

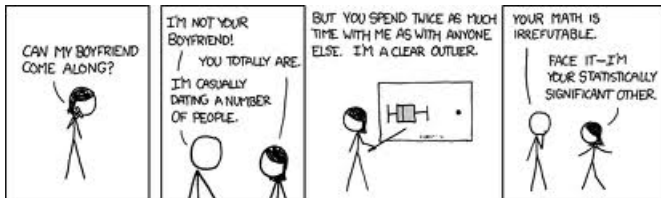
# Statistiek voor A.I.

College 13

Donderdag 25 Oktober

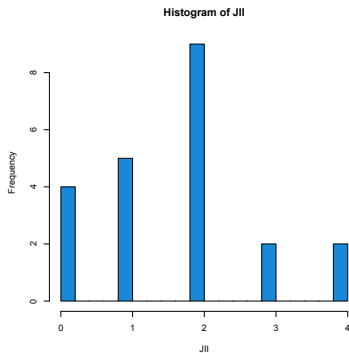
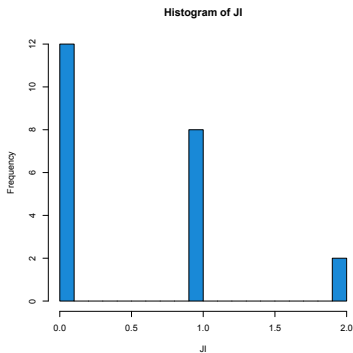
## 2 Deductieve statistiek

Orthodoxe statistiek



Links: Aantal broers/zussen van het andere geslacht.

Rechts: Aantal relaties tot nu toe.

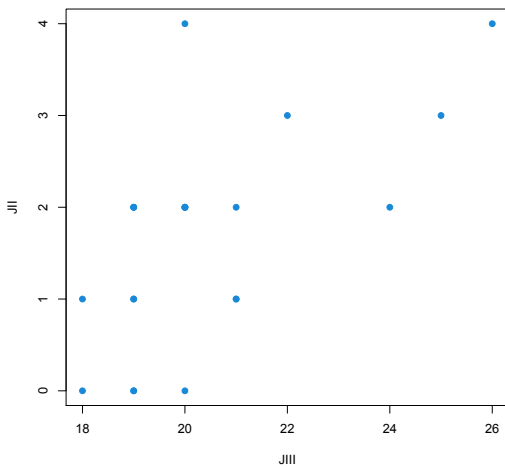


Geen noemenswaardige correlatie tussen deze twee variabelen.

## Jullie - onderzoek Tobias, Lody, Swen en Sander

Horizontaal: Leeftijd.

Verticaal: Aantal relaties tot nu toe.

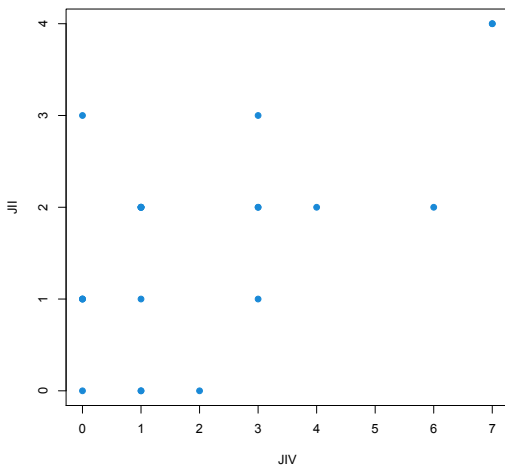


Correlatiecoëfficiënt: 0.61.

## Jullie - onderzoek Tobias, Lody, Swen en Sander

Horizontaal: Aantal keer verhuisd.

Verticaal: Aantal relaties tot nu toe.



Correlatiecoëfficiënt: 0.62.

## 2 Deductieve statistiek

Vandaag :

- Hypothese toetsen
- Schatten

## Hypothese toetsen

### Voorbeeld

- Een medicijn wordt aan een groep patiënten toegediend en aan een controlegroep niet. Er moet besloten worden of het medicijn werkt of niet.
- Men wil vaststellen of een zekere ziekte meer voorkomt onder Aziaten dan Afrikanen.
- Men wil vaststellen of het percentage defecte stofzuigers bij de ene fabriek groter is dan bij de andere.

## Hypothese toetsen: twee steekproeven

**Definitie** Een test naar de gelijkheid van de gemiddelden van twee populaties  $P_0$  en  $P_1$  heeft de volgende vorm.

$\mu_1$  en  $\mu_2$  zijn de gemiddelden van de populaties en  $\sigma_1^2$  en  $\sigma_2^2$  de varianties.

$\bar{X}_1, \bar{X}_2$  zijn de gemiddelden van de twee steekproeven en  $s_1^2$  en  $s_2^2$  de varianties.

De aanname dat  $\sigma_1^2 = \sigma_2^2$  (die wel of niet gehanteerd kan worden) is de *homogeniteit van variantie*.

De methode is hetzelfde als die voor het toetsen met 1 gemiddelde, alleen de verdelingen en hypothesen zelf hebben een andere vorm.

**Definitie** De hypothesen zijn van de vorm

$$\begin{array}{ll} \text{tweezijdig} & H_0 : \mu_1 - \mu_2 = 0 \\ & H_a : \mu_1 \neq \mu_2 \end{array}$$

$$\begin{array}{ll} \text{linkszijdig} & H_0 : \mu_1 - \mu_2 \geq 0 \\ & H_a : \mu_1 < \mu_2 \end{array}$$

$$\begin{array}{ll} \text{rechtszijdig} & H_0 : \mu_1 - \mu_2 \leq 0 \\ & H_a : \mu_1 > \mu_2 \end{array}$$

De statistiek is  $\bar{X}_1 - \bar{X}_2$ .

## Hypothese toetsen: twee steekproeven

**Definitie** Bij hypothese toetsen over het gemiddelden van twee populaties waarvan de varianties  $\sigma^2$  bekend en gelijk zijn, wordt de verdeling van  $\bar{X}_1 - \bar{X}_2$  benaderd door een normale verdeling met gemiddelde

$$\mu_1 - \mu_2$$

en standaardafwijking

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

$\sigma_{\bar{X}_1 - \bar{X}_2}$  is de *standaardfout van het verschil*.

Wanneer de varianties gelijk maar onbekend zijn, wordt de verdeling van  $\bar{X}_1 - \bar{X}_2$  benaderd door een  $t$ -verdeling bij  $n_1 + n_2 - 2$  vrijheidsgraden en gemiddelde

$$\mu_1 - \mu_2,$$

en standaardafwijking

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)},$$

waarbij  $s^2$  de *gezamenlijke schatting van de populatievariantie* is:

$$\frac{\sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{j2} - \bar{X}_2)^2}{n_1 + n_2 - 2}.$$

## Hypothese toetsen

**Voorbeeld 44** mannelijke en 44 vrouwelijke apen werd een pop en een speelgoedauto aangeboden.  $\bar{X}_1$  en  $\bar{X}_2$  zijn het percentage van de tijd dat de mannelijke en vrouwelijke apen met de pop speelden. De gevonden waarden zijn  $\bar{X}_1 = 9$  en  $\bar{X}_2 = 11$ .

De twee populaties zijn alle mannelijke en alle vrouwelijke apen op aarde.  $\mu_1$  en  $\mu_2$  zijn het percentage van de tijd dat alle mannelijke en vrouwelijke apen met de pop zouden spelen. Stel dat de variantie in speeltijd voor beide populaties bekend en gelijk aan 16% is.

Er wordt linkszijdig getoetst:  $H_0 : \mu_1 - \mu_2 \geq 0$      $H_a : \mu_1 < \mu_2$ . Er geldt

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{16\left(\frac{1}{44} + \frac{1}{44}\right)} = \sqrt{\frac{16 \times 2}{44}} = 0.85.$$

Bij significantieniveau  $\alpha = 0.1$  is de standaard kritische waarde  $z_{kw} = -1.282$ , want  $P_s(z \leq -1.282) = 0.1$ .

De kritische waarde is dus  $(\mu_1 - \mu_2) - 1.282\sigma_{\bar{X}_1 - \bar{X}_2} = -1.282 \times 0.85 = -1.1$ .

Het kritische gebied is  $(-\infty, -1.1]$  en de  $p$ -waarde van de statistiek is  $P(\bar{X}_1 - \bar{X}_2 \leq -2 \mid H_0) = P_s(z \leq -2.35) = 0.0094 < \alpha$ .

$H_0$  wordt verworpen: mannelijke apen spelen minder met poppen dan vrouwelijke apen.

## Hypothese toetsen: vrouwelijk of mannelijk

**Voorbeeld** Voelen mannelijke studenten KI aan de UU zich significant mannelijker dan vrouwelijke studenten zich vrouwelijk voelen (0=niet,5=zeer)? Als significantieniveau wordt 0.005 genomen. Aangenomen wordt dat de varianties van de populaties gelijk maar onbekend zijn. De hypothesen vormen een rechtszijdige toets:  $H_0 : \mu_1 - \mu_2 \leq 0$  en  $H_a : \mu_1 > \mu_2$ .

Er worden door Kimberly, Josca en Niels steekproeven  $X_1$  uit de mannelijke en  $X_2$  uit de vrouwelijke studenten genomen, beide ter grootte 9.  $\bar{X}_1 = 3.11$  en  $\bar{X}_2 = 2.22$ . De statistiek is daarmee  $3.11 - 2.22 = 0.98$ .

Er wordt getoetst met een t-verdeling met 16 vrijheidsgraden:  $t_{kw} = 2.921$ . De gezamenlijke schatting van de populatievariantie is:

$$s^2 = \frac{\sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{j2} - \bar{X}_2)^2}{n_1 + n_2 - 2} = 0.65.$$

Dat geeft

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 0.38.$$

De kritische waarde is  $x_{kw} = 2.921 \times 0.38 = 1.11$  en  $H_0$  wordt niet verworpen.

Schatten

**STATISTICS  
MEAN NEVER  
HAVING TO  
SAY YOU'RE  
CERTAIN**

## Vragen: liegen

61 Amerikanen werd gevraagd hoeveel % van de tijd zij liegen. Het gevonden gemiddelde was 0.4%.

Er werd hen ook gevraagd hoeveel % van de tijd zij liegen om iemands gevoelens te sparen. Het gevonden gemiddelde was 1%.

Zijn deze data consistent?

Statistisch schatten wordt toegepast wanneer men de waarde van een parameter wil benaderen.

Op grond van de statistiek van een steekproef wordt de corresponderende waarde van de populatie geschat/benaderd.

Sommige schatters produceren meerdere waarden (vaak een interval) en sommige schatters produceren één waarde.

Vraag: Hoe accuraat is de schatter?

## Schatten: betrouwbaarheidsinterval

**Definitie** Wanneer de standaardafwijking van de populatie  $\sigma$  bekend is dan is het *betrouwbaarheidsinterval* met *betrouwbaarheidsniveau*  $(1 - \alpha)$  op grond van een statistiek  $\bar{X}$  van een steekproef ter grootte  $n$ :

$$CI = (\bar{X} - z_{kw}\sigma_{\bar{X}}, \bar{X} + z_{kw}\sigma_{\bar{X}}),$$

waarbij  $z_{kw}$  de standaard kritische waarde bij significantieniveau  $\alpha$  voor een tweezijdige toets op grond van de normale verdeling is.

Wanneer de standaardafwijking van de populatie onbekend is dan is het betrouwbaarheidsinterval

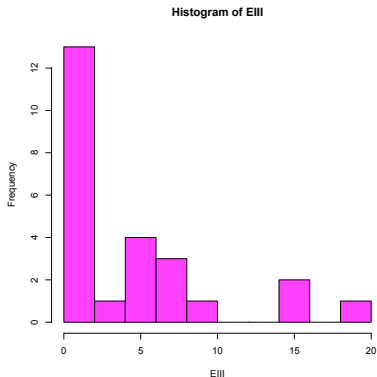
$$CI = (\bar{X} - t_{kw}s_{\bar{X}}, \bar{X} + t_{kw}s_{\bar{X}}).$$

waarbij  $t_{kw}$  de standaard kritische waarde bij significantieniveau  $\alpha$  voor een tweezijdig toets op grond van de  $t$ -verdeling bij  $(n - 1)$  vrijheidsgraden is.

$CI$  wordt ook wel het  $(1 - \alpha)100\%$ -betrouwbaarheidsinterval genoemd.

## Schatten

**Voorbeeld** Bjorn, Alex en Rianne vonden in een steekproef ter grootte 25 een gemiddelde van 4.680 uren gamen per week gedurende collegeweken, waarbij  $s = 5.528$ .



Het 95%-betrouwbaarheidsinterval is

$$\left(4.680 - 2.064 \times \frac{5.528}{5}, 4.680 + 2.064 \times \frac{5.528}{5}\right) = (2.398, 6.962).$$

**Stelling** Voor alle steekproeven van een vaste grootte uit een populatie met gemiddelde  $\mu$  waarvan het steekproefgemiddelde normaal verdeeld is, geldt dat  $100(1 - x)\%$  van de  $100(1 - x)\%$ -betrouwbaarheidsintervallen  $\mu$  bevat en  $100x\%$  van de intervallen bevat  $\mu$  niet.

**Bewijs**  $\mu$  ligt in het  $100(1 - x)\%$ -betrouwbaarheidsinterval rond  $\bar{X}$  als geldt dat

$$\mu - z_{kw}\sigma_{\bar{X}} < \bar{X} < \mu + z_{kw}\sigma_{\bar{X}}. \quad (1)$$

Voor  $z_{kw}$  geldt dat  $P_s(z \geq z_{kw}) = \frac{x}{2}$ . Omdat  $\bar{X}$  normaal verdeeld is met gemiddelde  $\mu$  en standaardafwijking  $\sigma_{\bar{X}}$  volgt daaruit dat de kans op (1) gelijk aan  $(1 - x)$  is.

**Merk op:** Voordat de steekproef heeft plaatsgehad is de kans dat het  $100(1 - x)\%$ -betrouwbaarheidsinterval  $\mu$  bevat  $(1 - x)$ . Nadat de steekproef heeft plaatsgehad, is die kans 0 of 1: het betrouwbaarheidsinterval bevat  $\mu$  niet of wel.

## Antwoord op een vraag: liegen

**Voorbeeld** Aan 61 Amerikanen werd gevraagd hoeveel % van de tijd zij liegen. Het gevonden gemiddelde was 0.4%. Aan een andere groep van 61 Amerikanen werd gevraagd hoeveel % van de tijd zij liegen om iemands gevoelens te sparen. Het gevonden gemiddelde was 1%.

Zijn die data 95% betrouwbaar, d.w.z. kunnen beide 95%-betrouwbaarheidsintervallen de ware gemiddeldes bevatten?

Een 95% betrouwbaarheidsinterval op grond van de eerste steekproef, waarbij gegeven is dat  $s_{\bar{X}} = 0.0015$ , is

$$CI_1 = (\bar{X} - t_{kw}s_{\bar{X}}, \bar{X} + t_{kw}s_{\bar{X}}) = (0.001, 0.007).$$

Een 95% betrouwbaarheidsinterval op grond van de tweede steekproef, waarbij gegeven is dat  $s_{\bar{X}} = 0.0013$ , is

$$CI_2 = (\bar{X} - t_{kw}s_{\bar{X}}, \bar{X} + t_{kw}s_{\bar{X}}) = (0.0074, 0.0126).$$

Stel, Amerikanen liegen in werkelijkheid  $\mu_1\%$  van de tijd en liegen  $\mu_2\%$  van de tijd om iemands gevoelens te sparen. Dan moet gelden  $\mu_1 \geq \mu_2$ .

Omdat  $CI_2$  rechts van  $CI_1$  ligt kan het niet zo zijn dat beide intervallen de werkelijke gemiddeldes bevatten ( $CI_i$  bevat  $\mu_i$ ).

Het onderzoek is niet 95% betrouwbaar.

**Definitie** Als een parameter  $\theta$  van een populatie benaderd/geschat wordt door een statistiek  $t$ , dan is  $t$  een *zuivere* schatter als

$$E(t) = \theta.$$

Hierbij is  $E(t)$  de verwachtingswaarde van  $t$  op grond van alle steekproeven van een vaste grootte uit de populatie.

## Schatten: zuivere schatters

**Stelling** Zij  $X$  een stochast met verwachtingswaarde/gemiddelde  $\mu$  en variantie  $\sigma^2$ . Dan geldt wanneer het bijbehorende experiment  $n$  maal herhaald wordt dat

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

een zuivere schatter van  $\mu$  en

$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

een zuivere schatter van  $\sigma^2$  is.

### Bewijs

Gebruik dat voor alle onafhankelijke stochasten  $X$  en  $Y$  en elke constante  $c$  geldt:  $E(cX + Y) = cE(X) + E(Y)$  en  $\text{Var}(cX + Y) = c^2\text{Var}(X) + \text{Var}(Y)$ .

Voor de variantie:

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

$$E(s_{\bar{X}}^2) = E\left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}\right) = E\left(\sum_{i=1}^n \frac{(X_i - \mu)^2}{n-1}\right) - E\left(\frac{n(\bar{X} - \mu)^2}{n-1}\right) = \frac{n\sigma^2}{n-1} - \frac{\sigma^2}{n-1} = \sigma^2.$$

**Definitie** Zij  $\Omega$  de verzameling van mogelijke waarden van een populatie parameter  $\theta$ . Gegeven een steekproef met bijbehorende statistiek  $t$  is de *likelihood* van  $t$  gegeven  $\theta \in \Omega$ :  $P(t | \theta)$ .

De *maximum likelihood schatter m.l.e.* is de schatter die op grond van de steekproef die  $\theta$  in  $\Omega$  kiest waarvoor  $t$  de grootste likelihood heeft. De schatter wordt aangeduid met  $m(t)$ :

$$m(t) = \operatorname{argmax}_{\tau \in \Omega} P(t | \tau)$$

Dus  $m(t) = \theta$  impliceert  $P(t | \theta) = \max_{\tau \in \Omega} P(t | \tau)$ .

**Merk op:** Een m.l.e. hoeft niet uniek te zijn.

**Stelling** Zij  $X$  de stochast behorende bij een Bernoulli experiment met kans op succes  $p$ . Dan geldt wanneer het bijbehorende experiment  $n$  maal herhaald wordt, dat het steekproefgemiddelde  $\bar{X}$  ( $\frac{\text{aantal maal succes}}{\text{totaal}}$ ) de m.l.e. van  $p$  is:

$$m(X_1 \dots X_n) = \bar{X}.$$

**Stelling** Zij  $X$  een normaal verdeelde stochast met gemiddelde  $\mu$  en variantie  $\sigma^2$ . Dan geldt wanneer het bijbehorende experiment  $n$  maal herhaald wordt dat het steekproefgemiddelde  $\bar{X}$  de m.l.e. van  $\mu$  en

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

de m.l.e. van  $\sigma^2$  is:

$$m(X_1 \dots X_n) = \bar{X} \quad m(X_1 \dots X_n) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

Een zuivere schatter is niet noodzakelijk gelijk aan een m.l.e.: de zuivere schatter en de m.l.e. van de variantie van de populatie zijn respectievelijk

$$\frac{\sum_{i=1}^n (\bar{X}_i - \bar{X})^2}{n-1} \quad \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

De eis van zuiverheid en de eis van maximum likelihood van een schatter ondersteunen de keuze om het gemiddelde van een populatie te benaderen met het steekproefgemiddelde op *verschillende* wijze.

Finis