

Gegevens kunnen gemeten worden op verschillende types schalen: welke? Indien je een dataset opmeet met de persoonlijke gegevens van de inwoners van Brussel, heb je data zoals: naam, geboortejaar, gewicht, lengte, geslacht, ...

Geef de schaal van deze gegevens.

- Nominale schaal: beschrijving in woorden, geen overlap
- Ordinale schaal: gerangschikte waarden (zeer goed, goed, slecht), geen precieze afstand
- Intervalschaal: idem ordinale schaal maar afstand tussen niveaus zijn betekenisvol en numeriek. (Temperatuur schalen)
- Verhoudingsschaal (ratioschaal): idem intervalschaal en verhouding der opmetingen direct interpreteerbaar als de verhouding der werkelijke waarden (kg, g)

Met opmerkingen [gw1]: P1.4

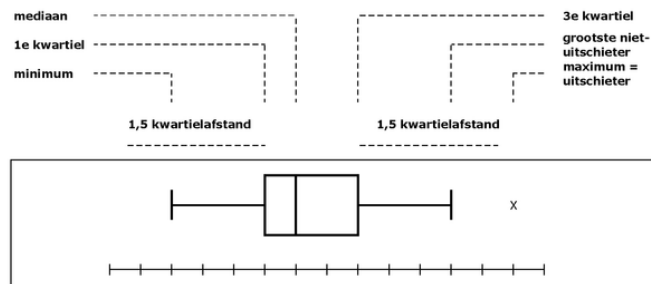
| Data         | Schaal                              |
|--------------|-------------------------------------|
| Naam         | Nominale schaal                     |
| Geboortejaar | Intervalschaal of verhoudingsschaal |
| Gewicht      | Verhoudingsschaal                   |
| Lengte       | Verhoudingsschaal                   |
| Geslacht     | Nominale schaal                     |

Welke kengetallen ken je om de spreiding van de gegevens in een steekproef weer te geven.

Definieer ze. Hoe maak je een boxplot?

Met opmerkingen [gw2]: P2.13-2.19

- Mediaan =  $x_{(n+1)/2}$  Als n oneven is  
 $= \frac{1}{2} (x_{n/2} + x_{(n+1)/2})$  Als n even is
- Range (R) =  $x_n - x_1$
- Variantie ( $\sigma^2$ ) =  $\frac{\sum (x_i - \mu)^2}{N}$  (enkel van een populatie)  
variantie ( $S_n^2$ ) =  $\frac{\sum (x_i - \bar{x})^2}{n-1}$  (enkel van een steekproef)
- Standaardafwijking ( $S_n$ ) =  $\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
- Mediaan absolute afwijking (MAD) =  $\text{mediaan}\{|x_i - \text{med}|\}_{i=1}^n$
- Gemiddelde absolute afwijking (MeanAD) =  $\frac{1}{n} \sum_{i=1}^n |x_i - \text{med}|$



uitschieters(\*)  $\begin{cases} < Q_1 - 1.5IQR \\ > Q_3 + 1.5IQR \end{cases}$  met IQR de interquartile range ( $Q_3 - Q_1$ )

**Welke kengetallen ken je om de centrale tendens van de gegevens in een steekproef weer te geven. Definieer ze.**

- Gemiddelde  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Mediaan  $med = \begin{cases} x_{\frac{n+1}{2}} & n \text{ oneven} \\ \frac{1}{2} \left( x_{\frac{n}{2}} + x_{\frac{n+1}{2}} \right) & n \text{ even} \end{cases}$
- Modus= waarde die het meest voorkomt
- Midrange=  $\frac{x_1 + x_n}{2}$
- Gewogen gemiddelde  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i w_i$  met  $w_i$  het gewicht van de meting

Met opmerkingen [gw3]: P2.11-2.13

**Welke kengetallen ken je om de positie van een datapunt in een steekproef weer te geven. Definieer ze.**

- De standaardscore of z-score  $z = \frac{x_i - \bar{x}}{s}$  (voor statistieken)  
 $z = \frac{X - \mu}{\sigma}$  (voor populaties)
- Kwartielen  $\xi_\alpha = x_p + \rho(x_{p+1} - x_p)$  met  $p + \rho = \frac{\alpha}{100}(n + 1)$  Met p geheel en  $\rho$  de overschot
- Kwantielen: idem kwartielen fractie i.p.v. procent (kwantiel = kwartiel/100)
- IQR (interquartile range) =  $Q_3 - Q_1$

Met opmerkingen [gw4]: P2.15-2.17

**Indien je over een steekproef beschikt, wat is dan de relatieve en de cumulatieve frequentie? Hoe kun je ze grafisch voorstellen (eventueel op meerdere manieren)? Wat is een empirische verdelingsfunctie?**

- Relatieve frequentie: frequentie uitgedrukt in percentages  
Horizontale as verdeeld in klassen  
Verticale as: percentages (dichtheidsschaal)
- Cumulatieve frequentie: Frequentie uitgedrukt in percentages, maar niet die van de eigen klasse wordt weergegeven, maar de frequentie van die klasse opgeteld bij alle frequenties van voorgaande klassen.  
Verticale as: percentages

Met opmerkingen [gw5]: P2.4-2.9

Hoe voorstellen:

- Histogram
- Polygoon: klassenmiddens worden verbonden van de relatieve frequenties.
- Ogief: rechterklassengrenzen van het cumulatief frequentiehistogram worden verbonden

Empirische verdelingsfunctie

$$f_n(x) = \frac{\text{aantal metingen} \leq x}{n}$$

Deze functie geeft de fractie van metingen die  $\leq$  dan x. Dit is een trapfunctie.

**Geef de definities van universum, gebeurtenis en het begrip kans. Wat is een voorwaardelijkekans?**

Universum → De verzameling van alle mogelijke uitkomsten van een experiment noemen we het universum ( $\Omega$ ).

Gebeurtenis → In de kanstheorie worden deelverzamelingen van het universum gebeurtenissen genoemd. Indien de deelverzameling slechts één element bevat, noemen we ze een elementaire gebeurtenis.

Kans → Een kans is een functie waarbij met elke gebeurtenis A van een universum  $\Omega$  een reëel getal  $P(A)$  geassocieerd wordt. Dit getal moet aan de volgende regels voldoen:

- 1)  $0 \leq P(A) \leq 1$
- 2)  $P(\Phi) = 0$  en  $P(\Omega) = 1$
- 3) Als  $A \cap B = \Phi$  dan is  $P(A \cup B) = P(A) + P(B)$

Voorwaardelijke kans → We noemen dit de voorwaardelijke kans op het optreden van gebeurtenis V als gebeurtenis B heeft plaatsgevonden en  $P(B) \neq 0$ .

$$P(V|B) = \frac{P(V \cap B)}{P(B)}$$

**Formuleer de wet van de totale kans en de regel van Bayes. Illustreer beiden met de hulp van een figuur. Bewijs de regel van Bayes.**

Wet van totale kans:

Onderstel dat  $E_1, E_2, \dots, E_n$  een partitie (1 mogelijke oplossing) is van het universum. Dan geldt voor A:

$$P(A) = \sum_{i=1}^n P(A|E_i)P(E_i)$$

Vb voorstelling zie p3.5 figuur 3.2

Bewijs:

$$\begin{aligned} P(A) &= P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n) \\ \Leftrightarrow P(A) &= P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \dots + P(A|E_n)P(E_n) \\ \Leftrightarrow P(A) &= \sum_{i=1}^n P(A|E_i)P(E_i) \end{aligned}$$

Regel van Bayes:

Onderstel dat  $E_1, E_2, \dots, E_n$  een partitie (1 mogelijke oplossing) is van het universum en zij A een willekeurige gebeurtenis met  $P(A) > 0$ . Dan geldt voor gelijk welke vaste k (met  $1 \leq k \leq n$ ):

$$P(E_k|A) = \frac{P(A|E_k)P(E_k)}{\sum_{i=1}^n P(A|E_i)P(E_i)}$$

Vb voorstelling zie p 3.6 figuur 3.3

Bewijs:

Met opmerkingen [gw6]: P3.1-3.4

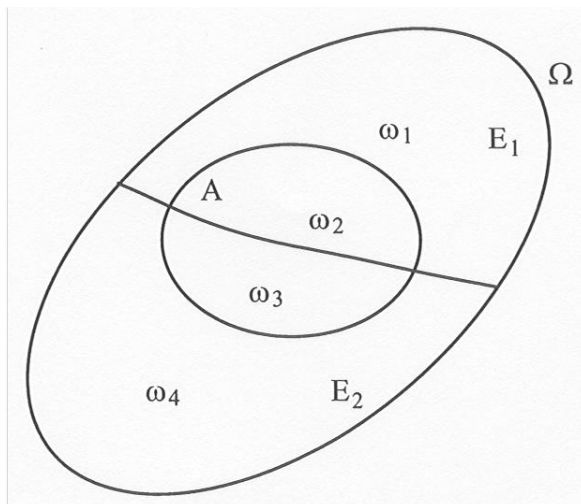
Met opmerkingen [gw7]: P3.4-3.7

$$\begin{aligned}
 &\begin{cases} P(E_k \cap A) = P(E_k | A)P(A) \\ P(E_k \cap A) = P(A | E_k)P(E_k) \end{cases} \\
 &\Leftrightarrow P(E_k | A)P(A) = P(A | E_k)P(E_k) \\
 &\Leftrightarrow P(E_k | A) = \frac{P(A | E_k)P(E_k)}{P(A)}
 \end{aligned}$$

Formuleer de regel van Bayes en illustreer deze wet met de hulp van een voorbeeld en een figuur.

Met opmerkingen [gw8]: P3.5-3.7

$$P(E_k | A) = \frac{P(A | E_k)P(E_k)}{\sum_{i=1}^n P(A | E_i)P(E_i)}$$



Vb1: 2 vazen X en Y met 10 ballen; in vaas X 6 witte en 4 zwarte; in vaas Y 2 witte en 8 zwarte  
Je trekt geblinddoekt een bal uit een vaas die wit is. Wat is de kans dat de vaas waaruit je trok vaas X was ?

Legende figuur:

A: bal is wit, E1: vaas is X; E2: vaas is y

$P(W | X) = 0.6$      $P(Z | X) = 0.4$

$P(W | Y) = 0.2$      $P(Z | Y) = 0.8$

$$P(X | W) = \frac{P(W | X)P(X)}{P(W | X)P(X) + P(W | Y)P(Y)} = \frac{(0.6)(0.5)}{(0.6)(0.5) + (0.2)(0.5)} = 0.75$$

$$P(Y | W) = \frac{P(W | Y)P(Y)}{P(W | X)P(X) + P(W | Y)P(Y)} = \frac{(0.2)(0.5)}{(0.6)(0.5) + (0.2)(0.5)} = 0.25$$

**Wat betekent het indien twee of meerdere gebeurtenissen onafhankelijk zijn? Geef ééngevolg.**

Onafhankelijkheid  $\rightarrow$  Twee gebeurtenissen A en B heten onafhankelijk als het voor de kans op A niet uitmaakt of B al dan niet gebeurd is.  $P(A) = P(A|B)$

$P(A|B) > P(A) \Rightarrow B$  draagt + info over A

$P(A|B) = P(A) \Rightarrow A$  en B zijn onafhankelijk

$P(A|B) < P(A) \Rightarrow B$  draagt - info over A

Gevolg

- Productregel:  $P(A \cap B) = P(A)P(B)$

**Wat is een stochastische variabele? Geef tevens de definitie van een verdelingsfunctie en van een dichtheidsfunctie. Bestaat de laatste altijd?**

Stochastische variabele:  $X: \Omega \rightarrow \mathbb{R}; \omega \rightarrow X(\omega)$ , De functie X wordt een stochastische variabele genoemd

Verdelingsfunctie  $\rightarrow$  Als X een stochastische variabele is, dan heet de functie  $F_X$ :

$$F_X(a) = P(X \leq a)$$

De (cumulatieve) verdelingsfunctie van X.

Dichtheidsfunctie  $\rightarrow$  we noemen een stochastische variabele X continu als de verdelingsfunctie  $F_X$  overal continu is en bovendien overal differentieerbaar (behalve eventueel in een eindig aantal punten). Hieruit wordt de dichtheidsfunctie  $f_X$  te definiëren als de afgeleide van  $F_X$ :

$$f_X = \frac{d}{dx} F_X(X)$$

De dichtheidsfunctie kan enkel gemaakt worden als  $F_X$  een continue verdeling is.

**Wat zijn onafhankelijke stochastische variabelen? Indien twee stochastische variabelen onafhankelijk zijn, wat weet je dan over de verwachtingswaarde van hun product en de variantie van hun som?**

Onafhankelijke stochastische variabelen  $\rightarrow$  twee stochastische variabelen X en Y heten onafhankelijk als de gebeurtenissen  $\{a_1 < X \leq b_1\}$  en  $\{a_2 < Y \leq b_2\}$  onafhankelijk zijn voor alle  $a_i, b_i$  in  $\mathbb{R} (i = 1, 2)$ , of,

equivalent, als:  $P(\{a_1 < X \leq b_1\} \cap \{a_2 < Y \leq b_2\}) = P(\{a_1 < X \leq b_1\})P(\{a_2 < Y \leq b_2\})$

$$E[X+Y] = E[X] + E[Y]$$

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] \text{ en } \text{Var}[X-Y] = \text{Var}[X] + \text{Var}[Y]$$

Met opmerkingen [gw9]: P4.1-4.5

Met opmerkingen [gw10]: P410-4.16

**X en Y zijn 2 stochastische variabelen. Hoe zijn we tot de definitie van het begrip covariantie van X en Y gekomen? Geef tevens de definitie van de correlatie van X en Y. Zijn twee ongecorreleerde stochastieken altijd onafhankelijk?**

Met opmerkingen [gw11]: P4.18-4.19

Voor 2 afhankelijke stochastieken kunnen we de som van de variantie als volgt berekenen:

$$\begin{aligned}\sigma^2_{X+Y} &= E[(X+Y - E[X+Y])^2] \\ \Leftrightarrow \sigma^2_{X+Y} &= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\ \Leftrightarrow \sigma^2_{X+Y} &\Leftrightarrow \sigma^2_X + \sigma^2_Y + 2E[(X - E[X])(Y - E[Y])]\end{aligned}$$

De term  $E[(X - E[X])(Y - E[Y])]$  geeft de mate van afhankelijk weer.

$$\text{Covariantie} \rightarrow \text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$\text{Correlatiecoëfficiënt } \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Twee stochastische variabelen zijn ongecorreleerd als  $\text{cov}(X,Y) = 0$  of  $E[XY] = E[X]E[Y]$

Twee onafhankelijke stochastische variabelen zijn ongecorreleerd, maar 2 ongecorreleerde stochastische variabelen zijn daarom niet steeds onafhankelijk.

**Men zegt dat een geometrisch verdeelde stochastische variabele geen geheugen heeft. Wat betekent dit? Ken je nog een andere verdeling, die geen geheugen heeft? Bewijs dat beide (geometrische + de andere) verdelingen geen geheugen hebben.**

Met opmerkingen [gw12]: P5.5-5.6

Een geometrische verdeling is het herhaaldelijk uitvoeren van een Bernoulli experiment, die wordt uitgevoerd tot men de eerste keer succes behaalt. Geen geheugen betekent dat de kans op mislukking of succes steeds even groot zijn na elk Bernoulli experiment. Ook de exponentiële verdeling heeft geen geheugen.

$$P(X \geq i+j | X \geq i) = \frac{P(X \geq i+j)}{P(X \geq i)} = \frac{q^{i+j}}{q^i} = q^j = P(X \geq j) \text{ voor de geometrische verdeling}$$

$$P(X \geq x_0 + x | X \geq x_0) = \frac{P(X \geq x_0 + x)}{P(X \geq x_0)} = \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda(x_0)}} = e^{-\lambda(x)} = P(X \geq x) \text{ voor de exp. Verdeling}$$

**Geef de centrale limietstelling. Illustreer hoe je deze stelling kunt gebruiken bij het benaderen van een Poissonverdeling.**

Met opmerkingen [gw13]: p5.24-5.27

Centrale limiet stelling  $\rightarrow$  Als  $X_1, X_2, \dots, X_n$  onafhankelijke identiek verdeelde stochastische variabelen zijn met reële variantie  $\sigma^2$  en verwachtingswaarde  $\mu$ .

Als  $Y_n$  hun som is met verwachtingswaarde  $\bar{\mu}_n$  en variantie  $\bar{\sigma}_n^2$ :

$$Y_n = \sum_{j=1}^n X_j, \bar{\mu}_n = n\mu, \bar{\sigma}_n^2 = n\sigma^2$$

Dan convergeert  $Z_n = \frac{Y_n - \bar{\mu}_n}{\bar{\sigma}_n}$  naar een standaard normaal verdeelde variabele:

$$\lim_{n \rightarrow \infty} Z_n = \lim_{n \rightarrow \infty} \frac{Y_n - \mu_n}{\sigma_n} = W \text{ met } W \sim N(0;1)$$

De poisson verdeling kunnen we door de centrale limietstelling als volgt benaderen:

$$\lim_{\lambda \rightarrow \infty} \frac{X_\lambda - \lambda}{\sqrt{\lambda}} \sim N(0;1)$$

Dit impliceert dat we van een Poisson verdeling kunnen over gaan naar een normale verdeling als er aan één voorwaarde is voldaan :  $n \geq 30$

$$P(\lambda) \sim N(\lambda; \lambda)$$

**Geef de centrale limietstelling. Illustreer hoe je deze stelling kunt gebruiken bij het benaderen van een binomiaalverdeling**

De binomiale verdeling kunnen we ook door een normale verdeling benaderen.  $B(n, p) \sim N(np, npq)$

Hierbij moeten ook enkele voorwaarden voldaan zijn:

$$\begin{cases} n \geq 30 \\ np > 5 \\ nq > 5 \end{cases}$$

Als deze niet zijn voldaan benaderen we deze met een Poisson verdeling. Er moet ook rekening worden gehouden met de continuïteitscorrectie.

**Geef de definitie van een statistiek, van een schatter en van een zuivere schatter. Geef tevens een zuivere schatter voor de populatievariantie en bewijs zijn zuiverheid.**

Met opmerkingen [gw14]: 6.2-6.4

Statistiek  $\rightarrow$  stochastische variabele, enkel functie van de steekproef (geen onbekende parameters  $\mu, \sigma$ )

Schatter  $\rightarrow$  statistiek die onbekende parameters benaderd

- Zuivere schatter  $\rightarrow$  een schatter  $T$  van een parameter  $\lambda$  heet zuiver, als  $E[T] = \lambda$ , dus als de verwachtingswaarde van de schatter gelijk is aan de parameter.

Bewijs zuiverheid steekproefgemiddelde  $\bar{X}_n$  voor  $\mu$

Zuivere schatter voor populatievariantie is  $\sigma^2 = n \text{ var}(\bar{x}_n)$

$$n \text{ var}(\bar{x}_n) = n \text{ var}\left(\frac{1}{n} \sum x_i\right) = \frac{1}{n} \sum \text{ var}(x_i) = \frac{1}{n} n \sigma^2$$

**Geef de definitie van een statistiek, van een schatter en van een zuivere schatter. Geef tevens een zuivere schatter voor het populatiegemiddelde en geef de verdeling van deze schatter.**

Idem vorige vraag

Statistiek → stochastische variabele, enkel functie van de steekproef (geen onbekende parameters  $\mu, \sigma$ )

Schatter → statistiek die onbekende parameters benaderd

Zuivere schatter → een schatter  $T$  van een parameter  $\lambda$  heet zuiver, als  $E[T] = \lambda$ , dus als de verwachtingswaarde van de schatter gelijk is aan de parameter.

Steekproefgemiddelde en gewogen gemiddelde zijn zuivere schatters van het populatiegemiddelde. Het steekproefgemiddelde is bernoulli verdeeld  $\sim B(n, p)$  Want  $E[\bar{X}_n] = \mu$

$$E[\bar{X}_n] = \frac{1}{n} \sum x_i = \frac{1}{n} \sum E[x_i] = \frac{1}{n} n\mu$$

**Je neemt een eerste steekproef van grootte  $n$  uit een normaal verdeelde populatie met gemiddelde  $\mu_1$  en variantie  $\sigma^2$ . Je neemt nadien een tweede steekproef van grootte  $m$  uit een andere normaal verdeelde populatie met gemiddelde  $\mu_2$  en een variantie  $\sigma^2$ . Geef de verdeling van het verschil van de 2 steekproefgemiddelden. Gebruik deze verdeling om een toets af te leiden voor het toetsen van de gelijkheid van de twee populatiegemiddelden? Wat kan je doen indien beide populatievarianties verschillend zijn?**

Als de populaties normaal verdeeld zijn dan is het steekproefgemiddelde als volgt normaal verdeelt:

$$\bar{Y}_m - \bar{X}_n \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

$$\text{Oftewel } Z = \frac{(\bar{Y}_m - \bar{X}_n) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

Toets: Ongepaarde t-toets, of F-toets kunnen we veronderstellen dat  $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Met  $\mu_2, p_2$  en  $\sigma^2$  onbekend.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2 \end{cases}$$

We schatten  $\sigma^2$  als  $\bar{S}^2 = \frac{(m-1)s_2^2 + (n-1)s_1^2}{m+n-2}$  (zuivere schatter)

$$\text{Bij nulhypothese } \bar{Y}_m - \bar{X}_n \sim N\left(0; \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right)$$

$$\text{Oftewel } U = \frac{(\bar{Y}_m - \bar{X}_n)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)$$

**Leid een betrouwbaarheidsinterval af voor het populatiegemiddelde indien de populatievariantie gekend is (de populatie is normaal verdeeld). Welke factoren bepalen de lengte van dit interval?**

**Wat verandert er aan de uitdrukking voor dit interval indien de populatievariantie niet gekend is?**

Afleiding zie p6.9-10 in boek

De lengte van het interval wordt bepaald door:

- Populatievariantie (st. afw) ( $\sigma$ )
- Significantieniveau ( $\alpha$ )
- Steekproefgrootte ( $n$ )

Wanneer populatievariantie ( $\sigma^2$ ) niet gekend is dan wordt deze geschat door  $S_n^2$  en krijgen we een t-verdeling i.p.v. een normale verdeling.

**Leid een betrouwbaarheidsinterval af voor de populatievariantie als de populatie normaalverdeeld is.**

Afleiding zie p 7 (boek p 6.14)

**Wat zijn de verschillende stappen bij het toetsen van hypothesen? Illustreer dit aan de hand van een toets voor het gemiddelde indien de variantie ongekend is.**

- 1) Het opstellen van de  $H_0$  en  $H_a$  hypothese.  $H_0$  zal men trachten te verwerpen ten voordele van de alternatieve hypothese
- 2) Het definiëren van een toetsingsgrootte  $T$ , een variabele die men o.b.v. de steekproef kan berekenen.  $T$  moet zodanig zijn dat de verdeling van  $T$  gekend is in de veronderstelling dat  $H_0$  waar is.
- 3) De verdeling van de toetsingsgrootte
- 4) Het beslissingscriterium

Uitgewerkt een éézijdige T-toets voor het gemiddelde indien de variantie ongekend is.

- 1) 
$$\begin{cases} H_0: \mu = E[X] \\ H_a: \mu > E[X] \end{cases}$$
- 2) 
$$T = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$
- 3) 
$$T = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t(n-1)$$
- 4) 
$$\left[ -t_{n-1, 1-\frac{\alpha}{2}}, t_{n-1, 1-\frac{\alpha}{2}} \right]$$

**Wat zijn de verschillende stappen bij het toetsen van hypothesen? Illustreer dit aan de hand van een toets voor de variantie.**

Idem vorige vraag

- 1) Het opstellen van de  $H_0$  en  $H_a$  hypothese.  $H_0$  zal men trachten te verwerpen ten voordele van de alternatieve hypothese
- 2) Het definiëren van een toetsingsgrootheid  $T$ , een variabele die men o.b.v. de steekproef kan berekenen.  $T$  moet zodanig zijn dat de verdeling van  $T$  gekend is in de veronderstelling dat  $H_0$  waar is.
- 3) De verdeling van de toetsingsgrootheid
- 4) Het beslissingscriterium

Uitgewerkt een toets voor de variantie (de tweezijdige  $\chi^2$  - toets):

- 1) 
$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_a: \sigma^2 \neq \sigma_0^2 \end{cases}$$
- 2) 
$$Y = \frac{(n-1)S_n^2}{\sigma_0^2}$$
- 3)  $\chi_{n-1}^2$  verdeeld
- 4)  $\left[ \chi_{n-1, \frac{\alpha}{2}}^2, \chi_{n-1, 1-\frac{\alpha}{2}}^2 \right]$

**Welke toetsen ken je om gemiddeldes van 2 verschillende populaties met elkaar te vergelijken? Geef een kort overzicht. Vermeld tevens wanneer je welke toets gebruikt.**

- F-toets: vergelijken van varianties in 2 normaal verdeelde onafhankelijke steekproeven met onbekende parameters
- Ongepaarde t-toets: vergelijken van gemiddeldes in 2 normaal verdeelde onafhankelijke steekproeven met onbekende parameters. Hier wordt aangenomen dat de varianties gelijk zijn (dit moet eerst worden nagegaan met een  $F$  - toets)
- Gepaarde t-toets: vergelijken van gemiddeldes in 2 afhankelijke steekproeven

**Bij het toetsen van hypothesen werden begrippen als kritisch punt, significantieniveau  $\alpha$  en een p-waarde geïntroduceerd. Wat betekenen deze begrippen? Illustreer dit aan de hand van een toets voor het vergelijken van varianties in 2 verschillende normaal verdeelde populaties.**

Kritisch punt  $\rightarrow$  punt dat de scheiding aangeeft tussen het aanvaardingsgebied en het verwerpingsgebied.

Significantieniveau  $\rightarrow$  de kans om een type I fout te maken.

p-waarde  $\rightarrow$  (overschrijdingskans) de kans om, als de nulhypothese waar is, uitkomsten te zien die "even extreem" zijn als de observatie die we hebben waargenomen.

Illustratie: tweezijdige F-toets :

$$\frac{s_1^2}{s_2^2} \sim F_{m-1, n-1} \text{ met } H_0: \sigma_1^2 = \sigma_2^2$$

$$\text{Dus onder de nulhypothese : } f = \frac{s_1^2}{s_2^2} \sim F_{m-1, n-1}$$

Kritisch punt:  $\left[ F_{m-1, n-1, \frac{\alpha}{2}}, F_{m-1, n-1, 1-\frac{\alpha}{2}} \right]$

Significantieniveau: Zelf te kiezen

p-waarde =  $2\min\{P(F < f), P(F > f)\} = 2\min\{F_{m-1, n-1}(f); 1 - F_{m-1, n-1}(f)\}$

**Definieer de correlatiecoëfficiënt van 2 steekproeven en van 2 stochastische variabelen. Beschrijf de toets die je kunt uitvoeren om te testen of de correlatie significant is. Wat betekent dit?**

$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E[X])(Y - E[Y])]}{\sigma_X \sigma_Y}$  voor 2 stochastische variabelen

$r(x, y) = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$  voor 2 steekproeven

Je doet een t-toets uit (kijken dat er geen correlatie is tussen X en Y op een bepaald sign niv)

$$H_0: \rho = 0$$

$H_a: \rho \neq 0$  er is een significant lineair verband tussen X en Y in de populatie

$$T = \rho \sqrt{\frac{n-2}{1-\rho^2}} \sim t_{n-2}$$

BI:  $\left[ t_{n-2, \frac{\alpha}{2}}, t_{n-2, 1-\frac{\alpha}{2}} \right]$

Kritische waarde:  $p = 2F_{t_{n-2}}^{-1}(-|t|)$

De correlatiecoëfficiënt ligt dicht bij -1 als er een negatief lineair verband is en dicht bij 1 als er een sterk positief lineair verband is.

**Welke toetsen ken je om op basis van een steekproef uit je populatie na te gaan of je populatie uniform verdeeld is? Geef hun voor- en nadelen. Beschrijf de Kolmogorov-Smirnov toets.**

- P-P-plot (theoretisch kans t.o.v. empirische kans) dicht genoeg bij rechte y=x  
Nadeel: je hebt een programma nodig om deze plot te tekenen  
Het al dan niet dicht bij de rechte liggen is subjectief
- Q-Q-plot (theoretische kwartielen uitgezet t.o.v. empirische waarde  $y_i$ )  
Nadeel: je hebt een programma nodig om deze plot te tekenen  
Het al dan niet dicht bij de rechte liggen is subjectief  
dicht genoeg bij rechte y=x
- Kolmogorov-Smirnov toets
- Chi-kwadraat toets

Kolmogorov-Smirnov toets:

Deze toets vergelijkt rechtstreeks de theoretische verdelingsfunctie F (bv een normale verdeling, hier uniforme verdeling) met de empirische verdelingsfunctie  $F_n$  van de verwachte verdeling. Het verschil

hiertussen moet minimaal zijn. Deze toets is handig bij kleinere steekproeven omdat daar een  $\chi^2$  – toets minder is aangewezen.

$$d_n \dots$$

$$\begin{cases} H_0: X = F \\ H_a: X \neq F \end{cases}$$

Nadeel : je hebt een tabel nodig om de kritische grenzen op te zoeken.

Voordeel : significantieniveau kan zelf gekozen W.

Bruikbaar bij kleinere steekproeven.

**Welke toetsen ken je om op basis van een steekproef uit je populatie na te gaan of je populatie uniform verdeeld is. Geef hun voor- en nadelen. Beschrijf de chi-kwadraat toets.**

Idem hierboven

Chi-kwadraat toets:

Geobserveerde frequenties worden vergeleken met verwachte frequenties. p.6.39-6.42

Eerst worden  $\bar{X}$  en  $\sigma^2$  bepaald om de verwachte freq te bepalen.

De toetsingsgrootheid  $T = \sum \frac{(O_i - E_i)^2}{E_i}$  met de som over k

BI :  $[0, \chi^2_{k-1, 1-\alpha}]$  met k aantal klassen

Of als we enkel voor uniforme doen, verdelen we onze meetwaarden in gelijke klassen. Als in elke klasse evenveel meetwaarden zitten is het uniform verdeeld.

**Wat is een kruistabel? Leid een toets af waarmee je de onafhankelijkheid van de kolom- en de rijvariabele kunt nagaan?**

Kruistabel → de weergave in tabel van telgegevens die volgens 2 variabelen zijn geclassificeerd.

$$\begin{cases} H_0: \text{Er is geen afhankelijkheid tussen rij en kolom} \\ H_a: \text{er is wel afhankelijkheid} \end{cases}$$

Toetsingsgrootheid  $T = \sum_i \frac{(O_i - E_i)^2}{E_i}$

BI :  $[0, \chi^2_{k-1, 1-\alpha}]$  met k aantal klassen

Welke types fouten heb je bij het toetsen van hypothesen. Wat is de macht van een toets? Leid de macht van de eenzijdige t-toets af. Welke factoren bepalen deze grootheid? Schetsdit.

|                            | $H_0$ is waar                         | $H_a$ is waar                        |
|----------------------------|---------------------------------------|--------------------------------------|
| $H_0$ wordt verworpen      | Type I fout<br>Kans $\alpha$          | Correcte uitspraak<br>Kans $1-\beta$ |
| $H_0$ wordt niet verworpen | Correcte uitspraak<br>Kans $1-\alpha$ | Type II fout<br>Kans $\beta$         |

Schets p 6.47

- Je kan de significantie veranderen : hoe kleiner significantie hoe kleiner de power
- De werkelijke verdeling : hoe groter het verschil ( $\Delta\mu$ ), hoe groter de power
- Grootte v/d steekproef aanpassen : hoe groter, hoe smaller  $\rightarrow$  minder overlap  $\rightarrow$  power stijgt

AFL:

$$\begin{aligned}
 H_0: \mu &= \mu_0 \\
 H_a: \mu &\neq \mu_0 \\
 T &= \frac{(\bar{X}_n - \mu_0)}{S_n/\sqrt{n}} \sim t_{n-1} \\
 BI &: [t_{n-1,1-\alpha}, +\infty[
 \end{aligned}$$

$$\begin{aligned}
 \text{De macht v/d toets } 1-\beta &= P(H_0 \text{ verworpen} | H_a) = P\left(\bar{X}_n > \mu_0 + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}} \mid H_a\right) \\
 &= 1 - F_{t_{n-1}}\left(\frac{\mu_0 - \mu_1 + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}}}{\frac{s_n}{\sqrt{n}}}\right) \\
 &= F_{t_{n-1}}\left(-t_{n-1,1-\alpha} + \frac{\mu_1 - \mu_0}{\frac{s_n}{\sqrt{n}}}\right)
 \end{aligned}$$

Leid de uitdrukkingen af voor de richtingscoëfficiënt en het snijpunt met de Y-as bij lineairregressie (kleinste kwadraten methode). Is er een verband met de covariantie?

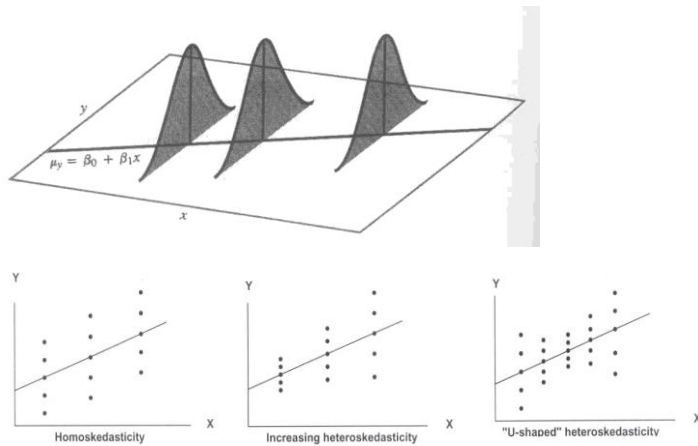
Afleiding zie boek 7.4

$b_1 = \frac{\text{Cov}(x,y)}{S_x^2}$  (dit is de correlatiecoëfficiënt) dus indien de mate van het lineair verband klein is zal de covariantie klein zijn en zal de regressierechte praktisch horizontaal zijn.

**Aan welke voorwaarden moet er voldaan zijn om de kleinste kwadraten toe te passen. Leg uit!**

- 1) Voor elke  $x_i$  is het residu een trekking uit een normaal verdeelde stochastiek met gemiddelde nul. M.a.w. geen systematische fout.
- 2) De residu's zijn onafhankelijk
- 3) De residu's bij verschillende waarden van  $x_i$  hebben dezelfde variantie. (homoscedasticiteit)

*"One picture is worth a thousand words"*



**Hoe kun je het lineaire kleinste kwadraten model verifiëren? Leg uit.**

- Testen op Lack Of Fit (LOF): residuele plot
- Normaliteit van de residus: een KS-toets of chi-kwadraat toets uitvoeren en residuele plot
- Homoscedasticiteit: residuele plot
- Analysis Of VAriance (ANOVA)

Als je naar de residuele plot kijkt mag je geen patronen opmerken. Als er een patron te zien is heb je ofwel te maken met een LOF, ofwel met heteroscedasticiteit. Een systematische fout kan hier ook aan de basis van liggen.

**Wat is de determinatiecoëfficiënt bij lineaire regressie? Hoe kun je de significantie van je regressie toetsen?**

$$\text{Determinatiecoëfficiënt } (R^2) \rightarrow R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} R \in [0,1]$$

Hoe regressie testen? P7.7-7.8

**In dien je een lineair model fit, krijg je soms te maken met outliers. Wat zijn dit? Welk effect hebben ze? Wat kun je doen?**

Grote impact op de lineaire regressie. Gebruik van robuuste lineaire regressie methode zoals de mediaan. Je kan kijken of je geen tikfouten hebt gemaakt of andere eenvoudige fouten zoals het werken met een ander toestel, een andere temperatuur,...

**Bij het toetsen van hypothesen werden begrippen als kritisch punt, significantieniveau  $\alpha$  en een p-waarde geïntroduceerd. Wat betekenen deze begrippen? Illustreer dit aan de hand van een toets om te toetsen of het populatiegemiddelde van een normaal verdeelde populatie gelijk is aan een bepaald getal  $a$ . Maak een grafiek waarop je al de bovenstaande begrippen aanduidt.**

Kritisch punt → punt dat de scheiding aangeeft tussen het aanvaardingsgebied en het verwerpingsgebied.

Significantieniveau → de kans om een type I fout te maken.

p-waarde → de kans om, als de nulhypothese waar is, uitkomsten te zien die “even extreem” zijn dan de observatie die we hebben waargenomen.

Schets zie achterkant p 6.24