Introductie tot de statistiek

Hogeschool Gent

04/05/2010

Inhoudsopgave

1 Basisbegrippen en beschrijvende statistiek

1.1 Onderzoek

Data verzamelen in een specifieke steekproef, representatief voor de populatie.

1.1.1 Data

- Data: p variabelen bij n observaties.
- Voorbeeld:

score	iq	motivatie	geslacht	werken
16	140	5	M	Neen
10	120	2	V	Ja
11	125	3	M	Ja
14	135	7	V	Neen
8	115	2	M	Neen
18	145	5	V	Neen
13	140	6	M	Ja
9	125	4	V	Neen
11	130	3	V	Neen
10	125	1	V	Neen

1.1.2 Variabelen

- Eigenschap die varieert: X
- scores zijn geobserveerde waarden van een variabele: x, vb. $x_2 = 10$

1.1.3 Meetniveau

- Categorische variabelen: nominaal of ordinaal (vb geslacht)
- Continue variabelen: minstens interval niveau (vb iq)
- Opm. Likert-schaal: ordinaal, maar als continue beschouwd.

1.2 Beschrijvende technieken: 1 variabele

1.2.1 Ordeningstechnieken

- frequentietabel

geslacht	freq.
M	4
V	6

- relatieve frequentieverdeling

geslacht	rel. freq.
M	0.4
V	0.6

- gegroepeerde frequentieverdeling

score	freq.
0-9	2
10-11	4
12-20	4

1.2.2 Reductietechnieken

- Maten van centrale tendentie
  1. modus (mo_x): waarde met grootste frequentie (vb iq: 125)
  2. mediaan: percentiel 50 ($md_x = P_{50}$) (vb iq: 127.5)
  3. rekenkundig gemiddelde: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

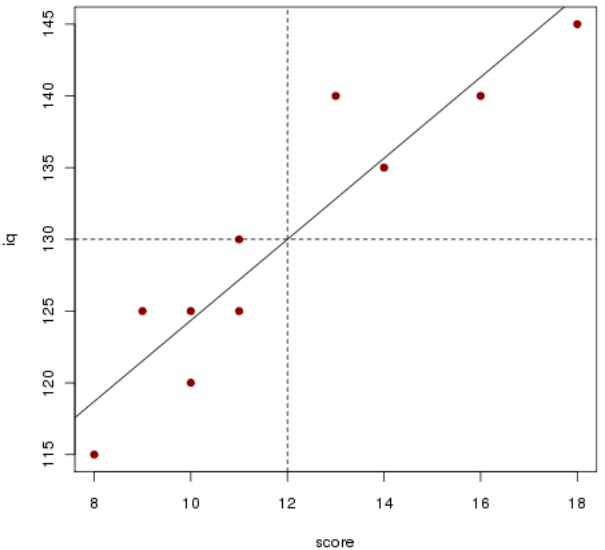
     vb $\bar{x} = \frac{16+10+11+...+10}{10} = 12$

- Maten van spreiding
  1. variatie of Sum of Squares: $\text{SS} = \sum_{i=1}^{n} (x_i - \bar{x})^2$
  2. variantie: $s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

     vb $s_x^2 = \frac{(16-12)^2+(10-12)^2+(11-12)^2+\ldots+(10-12)^2}{10} = 9.2$
  3. standaarddeviatie: $s_x = \sqrt{s_x^2}$ vb $s_x = \sqrt{9.2} = 3.03$

1.3 Beschrijvende technieken: associatiematen

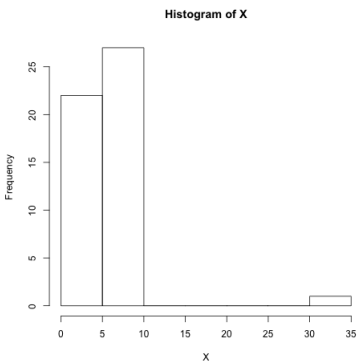
1. covariantie: lineaire samenhang
   $Cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
   vb score en iq: $Cov(x, y) = \frac{1}{10}260 = 26$
2. correlatie: normaliseren van covariantie
   $Cor_{x,y} = r_{xy} = \frac{Cov(x,y)}{\sqrt{s_x s_y}}$ vb score en iq: $r_{xy} = 0.93$

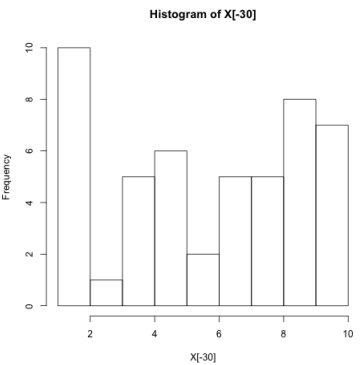

iq
score
115
120
125
130
135
140
145
8
10
12
14
16
18

1.4 Visualisatie

1.4.1 Histogram

- $X = [1, 10]$

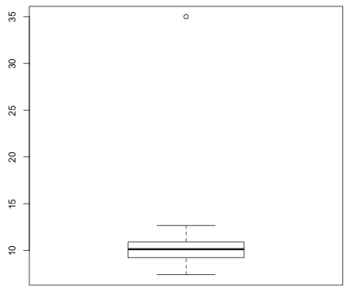


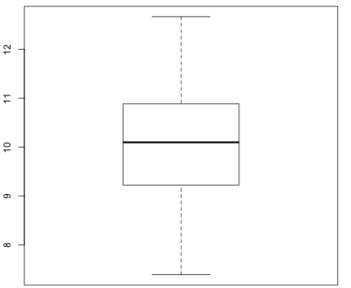


1.4.2 Boxplot

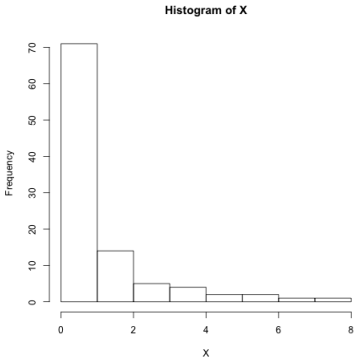
- $min - Q1 - Q2 - Q3 - max$
- $min - Q1$: 25% van de observaties
- box: 50% van de observaties
- $Q3 - max$: 25% van de observaties

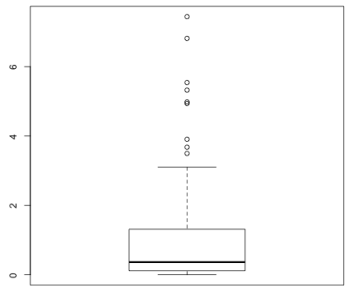
- Voorbeeld: $X \sim N(10, 1)$



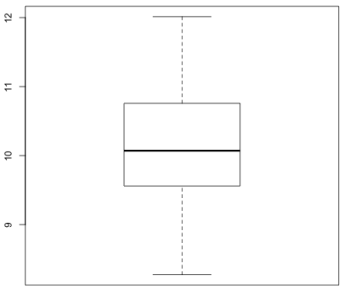


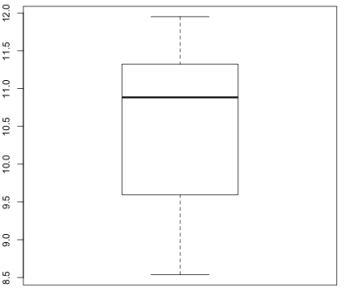
- Voorbeeld: $X \sim F(1, 15)$





- Voorbeeld: $X_1 \sim N(10,1)$,$X_2 \sim U(\min(X_1), \max(X_2))$





2 Toevalsvariabelen en kansverdelingen

2.1 Toevalsvariabelen

2.1.1 Toevalsproces en gebeurtenis

- toevalsproces: uitkomst is onvoorspelbaar
  - Kop of munt gooien
  - IQ meten bij een random gekozen persoon
- Gebeurtenis: deelverzameling van mogelijke uitkomsten voor een toevalsproces.
  - Kop of munt gooien: $\{munt\}$
  - IQ meten: 'meer dan 125'

2.1.2 Toevalsvariabele

Een toevalsvariabele of kansvariabele is een variabele waarvan de waarde in een toevalsproces onvoorspelbaar is.

- De kansvariabele 'score'

2.2 Kansen

- De kans van een gebeurtenis A bij een toevalsproces wordt gedefiniëerd als de relatieve frequentie van deze gebeurtenis als we het toevalsproces oneindig veel keer zouden herhalen.
- $P(A) = \lim_{n \to \infty} \frac{f_A}{n}$

2.3 Kansverdeling

2.3.1 Discrete kansverdeling

- Een toevalsvariabele is discreet indien de mogelijke waarden die de variabele kan aannemen een eindig (of telbaar) aantal vormen. vb ogen dobbelsteen, geslacht.
- De kansverdeling van een discrete kansvariabele geeft voor elke mogelijke waarde x_i de kans aan dat deze waarde voorkomt:
- $f_X(x_i) = f(x_i) = P[X = x_i]$

- Voorbeeld: ogen dobbelsteen

Ogen	$f(x_i)$	$F(x_i)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

- De cumulatieve verdelingsfunctie $F_X(x_i)$ drukt de kans uit dat de waarde van de toevalsvariabele X in een toevalsproces kleiner is of gelijk aan x:
- $F_X(x_i) = P(X \leq x_i) = \sum_{x \leq x_i} f(x_i)$

2.3.2 Continue kansverdeling

- De kansverdeling bestaat niet: $P[X = x] = 0$.
- Daarom Kansdichheidsfunctie:
  1. $P[a \leq x \leq b] = \int_a^b f(x)dx$
  2. $f(x) \geq 0$ voor alle x
  3. $\int_{-\infty}^{\infty} f(x)dx = 1$
- De cumulatieve verdelingsfunctie:
  $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
- Voorbeeld: De kans dat iemand kleiner of gelijk aan 80kg weegt:
  $P(X \leq 80) = 0.70$

2.4 Verwachting

- Het 'gemiddelde' van een toevalsvariabele X wordt de verwachting genoemd, $\mathrm{E}(X)$ of μ_X.
  - Discreet: $E(X) = \sum x_i f(x_i)$
    voorbeeld dobbelsteen:
    $E(X) = 1/6(1) + 1/6(2) + \ldots 1/6(6) = 3.5$
  - Continue: $E(X) = \int_{-\infty}^{+\infty} x f(x) dx$
  - Eigenschappen:
    1. $E(a) = a$
    2. $E(aX) = aE(X)$
    3. $E(a + X) = a + E(X)$
    4. $E(X \pm Y \pm Z) = E(X) \pm E(Y) \pm E(Z)$
    5. X en Y onafhankelijk: $E(XY) = E(X)E(Y)$

2.5 Variantie

- De 'mate van spreiding' van de verdeling van een kansvariabele X noemt men de variantie van X, $\mathrm{Var}(X)$ of σ_X^2.
- $Var(X) = E[X - E(X)]^2$

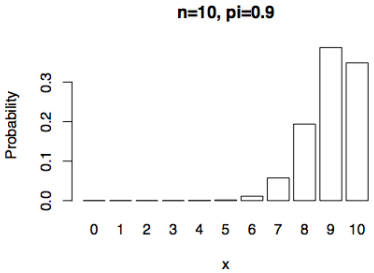
- Eigenschappen:
  1. $Var(a + X) = Var(X)$
  2. $Var(aX) = a^2 Var(X)$
  3. $Var(a) = 0$
  4. X en Y onafhankelijk: $Var(X \pm Y) = Var(X) + Var(Y)$
  5. X en Y afhankelijk: $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$

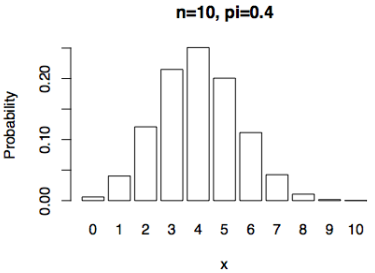
2.6 Kansverdelingen

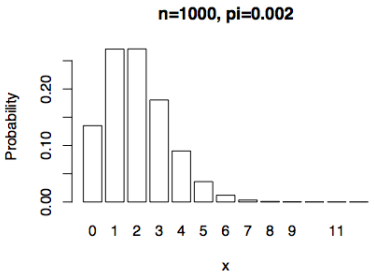
2.6.1 Binomiaal verdeling

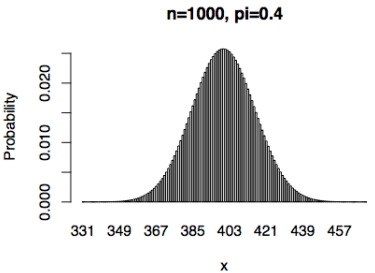
- $X \sim Binom(n, \pi)$
- Kansverdeling: $f(x) = \begin{pmatrix} n \\ x \end{pmatrix} \pi^x (1-\pi)^{n-x}$

  met $\begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{x!(n-x)!}$
- $E(X) = n\pi$
- $Var(X) = n\pi(1-\pi)$


n=10, pi=0.9
Probability
0.0
0.1
0.2
0.3
0 1 2 3 4 5 6 7 8 9 10
x


n=10, pi=0.4
Probability
0.00
0.10
0.20
0 1 2 3 4 5 6 7 8 9 10
x


n=1000, pi=0.002
Probability
0.00
0.10
0.20
0 1 2 3 4 5 6 7 8 9 11
x


n=1000, pi=0.4
Probability
0.000
0.010
0.020
331 349 367 385 403 421 439 457
x

2.6.2 Normaalverdeling

- $X \sim N(\mu, \sigma^2)$
- $f(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\}$
- $E(X) = \mu$
- $Var(X) = \sigma^2$

2.6.3 Standaardnormaalverdeling

- $\phi(x) \sim N(0, 1)$
- $z = \frac{X-\mu}{\sigma}$

2.6.4 t-verdeling

- $X \sim t(\nu)$
- $\nu =$ aantal vrijheidsgraden

2.6.5 χ^2-verdeling

- $X \sim \chi^2(\nu)$
- $\nu =$ aantal vrijheidsgraden
- som van ν onafhankelijke gekwadrateerde z-scores

2.6.6 F-verdeling

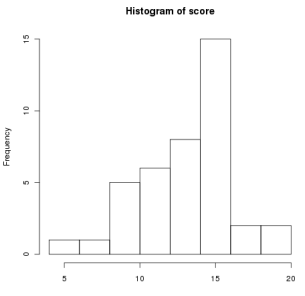
- $X \sim F(\nu_1, \nu_2)$
- ν_1 en ν_2 = vrijheidsgraden
- gebaseerd op ratio van twee χ^2-verdelingen

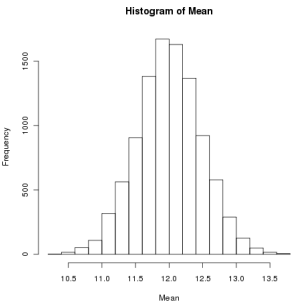
3 Statistische Inferentie: toetsen en schatten

3.1 De steekproevenverdeling

- Hypotheses: betrekking op onbekende parameters van de populatie
- Statistiek of steekproefgrootheid: maat gebaseerd op de gegevens van de steekproef: $S = f(X_1, X_2, X_3, \ldots, X_n)$
- Puntschatting: Gegeven S, schatten van parameter in de populatie?
- Intervalschatting: betrouwbaarheidsinterval
- Toetsen: geldigheid hypothese in de populatie?
- Steekproevenverdeling: verdeling van $S_1, S_2, \ldots, S_n$
- Standaardfout: op basis van steekproevenverdeling

- Voorbeeld: $n = 40, \mu = 12, sd = 3$





$\bar{X} = 12.38$ $\hat{\mu} = 12.00, \hat{\sigma} = 0.48$

3.2 De steekproevenverdeling voor $\bar{X}$

- Om de steekproevenverdeling voor $\bar{X}$ af te leiden doen we beroep op de centrale limietstelling.
- Gegeven n kansvariabelen $X_1, X_2, \ldots, X_n$ allen onafhankelijk en afkomstig van dezelfde (willekeurige) verdeling met gemiddelde μ en variantie $0 < \sigma^2 < \infty$ Stel:

  $S_n = X_1 + X_2 + X_3 + \ldots + X_n$

  Indien $n \to \infty$ dan is S_n normaal verdeeld met

  $E(S_n) = n\mu$ en $Var(S_n) = n\sigma^2$

- Gevolg 1:
  Stel $\bar{X} = \frac{S_n}{n} = \frac{X1+X2+X3+...+Xn}{n}$

  Indien $n \to \infty$ dan is $\bar{X}$ normaal verdeeld met

  $E(\bar{X}) = \mu$ en $Var(\bar{X}) = \sigma^2/n$
- Opmerkingen:
  Normaalverdeling goede benadering:
  - Vanaf $n > 30$
  - Indien $n \leq 30$ en oorspronkelijke scores zijn normaal verdeeld
- Voorbeeld:
  - Geobserveerde steekproefgemiddelde $\bar{X} = 12.38$
  - standaardafwijking of standaardfout: $\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{9}{40}} = 0.474$

- Gevolg 2:
  Stel $Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$

  Indien $n \to \infty$ dan is $Z_{\bar{X}}$ standaardnormaal verdeeld met

  $E(Z_{\bar{X}}) = 0$ en $Var(Z_{\bar{X}}) = 1$

3.3 De steekproevenverdeling voor $\bar{X}$ (σ^2 ongekend)

- Vervangen van σ^2 door steekproefschatter s^2 in $Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ dan:
- $t = \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$
- $t \sim t(\nu)$ met $\nu = n - 1$

3.4 Intervalschatting

3.4.1 Puntschatting

- De geschatte waarde $\hat{\theta}$ weerspiegelt:
  1. de waarde θ in de populatie
  2. de steekproeffout ε: $\hat{\theta} = \theta + \varepsilon$

3.4.2 Het betrouwbaarheidsinterval

- Hoe smaller, hoe nauwkeurig de schatting
- Confidentie niveau: $100(1 - \alpha)\%$, met $\alpha = 0.05$, $\alpha = 0.01$ of ...

3.4.3 Opstellen van betrouwbaarheidsinterval

1. Trek random steekproef
2. Puntschatting θ: $\hat{\theta}$
3. Berekenen onder- en bovengrens:
   - ondergrens $= \hat{\theta} - (|g_1^{\alpha/2}| \times s)$
   - bovengrens $= \hat{\theta} + (|g_1^{\alpha/2}| \times s)$
4. ... 95% van de intervallen zal θ bevatten

3.5 Toetsen van hypotheses

3.5.1 Nulhypothese

- Is populatieparameter θ gelijk aan vooropgestelde waarde θ_0?
- H_0 is de hypothese die effectief getoets wordt: $H_0 : \mu = 110$
- H_a is de alternatieve hypothese:
  1. tweezijdig: $H_a : \mu \neq 110$
  2. linkszijdig: $H_a : \mu < 110$
  3. rechtszijdig: $H_a : \mu > 110$

3.5.2 Toetsingsgrootheid G

1. Verdeling $G \sim$ theoretische verdeling vb t, F, ...
2. Verdeling van G onder de assumptie dat H_0 waar is.

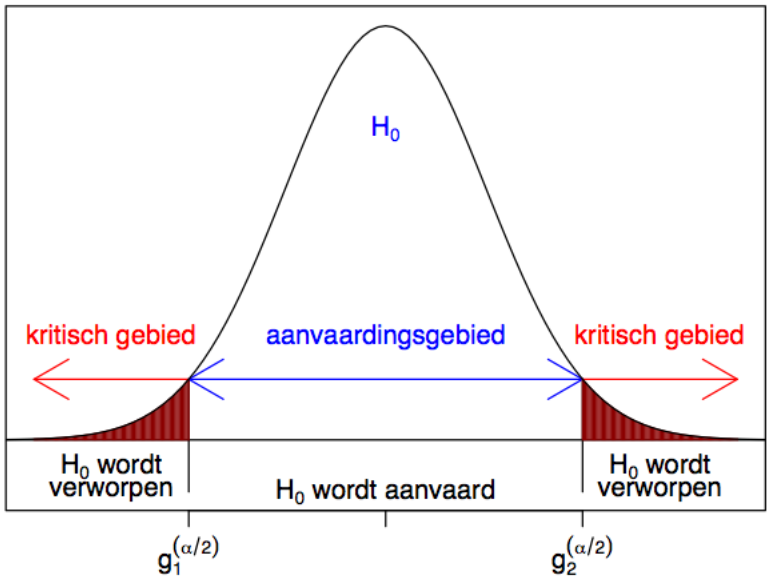
3.5.3 Kies betrouwbaarheid $(1 - \alpha)$

- $1 - \alpha$: conditionele kans om H_0 te aanvaarden op voorwaarde dat H_0 juist is
- α: significantieniveau is de conditionele kans om de nulhypothese te verwerpen op voorwaarde dat de nulhypothese juist is.

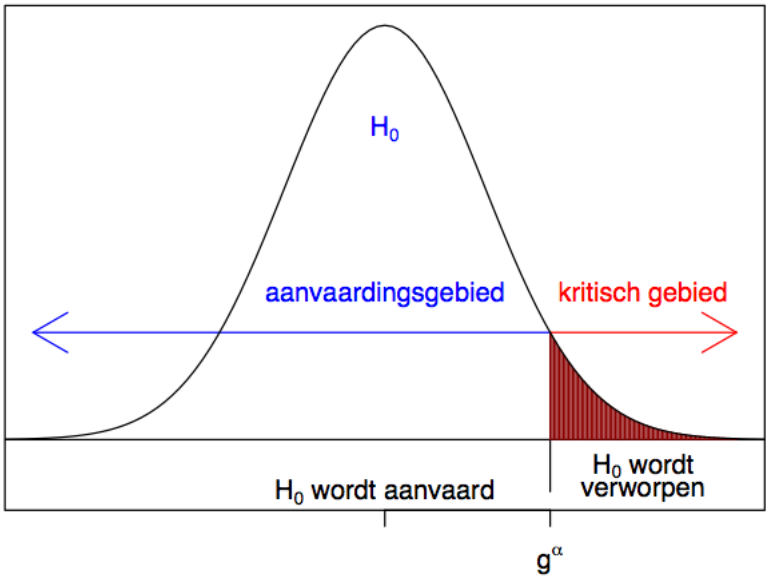
3.5.4 H_0 aanvaarden of verwerpen

- Tweezijdig toetsen:
  Bepaal kritische waarden $g_1^{\alpha/2}$ en $g_2^{\alpha/2}$:

  $P(G \leq g_1^{\alpha/2}) = \alpha/2$ en $P(G \geq g_2^{\alpha/2}) = \alpha/2$
  aanvaardingsgebied: $g_1^{\alpha/2} \leq G \leq g_2^{\alpha/2}$
  kritisch gebied: gebied buiten deze twee waarden


H_0
kritisch gebied
aanvaardingsgebied
kritisch gebied
H_0 wordt verworpen
H_0 wordt aanvaard
H_0 wordt verworpen
$g_1^{(\alpha/2)}$
$g_2^{(\alpha/2)}$

- Eenzijdig toetsen:
  Bepaal kritische waarde g^α:
  $P(G \leq g^\alpha) = \alpha$ OF $P(G \geq g^\alpha) = \alpha$


H_0
aanvaardingsgebied
kritisch gebied
H_0 wordt aanvaard
H_0 wordt verworpen
g^{α}

3.5.5 H_0 aanvaarden of verwerpen met p-waarde

- Bereken kans dat onder de verdeling van G onder H_0 dat g of een waarde groter dan g zich voordoet.
  - Eenzijdig: $p = P(G \geq g)$ of $p = P(G \leq g)$
  - Tweezijdig: $p_{2zijdig} = 2 \times p_{1zijdig}$

3.6 Toetsen van hypotheses

3.6.1 One-sample t-test

- Gebruik: Nagaan of het gemiddelde van een continue variabele afwijkt van een gegeven waarde μ_0.
- assumpties:
  1. Onafhankelijke observaties.
  2. Normaalverdeelde observaties of een 'grote' steekproef.

- $H_0 : \mu = \mu_0$
- toetsingsgrootheid: $t = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$
- betrouwbaarheidsinterval:
  - ondergrens $= \bar{X} - (|t_{n-1}^{\alpha/2}| \times s/\sqrt{n})$

- bovengrens $= \bar{X} + (|t_{n-1}^{\alpha/2}| \times s/\sqrt{n})$

Voorbeeld:

- $n = 100$, $\bar{x} = 116$ en $s^2 = 400$
- $H_0 : \mu = 110$, $H_a : \mu \neq 110$
- $t = \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{116-110}{20/\sqrt{100}} = 3$
- $\alpha = 0.05$, $t_{99}^{0.025} = +2$ en -2, $p = 0.0034$
- ondergrens $= 116 - (2 \times 20\sqrt{100})$, bovengrens $= 116 + (2 \times 20\sqrt{100})$
- 95% betrouwbaarheidsinterval is $[112, 120]$, μ_0 ligt niet in dit interval.

3.6.2 two-sample t-test

- Gebruik: Nagaan of het gemiddelde van een continue variabele gelijk is in twee onafhankelijke populaties.
- assumpties:
  1. Onafhankelijke observaties.
  2. Normaalverdeelde observaties of een 'grote' steekproef in elke groep.

- $H_0 : \mu_1 = \mu_2$ en varianties homogeen ($\sigma_1^2 = \sigma_2^2 = \sigma$)
- toetsingsgrootheid: $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2_{pooled}(\frac{1}{n_1} + \frac{1}{n_2})}}$
- $s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
- betrouwbaarheidsinterval:
  - ondergrens $= (\bar{X}_1 - \bar{X}_2) - (|t^{\alpha/2}_{n_1+n_2-2}| \times s_{(\bar{X}_1 - \bar{X}_2)})$

– bovengrens $= (\bar{X_1} - \bar{X_2}) + (|t^{\alpha/2}_{n_1+n_2-2}| \times s_{(\bar{X_1}-\bar{X_2})})$

Voorbeeld:

- $n_1 = 4$,$n_2 = 6$,$\bar{x_1} = 14.75$,$\bar{x_2} = 10.33$ en $s^2_{pooled} = 5.26$
- $H_0 : \mu_A = \mu_B$
- $t = \frac{(\bar{X_1}-\bar{X_2})-(\mu_1-\mu_2)}{\sqrt{s^2_{pooled}(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{4.417-0}{\sqrt{5.26(\frac{1}{4}+\frac{1}{6})}} = 2.983$
- $\alpha = 0.05$, $t^{0.025}_8 = 2.306$, $p = 0.0175$
- ondergrens $= 4.417 - (2.306 \times 1.48) = 1.003$
- bovengrens $= 4.417 + (2.306 \times 1.48) = 7.831$
- 95% betrouwbaarheidsinterval is $[1.003, 7.831]$, $(\mu_A - \mu_B)$ ligt niet in dit interval.

3.6.3 One-way analysis of variance (Anova)

- Gebruik: Nagaan of het gemiddelde van een continue variabele gelijk is in twee of meer (k) onafhankelijke populaties.
- Uitbreiding van de two-sample t-test
- assumpties:
  1. Onafhankelijke observaties.
  2. Normaalverdeelde observaties of een 'grote' steekproef in elke groep.
  3. Gelijke variantie in elke groep.
- principe: is de variate tussen (between) groepen groot indien vergeleken met de variatie binnen (within) groepen?
- within MSE $= \frac{withinSS}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j}(Y_{ij}-\bar{Y}_i)^2}{n-k}$

- between MSE $= \frac{betweenSS}{k-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j}(\bar{Y}_i - \bar{Y})^2}{k-1}$

- $H_0 : \mu_1 = \mu_2 = \ldots, \mu_k$
- toetsingsgrootheid: $F = \frac{betweenMSE}{withinMSE}$, met onder $H_0 \sim F(k-1, n-k)$.

- Voorbeeld:
  Data:

Groep1	Groep2	Groep3
1	2	4
2	2	3
2	3	5
2	4	4
3	4	4
2	4	5
1	3	4
2	2	5
3	3	6
3	3	5
$\bar{y_1} = 2.1$	$\bar{y_2} = 3$	$\bar{y_3} = 4.5$
$\bar{y} = 3.2$		

Output:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	29.400	2	14.700	22.810	.000
Within Groups	17.400	27	.644		
Total	46.800	29			

4 Categorische data-analyse

4.1 Inleiding

- Afhankelijke variabele: categorisch (nominaal of ordinaal)
- vb geslacht, opleidingsniveau
- aantallen, frequenties, proporties, percentages

4.2 1 Categorische variabele

4.2.1 1 Categorische variabele met 2 niveaus

- Voorbeeld:

Vrouwen	Mannen	Totaal	Vrouwen	Mannen	Totaal
11	19	30	0.3666	0.6333	1.0000

- De binomiaaltoets:
  $H_0 : \pi = \pi_0$ en stel $\pi_0 = 0.56$

  $H_a : \pi < 0.56$

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} \pi^x (1-\pi)^{n-x}$$

De kans dat er exact 10 vrouwen zijn:

$$P(X = 10) = \begin{pmatrix} 30 \\ 10 \end{pmatrix} 0.56^x (1-0.56)^{30-10} = 0.0067$$

De kans dat er 11 vrouwen of minder zijn:

$P(X \leq 11) = P(X = 0) + P(X = 1) + \ldots + P(X = 11) = 0.0256$

$p_{tweezijdig} = 0.0256 \times 2 = 0.052$

Indien $\min n\pi_0, n(1-\pi_0) > 5$: benaderen via normaalverdeling

$z = \frac{|x-n\pi|}{\sqrt{n\pi_0(1-\pi_0)}}$

$z = \frac{|11-30\times 0.56|}{\sqrt{30\times 0.56(1-0.56)}} = 2.133$

$P(Z > 2.133) = 0.016$

Soms continuiteits-correctie: $z = \frac{|11-30\times 0.56|-0.5}{\sqrt{30\times 0.56(1-0.56)}} = 1.95$

$P(Z > 1.95) = 0.0256$

4.2.2 1 Categorische variabele met $J \geq 2$ niveaus

- Voorbeeld:

	Klinische	Bedrijfs	Experimentele	Totaal
n_j	258	69	19	346
p_j	0.75	0.20	0.05	1.00
π_j	0.70	0.28	0.02	1.00
$\mu_j (= n \times \pi_j)$	242.20	97.88	6.92	3.46

- De Pearson chi-kwadraat toets: $H_0 : p_j = \pi_j$ of $n_j = \mu_j$, voor alle j.
- $\chi^2 = \sum_{j=1}^{J} \frac{(n_j - \mu_j)^2}{\mu_j}$, met $df = J - 1$.
- $\chi^2_2 = \frac{(258-242.20)^2}{242.20} + \frac{(69-96.88)^2}{96.88} + \frac{(19-6.92)^2}{6.92} = 30.1416$, $p < 0.0001$

4.3 2 Categorische variabelen

4.3.1 2-Wegs kruistabel: geobserveerde frequenties

- Voorbeeld:

	Klinische	Bedrijfs	Experimentele	Totaal
geslaagd $= 0$	120	34	5	159
geslaagd $= 1$	138	35	14	187
totaal	258	69	19	346

- Notatie:

	Klinische	Bedrijfs	Experimentele	Totaal
geslaagd $= 0$	n_{11}	n_{12}	n_{13}	n_{1+}
geslaagd $= 1$	n_{21}	n_{22}	n_{23}	n_{2+}
totaal	n_{+1}	n_{+2}	n_{+3}	n

4.3.2 Test voor onafhankelijke variabelen

- Is er een verband tussen X en Y? Zo niet: statistisch onafhankelijk
- $H_0 : \pi_{ij} = \pi i+ \times \pi+j$, voor alle i, j.
- $\sim H_0 : \pi_{i|j} = \pi+j$, voor alle i, j.
- Onder $H_0 : \mu_{ij} = n\pi ij = n \times \pi i+ \times \pi+j$.
- π_{i+} en $\pi+j$ onbekend:
  $\hat{\mu}_{ij} = np_{i+}p+j = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$.
- $\hat{\mu}_{ij}$: geschatte verwachte frequenties.

$$\hat{\mu}_{11} = \frac{159 \times 258}{346} = 118.56$$

$$\hat{\mu}_{12} = \frac{159 \times 69}{346} = 31.71$$

$$\hat{\mu}_{13} = \frac{159 \times 19}{346} = 8.73$$

$$\hat{\mu}_{21} = \frac{187 \times 258}{346} = 139.44$$

$$\hat{\mu}_{22} = \frac{187 \times 69}{346} = 37.29$$

$$\hat{\mu}_{23} = \frac{187 \times 19}{346} = 10.27$$

	Klinische	Bedrijfs	Experimentele	Totaal
geslaagd = 0	118.56	31.71	8.73	159
geslaagd = 1	139.44	37.29	10.27	187
totaal	258	69	19	346

- $\chi^2 \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
- $df = (I-1)(J-1)$
- $\chi^2 = 3.2891, df = 2, p = 0.1931$

4.4 Veralgemeend lineaire modellen

- Afhankelijke variabele is categorisch, maar meerdere predictoren
- Regressie, anova niet meer mogelijk

4.4.1 Logistische regressie

- Afhankelijke variabele is dichotoom, of binair
- Alternatief: probit regressie
- Indien afhankelijke variabele meerdere niveaus: multinomiale regressie

4.4.2 Poisson regressie

- Afhankelijke variabele is een frequentie die een poisson verdeling volgt
- Aantal ongevallen/uur, Aantal klanten per dag,...

4.4.3 Loglineaire analyse

- Speciaal geval van poisson regressie
- Associatie tussen verschillende nominale variabelen in kaart brengen

5 Enkelvoudige Lineaire Regressie

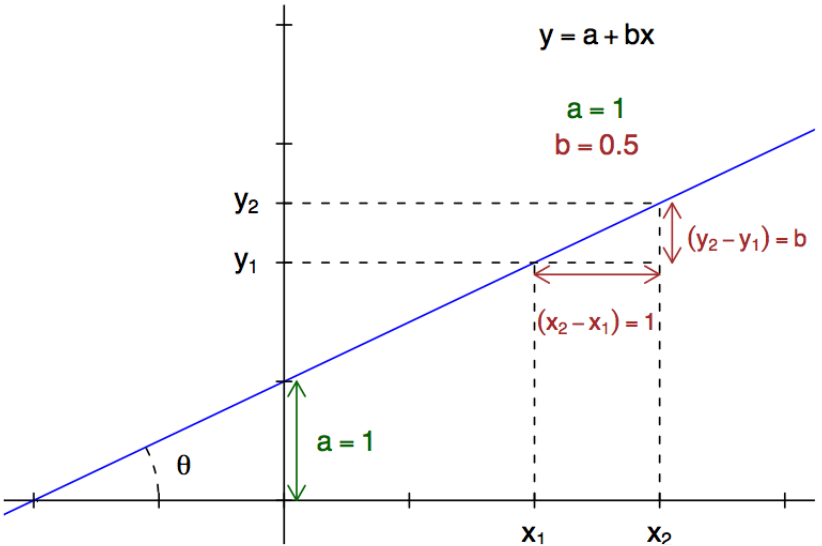
5.1 Inleiding

5.1.1 doel

- Modelleren van lineaire relatie tussen een afhankelijke variabele Y en een onafhankelijke variabele X
- X en Y gemeten op minstens interval niveau
- Lineaire regressie laat toe:
  1. variatie in Y te verklaren in termen van variatie in X
  2. Y te voorspellen op basis van X
  3. nagaan of X een significante predictor is

5.1.2 Vergelijking van een rechte

- $y = a + bx$
- a = intercept: indien $x = 0$, dan $y = a$
- b = helling of slope: indien de waarde van x stijgt met één eenheid, dan stijgt de waarde van y met b


y = a + bx
a = 1
b = 0.5
y2
y1
(y2 – y1) = b
(x2 – x1) = 1
a = 1
θ
x1
x2

5.2 Het regressiemodel

5.2.1 Structuur

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$
- β_0 en β_1 zijn de regressiecoëfficiënten
- ε_i is de foutterm voor observatie i

5.2.2 assumpties

- $E(\varepsilon_i) = 0 \Rightarrow E(Y_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_1 X_{pi}$
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$ voor alle $i \Rightarrow Var(Y_i) = \sigma_{\varepsilon_i}^2$
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ voor alle $i \neq j$

5.2.3 Onderzoeksvragen

- Wat is de bijdrage van X in het model? Is dit significant?

  $H_0 : \beta_1 = 0$

- Hoeveel variantie in Y wordt verklaard door het model?

  $H_0 = R^2 = 0$,
  met R^2 =determinatiecoëfficiënt

5.3 Parameters

- Enkelvoudig regressiemodel telt drie vrije parameters:
  1. de regressieconstante β_0
  2. de regressiecoëfficiënt β_1
  3. de variantie van de fouttermen σ_ε^2
- Schatten van parameters? Methode van kleinste kwadraten, maximum likelihood
- Minimaliseren van $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, met $\hat{y}_i = b_0 + b_1 x_i$

5.4 Toetsen van hypotheses

- $H_0 : \beta_0 = 0$: $t = \frac{b_0 - \beta_0}{s_{b_0}}$ met $n - 2$ vrijheidsgraden
- $H_0 : \beta_1 = 0$: $t = \frac{b_1 - \beta_1}{s_{b_1}}$ met $n - 2$ vrijheidsgraden
- Voorbeeld score en iq:

	B	Std.Error	t	Sig
constant	-27.765	5.58	-4.975	0.001
iq	0.306	0.043	7.143	0.000

- ondergrens: $b_i - (|t_{n-2}^{\alpha/2}| \times s_{b_i})$
- bovengrens: $b_i + (|t_{n-2}^{\alpha/2}| \times s_{b_i})$

5.5 De determinatiecoëfficiënt R^2

- Nulmodel: $Y_i = \beta_0 + \varepsilon_i \Rightarrow b_0 = \bar{y}$
- Total sum of squares (SST): $E_0 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$
- Residual sum of squares (SSE): $E_p = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Regression sum of squares (SSR) = SST-SSE
- $R^2 = \frac{E_0 - E_p}{E_0} \Rightarrow 0 < r^2 < 1$
- $H_0 : R^2 = 0$: $F = \frac{(E_0 - E_p)/(df_0 - df_p)}{E_p/df_p}$
- Voorbeeld score en iq: $R^2 = 0.864$

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	79.529	1	79.529	51.019	.000
Residual	12.471	8	1.559		
Total	92.000	9			

6 Meervoudige Lineaire Regressie

6.1 Structuur

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i$, $i = 1, 2, \ldots, n$
- $\beta_0, \ldots, \beta_1$ zijn de regressiecoëfficiënten
- ε_i is de foutterm voor observatie i

6.2 Onderzoeksvragen

- Wat is de bijdrage van X_p in het model? Is dit significant? H_0 : $\beta_p = 0$
- Hoeveel variantie in Y wordt verklaard door het model? $H_0 = R^2 = 0$, met R^2 =determinatiecoëfficiënt

6.3 Parameters

- Schatten van vrije parameters: Cfr. Enkelvoudige lineaire regressie

6.4 Toetsen van hypotheses

- $H_0 : \beta_p = 0$: $t = \frac{b_p - \beta_p}{s_{b_p}}$ met $n - p - 1$ vrijheidsgraden
- Voorbeeld score, iq en leeftijd:

	score	iq	leeftijd
1	16.00	140.00	22.00
2	10.00	120.00	24.00
3	11.00	125.00	25.00
4	14.00	135.00	31.00
5	8.00	115.00	30.00
6	18.00	145.00	26.00
7	13.00	140.00	26.00
8	9.00	125.00	29.00
9	11.00	130.00	33.00
10	10.00	125.00	27.00

	B	Std.Error	t	Sig
constant	-22.513	7.243	-3.108	0.017
iq	0.295	0.043	6.784	0.000
leeftijd	-0.138	0.124	-1.114	0.302

6.5 De determinatiecoëfficiënt R^2

- Nulmodel: $Y_i = \beta_0 + \varepsilon_i \Rightarrow b_0 = \bar{y}$
- Total sum of squares (SST): $E_0 = \sum_{i=1}^{n} (y_i - \bar{y_i})^2$
- Residual sum of squares (SSE): $E_p = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$
- Regression sum of squares (SSR) = SST-SSE
- $R^2 = \frac{E_0 - E_p}{E_0} \Rightarrow 0 < r^2 < 1$
- $H_0 : R^2 = 0$: $F = \frac{(E_0 - E_p)/(df_0 - df_p)}{E_p/df_p}$
- Voorbeeld score, iq en leeftijd: $R^2 = 0.885$

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	79.529	1	79.529	26.900	.000
Residual	12.471	8	1.559		
Total	92.000	9			