# Het toetsen van hypothesen

Sandra Van Aert

17 november 2011

# Toetsen van kansverdelingen

- is aantal voorbijrijdende auto's X op een snelweg Poisson verdeeld met parameter $\lambda = 2$ (per tijdsinterval van 10 sec)?
- $H_0$: X $\sim$ Poisson($\lambda = 2$) $H_a$: X $\nsim$ Poisson($\lambda = 2$)

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | ≥6 |
|---|---|---|---|---|---|---|---|
| $O_i$ | 19 | 38 | 28 | 20 | 7 | 4 | 4 |
| kans | 0.135 | 0.271 | 0.271 | 0.180 | 0.090 | 0.036 | 0.017 |
| $E_i$ | 16.24 | 32.48 | 32.48 | 21.65 | 10.83 | 4.33 | 1.99 |

- $E_i = 120\times$ kans $p_X(x; 2)$
- geobserveerde frequentie $O_i$ en theoretische frequentie $E_i$ vergelijken
- verschillen $O_i - E_i$ berekenen
- weging: $\dfrac{(O_i - E_i)^2}{E_i}$

# $\chi^2$-hypothesetoets

- indien $H_0$ waar is, dan is toetsingsgrootheid

$$\chi = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  $\chi^2_{k-1}$ verdeeld waarbij $k = \#$ uitkomsten
- indien $H_0$ waar is, dan liggen alle $O_i$ dicht bij $E_i$
- indien $H_0$ waar is, is toetsingsgrootheid niet veel groter dan 0
- berekende toetsingsgrootheid vergelijken met $\chi^2_{\alpha;k-1}$

# Voorbeeld

- $\alpha = 0.05$
- berekende toetsingsgrootheid

$$x = \frac{(19-16.24)^2}{16.24} + \frac{(38-32.48)^2}{32.48} + \cdots + \frac{(4-1.99)^2}{1.99}$$
$$= 5.57$$

- $\chi^2_{0.05;6} = 12.59$
  vermits $5.57 < 12.59$ wordt $H_0$ aanvaard
- $p$-waarde $= P(\chi^2_6 > 5.57) = 0.473$
  $p > \alpha \Rightarrow H_0$ aanvaarden
- probleem: bepaling van $\lambda$

# Vervolg voorbeeld

- $\lambda$ schatten op basis van steekproefgegevens:

$$\text{schatting voor } \lambda : \qquad \overline{x} = 1.88$$

- toetsingsgrootheid nog steeds $\chi^2$ verdeeld, maar nu slechts $k-2$ vrijheidsgraden
- we verliezen één vrijheidsgraad omdat we hier één parameter moeten schatten

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|---|
| $O_i$ | 19 | 38 | 28 | 20 | 7 | 4 | 4 |
| $p_X(x)$ | 0.152 | 0.287 | 0.270 | 0.169 | 0.079 | 0.030 | 0.013 |
| $E_i$ | 18.31 | 34.42 | 32.36 | 20.28 | 9.53 | 3.58 | 1.51 |

# Vervolg voorbeeld

- berekende toetsingsgrootheid $x = 5.79$
- $\chi^2_{0.05;5} = 11.07$
  vermits $5.79 < 11.07$ wordt $H_0$ aanvaard
- $p$-waarde $= P(\chi^2_5 > 5.79) = 0.3272$
  $p > \alpha \Rightarrow H_0$ aanvaarden

# Toetsen van kansdichtheden

- klemtoon op normale kansdichtheid
- grafische aanpak

- optie 1:
  → histogram
  → stam- en bladdiagram
  → polygoon

- optie 2:
  → kwantieldiagram of $Q-Q$–plot:
  theoretische en geobserveerde kwantielen
  of percentielen vergelijken

# Voorbeeld kwantieldiagram

- komen data uit normale verdeling?
- $n = 10$, $\overline{x} = 49.8$ en $s = 12.4$
- $61, 50, 26, 47, 38, 46, 60, 65, 62, 43$
- rangschikken: $26, 38, 43, 46, 47, 50, 60, 61, 62, 65$

  $\uparrow$ 10de kwantiel $\uparrow$ 20ste … $\uparrow$ 90ste
- $26 \approx$ 5de kwantiel $\rightarrow \mathrm{cf}_1$
- $38 \approx$ 15de kwantiel $\rightarrow \mathrm{cf}_2$
- …
- $\underbrace{65 \approx \text{95ste kwantiel}}_{\text{empirische of geobserveerde kwantielen}} \rightarrow \mathrm{cf}_{10}$

# Vervolg voorbeeld

- cumulatieve frequenties:

$$\mathrm{cf}_j = \frac{j - 0.5}{n} \qquad j = 1, \ldots, n$$

- theoretisch: 5de kwantiel
  vertrek van $z_{0.95}$
  vertaal naar $N(\overline{x}, s^2) = N(49.8; (12.4)^2)$

$$\overline{x} + z_{0.95} s = 49.8 + z_{0.95} \times 12.4 = 29.4$$

- theoretisch: 15de kwantiel

$$\overline{x} + z_{0.85} s = 49.8 + z_{0.85} \times 12.4 = 36.95$$

- algemeen:

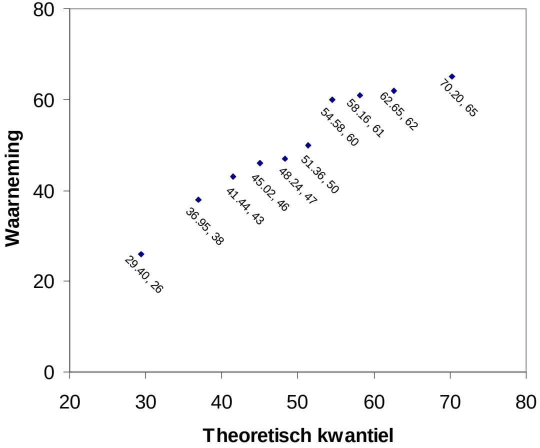
$$\overline{x} + z_{1-\mathrm{cf}_i} s$$

# Constructie kwantieldiagram

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_i$ | 26 | 38 | 43 | $\ldots$ | 62 | 65 |
| $cf_i$ | 0.05 | 0.15 | 0.25 | $\ldots$ | 0.85 | 0.95 |
| $z_{1-cf_i}$ | -1.64 | -1.04 | -0.67 | $\ldots$ | 1.04 | 1.64 |
| $49.8 + 12.4 z_{1-cf_i}$ | 29.40 | 36.95 | 41.44 | $\ldots$ | 62.65 | 70.20 |

# Elementair kwantieldiagram

$X$-as: theoretische kwantielen
$Y$-as: geobserveerde kwantielen



# Verbeterde constructie van een kwantieldiagram

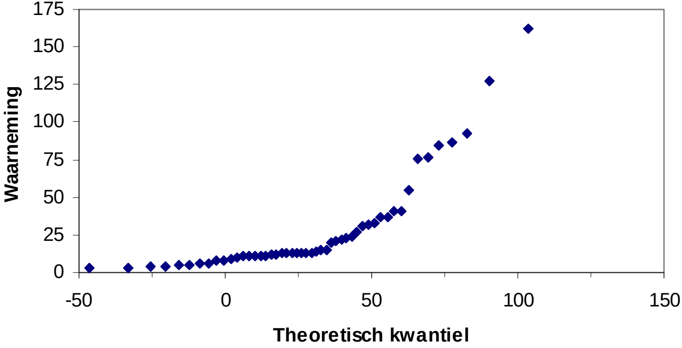
beter om

$$cf_j^* = \frac{j - 0.375}{n + 0.25} \qquad j = 1, \ldots, n$$

te gebruiken

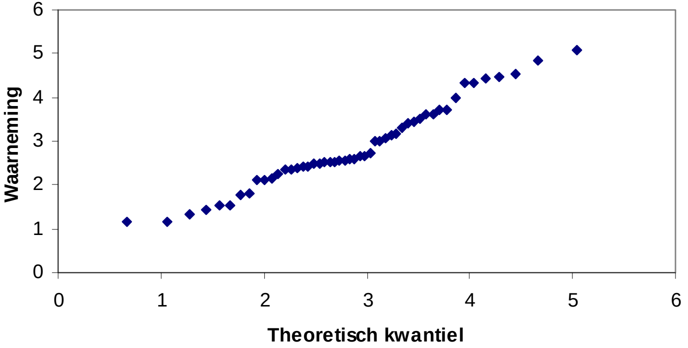
| $x_i$ | 26 | 38 | 43 | 46 | $\ldots$ | 65 |
|---|---|---|---|---|---|---|
| $cf_i^*$ | 0.061 | 0.159 | 0.256 | 0.354 | $\ldots$ | 0.9390 |
| $z_{1-cf_i^*}$ | -1.55 | -1.00 | -0.66 | -0.38 | $\ldots$ | 1.55 |
| $49.8 + 12.4 z_{1-cf_i}$ | 30.62 | 37.39 | 41.67 | 45.14 | $\ldots$ | 68.98 |

# Voorbeeld breeksterktes




Kwantieldiagram “breeksterkte”

# Voorbeeld breeksterktes




Kwantieldiagram “ln(breeksterkte)”

# Toetsen van kansdichtheden

- klemtoon op normale kansdichtheid
- formele hypothesetoets
- optie 1: Lillieforstoets
  → ook Kolmogorov-Smirnov-toets genoemd
  → vergelijkt theoretische cumulatieve verdelingsfunctie met empirische
- optie 2: Shapiro-Wilk toets
  → gebruikt de correlatiecoëfficiënt van het Q-Q-diagram om na te gaan of de gegevens al dan niet normaal verdeeld zijn

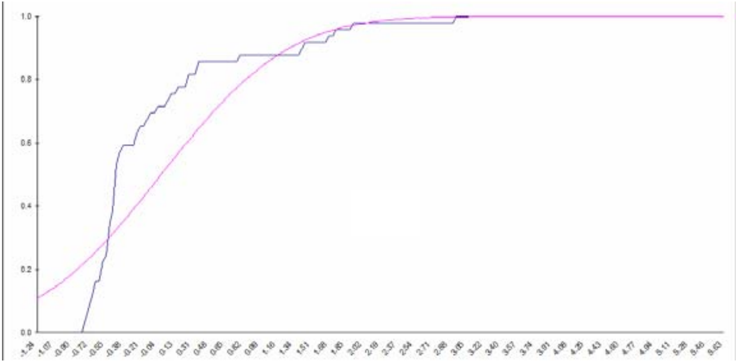
# Lillieforstoets

- $S_X(x)$ = empirische, relatieve cumulatieve frequentie van $x$
- $F_X(x)$ = theoretische cumulatieve verdelingsfunctie in $x$
- toetsingsgrootheid

  $$D = \max_x |F_X(x) - S_X(x)|$$

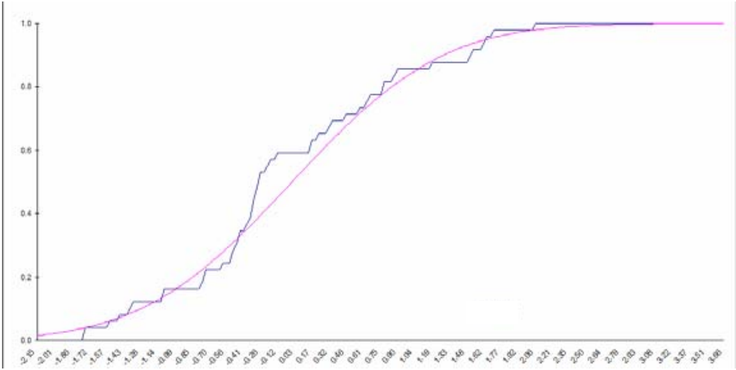
  dit is de maximale verticale afstand tussen de grafieken van $F_X(x)$ en $S_X(x)$
- berekende toetsingsgrootheid $d$ vergelijken met kritieke waarde
- $p$-waarde = $P(D > d)$ vergelijken met $\alpha$

# Lillieforstoets




Theoretische en empirische cumulatieve verdelingsfunctie voor de variabele “breeksterkte”

# Lillieforstoets




Theoretische en empirische cumulatieve verdelingsfunctie voor de variabele "ln(breeksterkte)"

# Hypothesetoetsen voor twee populaties

Sandra Van Aert

17 november 2011

# Inleiding

- doel: gelijkenissen/verschillen tussen 2 populaties ontdekken en meten
- meest algemeen: cumulatieve verdelingsfuncties vergelijken

$$H_0 : F_1 = F_2 \text{ versus } H_a : F_1 \neq F_2$$

- meestal beperkt men zich tot het vergelijken van gemiddeldes, varianties, medianen, proporties,…

# Voorbeelden

- populatiegemiddeldes

$$\begin{aligned} H_0 : \mu_1 = \mu_2 \text{ versus } H_a &: \mu_1 > \mu_2 \\ H_a &: \mu_1 < \mu_2 \\ H_a &: \mu_1 \neq \mu_2 \end{aligned}$$

- populatievarianties

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_a : \sigma_1^2 \begin{smallmatrix} < \\ > \\ \neq \end{smallmatrix} \sigma_2^2$$

- populatieproporties

$$H_0 : \pi_1 = \pi_2 \text{ versus } H_a : \pi_1 \begin{smallmatrix} < \\ > \\ \neq \end{smallmatrix} \pi_2$$

- ...

# Belangrijk

gepaarde waarnemingen

of afhankelijke steekproeven

$\updownarrow$

niet-gepaarde waarnemingen

of onafhankelijke steekproeven

# Voorbeelden

- vergelijken van 2 types zolen (slijtvastheid)
  → 20 proefpersonen
  → 10 testen zool A
  → andere 10 testen zool B
  → onafhankelijke steekproeven
- vergelijken van 2 types zolen
  → 20 proefpersonen
  → iedere proefpersoon
  krijgt 1 schoen met zool A
  1 schoen met zool B
  → gepaarde waarnemingen

# Voorbeelden (vervolg)

- vergelijk aantal bacteriën in melk vóór opwarming en na opwarming
  → 10 melkstalen
  → meet aantal bacteriën in elk staal
  → warm elk staal op
  → meet aantal bacteriën opnieuw
  → gepaarde waarnemingen

# Twee populatiegemiddeldes - onafhankelijk

- algemeen

$$
\begin{aligned}
H_0: \mu_1 = \mu_2 \text{ versus } H_a &: \mu_1 > \mu_2 && \text{rechtseenzijdig}\\
H_a &: \mu_1 < \mu_2 && \text{linkseenzijdig}\\
H_a &: \mu_1 \neq \mu_2 && \text{tweezijdig}
\end{aligned}
$$

- herschrijven als

$$
\begin{aligned}
H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_a &: \mu_1 - \mu_2 > 0 && \text{rechts}\\
H_a &: \mu_1 - \mu_2 < 0 && \text{links}\\
H_a &: \mu_1 - \mu_2 \neq 0 && \text{tweezijdig}
\end{aligned}
$$

# Twee populatiegemiddeldes (vervolg)

- voor de twee steekproeven bepalen we steekproefgemiddelde: $\overline{x}_1$ en $\overline{x}_2$
$\rightarrow \overline{X}_1$ en $\overline{X}_2$
- $E(\overline{X}_1) = \mu_1$ en $E(\overline{X}_2) = \mu_2$
$\mathrm{E}(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$
- $\mathrm{var}(\overline{X}_1) = \frac{\sigma_1^2}{n_1}$ en $\mathrm{var}(\overline{X}_2) = \frac{\sigma_2^2}{n_2}$
stelling 10.2 $\Rightarrow \mathrm{var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

# Overzicht

- onderscheid
  - kleine steekproeven
    - normaal verdeeld
      - gelijke varianties 15.1.1
      - ongelijke varianties 15.1.2
    - niet-normaal verdeeld 15.5.1
  - grote steekproeven

# Kleine steekproeven, gelijke varianties, normaal verdeelde gegevens (15.1.1)

- $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$
- $$\begin{aligned}\mathrm{var}(\overline{X}_1 - \overline{X}_2) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &\qquad\downarrow \sigma_1^2 = \sigma_2^2 = \sigma^2 \\ &= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\end{aligned}$$

# Kansdichtheid $\overline{X}_1 - \overline{X}_2$

- data steekproef 1: $X_{11}, X_{12}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma^2)$

$$\Downarrow$$

$$\overline{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n} \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

- data steekproef 2: $X_{21}, X_{22}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma^2)$

$$\Downarrow$$

$$\overline{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n} \sim N(\mu_2, \frac{\sigma^2}{n_2})$$

**gevolg**: $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$

# Toetsingsgrootheid

- als nulhypothese $H_0 : \mu_1 = \mu_2$ of $H_0 : \mu_1 - \mu_2 = 0$ waar is, dan geldt

$$\overline{X}_1 - \overline{X}_2 \sim N(0, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

en dus

$$\frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0, 1)$$

- probleem: $\sigma^2 = ?$

  beide steekproeven bevatten informatie over $\sigma^2$

# Gepoolde variantie

- steekproefvariantie steekproef 1: $S_1^2$
- steekproefvariantie steekproef 2: $S_2^2$
- $S_1^2$ en $S_2^2$ combineren, rekening houdend met steekproefgroottes: gepoolde schatter

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

- $$\underbrace{\frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}}_{\downarrow} = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

toetsingsgrootheid $T \sim t_{n_1+n_2-2}$

# Rechtseenzijdige toets

- $H_a : \mu_1 > \mu_2$
  $H_a : \mu_1 - \mu_2 > 0$
- beslissingsregel 1:
  berekende toetsingsgrootheid $t = \dfrac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
  $t > t_{\alpha;n_1+n_2-2}$: verwerp $H_0$
  $t \leq t_{\alpha;n_1+n_2-2}$: aanvaard $H_0$
- beslissingsregel 2:
  $$p = P(t_{n_1+n_2-2} > \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}) = P(t_{n_1+n_2-2} > t)$$
  $p < \alpha$: verwerp $H_0$
  $p \geq \alpha$: aanvaard $H_0$

# Linkseenzijdige toets

- $H_a: \mu_1 < \mu_2$
  $H_a: \mu_1 - \mu_2 < 0$
- beslissingsregel 1:
  berekende toetsingsgrootheid $t = \dfrac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

  $t < t_{1-\alpha;n_1+n_2-2} = -t_{\alpha;n_1+n_2-2}$: verwerp $H_0$
  $t \geq -t_{\alpha;n_1+n_2-2}$: aanvaard $H_0$
- beslissingsregel 2:
  $p = P(t_{n_1+n_2-2} < t)$

  $p < \alpha$: verwerp $H_0$
  $p \geq \alpha$: aanvaard $H_0$

# Tweezijdige toets

- $H_a: \mu_1 \neq \mu_2$
  $H_a: \mu_1 - \mu_2 \neq 0$
- beslissingsregel 1:

  berekende toetsingsgrootheid $t = \dfrac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

  $t < -t_{\alpha/2;n_1+n_2-2}$ of $t > t_{\alpha/2;n_1+n_2-2}$: verwerp $H_0$
  $-t_{\alpha/2;n_1+n_2-2} \leq t \leq t_{\alpha/2;n_1+n_2-2}$: aanvaard $H_0$
- beslissingsregel 2:
  $p = 2P(t_{n_1+n_2-2} > |t|)$

  $p < \alpha$: verwerp $H_0$
  $p \geq \alpha$: aanvaard $H_0$

# Voorbeeld

- $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$
- 2 machines → identieke kabels
- normaal verdeelde populaties
- gelijke populatievarianties

| steekproef 1 | steekproef 2 |
|---|---|
| $n_1 = 20$ | $n_2 = 20$ |
| $\overline{x}_1 = 50.9033$ | $\overline{x}_2 = 50.2525$ |
| $s_1^2 = 0.1848$ | $s_2^2 = 0.2837$ |

$$\Rightarrow s_p^2 = \frac{(20-1)0.1848 + (20-1)0.2837}{20+20-2} = 0.2342$$

$$\Rightarrow s_p = 0.4840$$

# Vervolg voorbeeld

- berekende toetsingsgrootheid

$$t = \frac{50.9033 - 50.2525}{0.4840\sqrt{\frac{1}{20}+\frac{1}{20}}} = 4.2521$$

- $p$-waarde

$$p = 2P(t_{20+20-2} > t) = 2P(t_{38} > 4.2521) = 0.000133$$

$$p < \alpha = 5\% \Rightarrow \text{verwerp } H_0$$

- kritieke waarde

$$t_{0.025;38} = 2.0244$$

$$t_{0.025;38} < t = 4.2521 \Rightarrow \text{verwerp } H_0$$

# Kleine steekproeven, ongelijke varianties, normaal verdeelde gegevens (15.1.2)

- $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$
- $\text{var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$
- $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$

**gevolg**: $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

# Kansdichtheid

- als $H_0 : \mu_1 = \mu_2$ of $H_0 : \mu_1 - \mu_2 = 0$ waar is, dan

$$\overline{X}_1 - \overline{X}_2 \sim N(\color{red}{0}, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

  en dus

$$\frac{\overline{X}_1 - \overline{X}_2 \color{red}{-0}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

- probleem: $\underset{\underset{\color{blue}{S_1^2}}{\color{blue}{\uparrow}}}{\sigma_1^2}$ ?, $\underset{\underset{\color{blue}{S_2^2}}{\color{blue}{\uparrow}}}{\sigma_2^2}$ ?

# Toetsingsgrootheid

- $$\frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \cancel{t_{n_1+n_2-2}}$$

$$\sim \text{ Behrens-Fisher}$$

ongeveer $\sim t_\nu$

met $\nu$ geheel getal dichtst bij $\dfrac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$

# Rechtseenzijdige toets

- $H_a: \mu_1 > \mu_2$
  $H_a: \mu_1 - \mu_2 > 0$
- beslissingsregel 1:
  berekende toetsingsgrootheid $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

  $t > t_{\alpha;\nu}$: verwerp $H_0$
  $t \leq t_{\alpha;\nu}$: aanvaard $H_0$
- beslissingsregel 2:
  $p = P(t_\nu > t)$

  $p < \alpha$: verwerp $H_0$
  $p \geq \alpha$: aanvaard $H_0$

# Andere hypothesen

- linkseenzijdige toets
- tweezijdige toets verlopen analoog

# Grote steekproeven (15.1.3)

$$\frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

ongeacht of $\sigma_1^2$ en $\sigma_2^2$ gelijk zijn of niet

# Gekende varianties (15.1.4)

$$\frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\text{bij} \begin{cases} \text{normaal verdeelde populatie} \\ \text{grote steekproeven} \end{cases}$$