van $N(\bar{x}, s^2)$ , d.w.z., $P_{TCP}(X \leq x)$

| x | Emp. cum. prob. | Theor. cum. prob. |
|---|---|---|
| 9.25 | 3/10 = 0.30 | 0.32 |
| 12.90 | 8/10 = 0.80 | 0.71 |
| 7.64 | 2/10 = 0.20 | 0.18 |
| **18.73** | **10/10 = 1.00** | 0.99 |
| 9.45 | 4/10 = 0.40 | 0.34 |
| 10.45 | 7/10 = 0.70 | 0.45 |
| 14.26 | 9/10 = 0.90 | 0.83 |
| 10.23 | 6/10 = 0.60 | 0.42 |
| 9.61 | 5/10 = 0.50 | 0.36 |
| **6.67** | **1/10 = 0.10** | 0.11 |



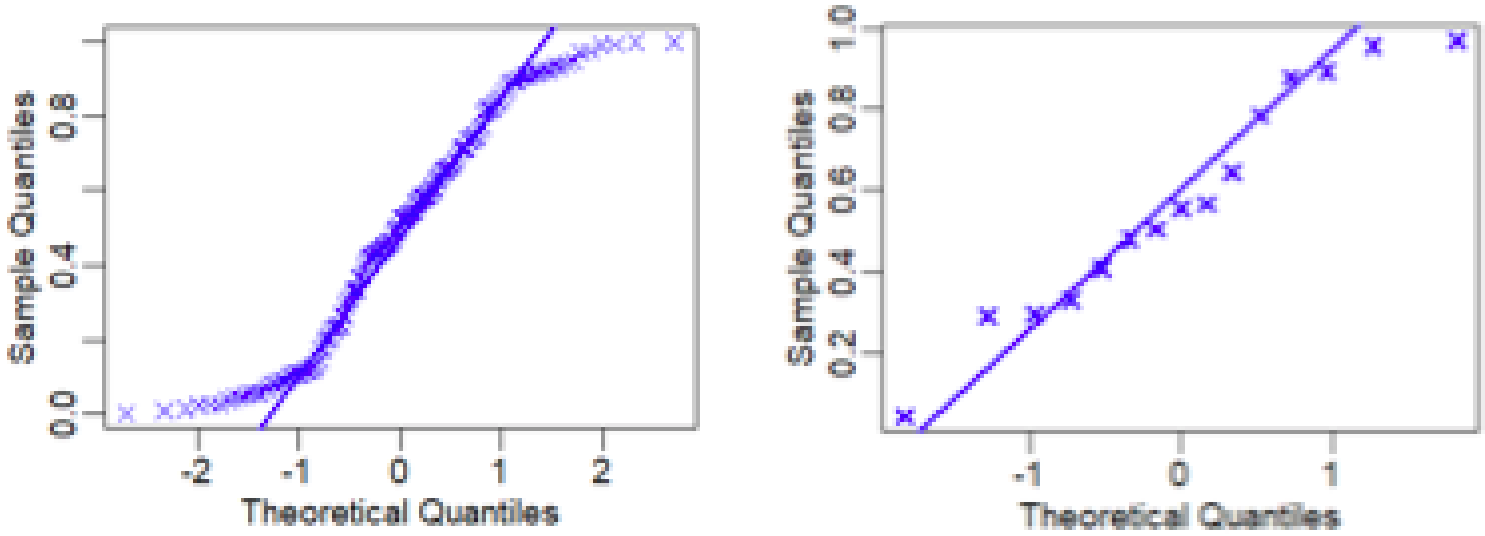
## Methode 3: Quantile-quantile plot (qq-plot)

- Net als bij de pp-plot kunnen de theoretische kwantielen uit de normale verdeling worden berekend en vergeleken met de empirische kwantielen
- Dit geeft aanleiding tot de zogenaamde **qq-plot**



**Belangrijk:** Wees voorzichtig met de interpretatie!

- **Twee voorbeelden:**




  - Op het eerste gezicht lijken beide plots te corresponderen met een normale verdeling

  - Beide voorbeelden zijn echter getrokken uit een andere verdeling

# 6.2.2 Formele testen

- Als alternatief kan men zich wenden tot test statistieken om de normaliteit van de gegevens te onderzoeken

- Hier zullen we ons concentreren op de twee meest populaire formele testen voor normaliteit, namelijk de **Kolmogorov-Smirnov** en **Shapiro Wilk's test**

- Beide testen worden gebruikt voor de **hypothese**

$$H_0 : \text{De gegevens zijn normaal verdeeld}$$

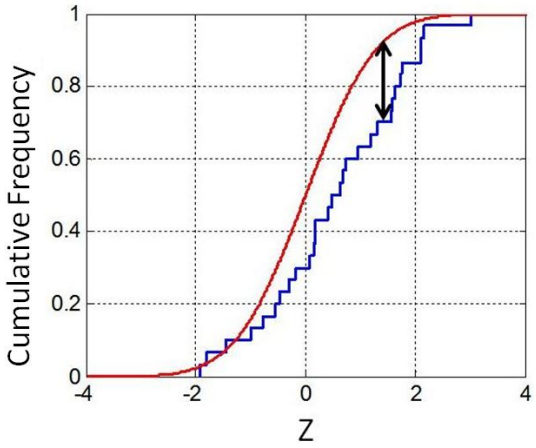
$$H_a : \text{De gegevens zijn niet normaal verdeeld}$$

## Kolmogorov-Smirnov test:

- De **Kolmogorov-Smirnov test** is gebaseerd op het verschil tussen de empirische cumulatieve kansfunctie en de theoretische cumulatieve kansfunctie afkomstig van de normaal verdeling

- **Test statistiek: Kolmogorov-Smirnov test**

$$KS = \max_{\forall x} \mid P_{ECP}(X \leq x) - P_{TCP}(X \leq x) \mid$$



  - Als $H_0$ waar is, verwachten we dat $KS \approx 0$
  - Verwerp $H_0$ wanneer $KS$ significant groter is dan 0

- De **"zwakke" Kolmogorov-Smirnov test** assumeert dat, onder $H_0$, $KS$ een normale verdeling benadert.

- Om een hogere power te bereiken is een correctie toegepast zodat de p-waarde gebaseerd is op een meer exacte verdeling van KS, namelijk de **Lilliefors verdeling** (bekomen door simulaties).

## Shapiro Wilk's test:

- **Idee van de Shapiro Wilk's test:**
  Een normale verdeling heeft specifieke staarten die goed bekend zijn en volledig kunnen worden gekarakteriseerd door zijn variantie. Aangezien de SW test twee schattingen voor de variantie vergelijkt, de klassieke steekproefvariantie en een schatting op basis van de rank statistieken

- De **test statistiek** $SW$ meet in wezen de correlatiecoëfficiënt van de normale $qq$-plot
  - Als $H_0$ waar is, verwachten we dat $SW \approx 1$
  - Verwerp $H_0$ wanneer $SW$ significant kleiner is dan 1

- **Opmerking:** Vanwege de benaderingen die zijn gebruikt om deze test af te leiden, mag de test alleen worden gebruikt voor steekproefgrootte tot 50. Correctie van de Shapiro-Wilk kan echter in sommige softwarepakketten (SPSS, SAS, R) steekproeven tot 2000 mogelijk maken. Voor alle andere steekproefomvang wordt de Kolmogorov-Smirnov test aanbevolen.

- Bovendien bevat de **Shapiro-Wilk's test** vaak de grootste power in vergelijking met de **Kolmogorov-Smirnov test**

# 6.3 Homogeniteit nagaan

- Tot nu toe gebruikten we de **Levene's test** om de homogeniteit te testen
  - **Hypothese:** $H_0 : \sigma_1^2 = \cdots = \sigma_k^2 \qquad H_a : \sigma_i^2 \neq \sigma_j^2$ for any $i \neq j$
  - **Levene's test** is in feite een one-way ANOVA op de varianties. Maar met enkele aanpassingen om het robuuster te maken
  - **Test statistiek:**

$$L = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2} \overset{H_0}{\approx} F_{k-1,n-k}$$

$$Y_{ij} = \mid X_{ij} - \bar{X}_i \mid, \qquad \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \qquad \bar{Y} = \frac{1}{k}\sum_{i=1}^{k} n_i \bar{Y}_i$$

- Er bestaan ook andere testen

## Brown-Forsythe test:

- **Hypothese:** $H_0 : \sigma_1^2 = \cdots = \sigma_k^2 \qquad H_a : \sigma_i^2 \neq \sigma_j^2$ voor $i \neq j$
- Deze test is een andere wijziging van Levene's test, in die zin dat $Y_{ij}$ wordt gedefinieerd als

$$| X_{ij} - \tilde{X}_i |$$

  - $\tilde{X}_i$: Mediaan van group $i$

- Deze test is meer robuust voor niet-normaliteit in vergelijking met de **Levene's test**

**F test:**

- **Hypothese:** $H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_a : \sigma_1^2 \neq \sigma_2^2$
- **Test statistiek:** F-test

$$S = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F_{n_1-1,n_2-1}$$

- Generalisatie naar meer dan 2 groepen heet **Hartley's F-test**
- Houd er rekening mee dat deze test extreem gevoelig is voor niet-normaliteit!

# 6.5 Alternatieve methodes

- In het geval dat aannames niet voldaan zijn, kan men vertrouwen op **twee oplossingen**, d.w.z.,
  - **Transformaties**, zodat aan de aannames wordt voldaan
  - **Niet-parametrische versies van de testen:**

| Parametric test | Non-parametric alternative |
|---|---|
| One sample t-test | Binomial test (sign test)<br>Wilcoxon signed-rank test |
| Paired t-test | Wilcoxon signed-rank test on differences |
| Two sample t-test | Wilcoxon rank-sum test<br>=<br>Mann-Whitney U test |
| One-way ANOVA | Kruskall Wallis test |