

Voorbeeldvragen:

H1

Gegevens kunnen gemeten worden op verschillende types schalen: welke ?

Nominale, ordinale, interval en verhoudingsschaal.

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.

Verhoudingsschaal

H2

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

- **bereik/range** = hoogste – laagste waarde
 - **variantie** :
 - $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$ voor een populatie
 - $s_n^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ voor een steekproef
 - **standaardafwijking** : wortel uit de variantie
 - **mediaan en gemiddelde absolute afwijking** (MAD en MeanAD)
 - $MAD = \text{mediaan } \{|x_i - med|\}_{i=1}^n$
 - $MeanAD = \frac{1}{n} \sum_{i=1}^n |x_i - med|$
-

Hoe maak je een boxplot ?

We tekenen een as gaande van de kleinste naar de grootste meting (het bereik). Op deze as worden de mediaan, het eerste en het derde kwartiel aangegeven met een dwarse streep. Van het stuk tussen het eerste en het derde kwartiel wordt een doosje (box) gemaakt. Er kunnen uitschieters voorkomen. Dit komt wanneer een datapunt kleiner is dan $Q1 - 1,5 * \text{interkwartiel (IQR)}$ of groter dan $Q3 + 1,5 * \text{IQR}$ en die duiden we aan met een sterretje. De as gaat van de kleinste naar de grootste meting, die geen uitschieter is.

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

- **gemiddelde** : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- $med = \begin{cases} \frac{x_{n+1}}{2} & n \text{ oneven} \\ \frac{1}{2} \left(\frac{x_n}{2} + \frac{x_{n+1}}{2} \right) & n \text{ even} \end{cases}$
- **mediaan** :
- **modus** : de waarde die het meest voorkomt
- **midrange** : gemiddelde van hoogste en laagste waarde, dus: $\frac{x_1 + x_n}{2}$
- **gewogen gemiddelde** : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i w_i$

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

- **z-score** : $z = \frac{x_i - \bar{x}}{s}$, voor populaties : $Z = \frac{X - \mu}{\sigma}$
- **percentielen, kwantielen en interkwartiel** : $\frac{\alpha}{100}(n+1)$

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: f_i / n , de dichtheidsschaal, som van hoogtes = 1.

Cumulatieve frequentie: som van de tot-nu-toe behandelde waarden.

Met behulp van een **histogram** voorstellen. Ook met **polygoon** en **ogieven**.

De **empirische verdelingsfunctie** stelt het aantal metingen voor met een waarde kleiner dan of gelijk aan een bepaalde waarde (x) gedeeld door het totale aantal metingen:

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

H3

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

De verzameling van alle mogelijke uitkomsten van een experiment noemen we het **universum** en we noteren het door Ω (hoofdletter omega).

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

Een **kans** (waarschijnlijkheid, probabiliteit) is een functie waarbij met elke gebeurtenis A van een universum Ω een reëel getal $P(A)$ geassocieerd wordt.

Dit getal $P(A)$ moet aan de volgende **regels** voldoen:

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

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

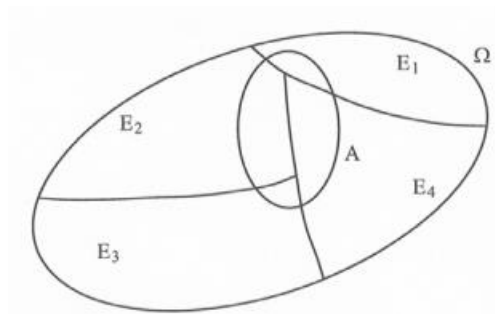
We noemen dit de **voorwaardelijke kans** op het optreden van gebeurtenis V als gebeurtenis B heeft plaatsgevonden (en $P(B) \neq 0$).

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

Wet der totale kans

Onderstel dat $E_1, E_2, \dots, E_n$ een partitie is van het universum Ω (wat wil zeggen dat $E_i \cap E_j = \emptyset$ voor $i \neq j$ en dat $\bigcup_{i=1}^n E_i = \Omega$). Neem nu een willekeurige gebeurtenis A , dan geldt:

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



Bewijs:

$$\begin{aligned}
 A &= A \cap [E_1 \cup E_2 \cup \dots \cup E_n] = [A \cap E_1] \cup \dots \cup [A \cap E_n] \\
 \uparrow & \quad \quad \quad \text{— } E_i \text{'s zijn disjunct} \\
 A &= A \cap \Omega = \quad \quad \quad (A \cap E_i) \text{ 2 aan 2 disjunct} \\
 P(A) &= P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n) \\
 &= P(A|E_1)P(E_1) + \dots + P(A|E_n)P(E_n) \\
 &= \sum_{i=1}^n P(A|E_i)P(E_i)
 \end{aligned}$$

Regel van Bayes

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

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

Bewijs: m.b.v. de produktregel en de wet der totale kans

Formuleer de regel van Bayes en illustreer deze wet met de hulp van een voorbeeld en een figuur. (zie vorige vraag)

Wat betekent het indien twee of meerdere gebeurtenissen onafhankelijk zijn? Geef één gevolg. Twee gebeurtenissen A en B heten **(stochastisch) onafhankelijk** als het voor de kans op A niet uitmaakt of B al dan niet gebeurd is. $P(A) = P(A|B) = P(A|B^c)$

Gevolg A en B onafhankelijk $\Leftrightarrow P(A \cap B) = P(A)P(B)$

We noemen bovenstaande regel de **produktregel** voor twee onafhankelijke gebeurtenissen.

H4

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

Wiskundig gebeurt de overgang van het universum Ω naar $\mathbb{R}$ door een reële functie X te definiëren: $X: \Omega \rightarrow \mathbb{R}: \omega \rightarrow X(\omega)$. In de kansentheorie wordt zo'n functie X een **stochastische veranderlijke/variabele** genoemd.

Als X een stochastische variabele is, dan heet de functie F_X : $F_X(a) = P(X \leq a)$ de **verdelingsfunctie** van X .

De **dichtheidsfunctie** is de afgeleide van de verdelingsfunctie. De eis is dat deze overal continu is en bovendien overal differentieerbaar (behalve eventueel in een eindig aantal punten).

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



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:

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

(

Gevolg 1

De componenten van een tweedimensionale kansvector $\mathbf{Z} = (X, Y)$ zijn onafhankelijk als en slechts als de verdelingsfunctie van $\mathbf{Z}$ het product is van de marginale verdelingsfuncties:

$$F_Z(x, y) = F_X(x) F_Y(y)$$

Gevolg 2

De componenten van een tweedimensionale continue kansvector $\mathbf{Z} = (X, Y)$ zijn onafhankelijk als en slechts als de dichtheidsfunctie van $\mathbf{Z}$ het product is van de marginale dichtheidsfuncties:

$$f_Z(x, y) = f_X(x) f_Y(y)$$

Stelling: $\text{Var}[X] = E[(X - E[X])^2] = \text{Var}[X] = E[X^2] - E[X]^2$

Bewijs:

$$\begin{aligned} \text{Var}[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2 - 2E[X]X] + E[X]^2 \\ &= E[X^2] + E[-2E[X]X] + E[X]^2 \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

)

$$- \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$$

Bewijs :

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy$$

Omdat X en Y onafhankelijk zijn, kunnen we $f_{X,Y}(x,y)$ vervangen door $f_X(x)f_Y(y)$:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy$$

Nadien kunnen we de integralen scheiden en we bekomen $\mathbb{E}[X] \mathbb{E}[Y]$.

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

Bewijs :

$$\begin{aligned} \text{Var}[X + Y] &= \mathbb{E}[(X + Y - \mathbb{E}[X + Y])^2] = \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\mathbb{E}[XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] - 2\mathbb{E}[Y]\mathbb{E}[X] + 2\mathbb{E}[X]\mathbb{E}[Y] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] \\ &= \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

$$- \left(\begin{aligned} \text{Var}[X - Y] &= \text{Var}[X] + \text{Var}[Y] \end{aligned} \right.$$

Bewijs :

$$\begin{aligned} \text{Var}[X - Y] &= \text{Var}[X + (-Y)] \\ &= \text{Var}[X] + \text{Var}[(-1)Y] \\ &= \text{Var}[X] + (-1)^2 \text{Var}[Y] = \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

)

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.

Dit wilt zeggen dat er geen verandering is van een kans gedurende een experiment. Bijvoorbeeld: het gooien van een muntstuk en het aantal kruis en munt optellen.

De exponentiële verdeling heeft ook geen geheugen.

Bewijs:

$$P(X \geq x_0 + x | X \geq x_0) = e^{-\lambda x} = P(X \geq x)$$

$$\begin{aligned} LL &= \frac{P(X \geq x_0 + x \text{ en } X \geq x_0)}{P(X \geq x_0)} = \frac{P(X \geq x_0 + x)}{P(X \geq x_0)} = \frac{e^{-\lambda(x_0 + x)}}{e^{-\lambda x_0}} \\ &= e^{-\lambda x} = P(X \geq x) \end{aligned}$$

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?

Neem 2 stochastische variabelen X en Y. Als X en Y onafhankelijk zijn, dan geldt, zoals we gezien hebben, dat $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$. In het algemeen (voor afhankelijke stochastieken) hebben we:

$$\begin{aligned}\sigma_{X+Y}^2 &= E[(X+Y - E[X+Y])^2] \\ &= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\ &= \sigma_X^2 + \sigma_Y^2 + 2E[(X - E[X])(Y - E[Y])]\end{aligned}$$

De term $2E[(X - E[X])(Y - E[Y])]$ geeft dus een idee van de mate van onderlinge afhankelijkheid van X en Y. Dit leidt tot de begrippen covariantie en correlatie.

Definitie

De covariantie van twee stochastische variabelen wordt gegeven door:

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

Definitie

De correlatiecoëfficiënt van twee stochastische variabelen wordt gegeven door:

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

Gevolg

De correlatiecoëfficiënt is begrensd: $1 \geq \rho \geq -1$

Twee onafhankelijke stochastische variabelen zijn ongecorreleerd. Het omgekeerde geldt echter niet zoals blijkt uit het volgende voorbeeld.

Voorbeeld

$$Z = (X, Y) \text{ met dichtheidsfunctie: } f_Z(x, y) = \begin{cases} 1/\pi & \text{als } x^2 + y^2 \leq 1 \\ 0 & \text{als } x^2 + y^2 > 1 \end{cases}$$

X en Y zijn niet gecorreleerd want $E[XY] = 0$ en $E[X] = 0 = E[Y]$

X en Y zijn niet onafhankelijk want $P(X > \sqrt{2}/2, Y > \sqrt{2}/2) = 0$ en

$$P(X > \sqrt{2}/2) P(Y > \sqrt{2}/2) \neq 0$$

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

Als $X_1, X_2, \dots, X_n$ onafhankelijke identiek verdeelde stochastische variabelen zijn met reële variantie σ^2 en verwachtingswaarde μ . Als Y_n hun som is met verwachtingswaarde μ_n en

variantie σ_n^2 : $Y_n = \sum_{j=1}^n X_j$, $\mu_n = n\mu$, $\sigma_n^2 = n\sigma^2$ dan convergeert $Z_n = \frac{Y_n - \mu_n}{\sigma_n}$ naar een

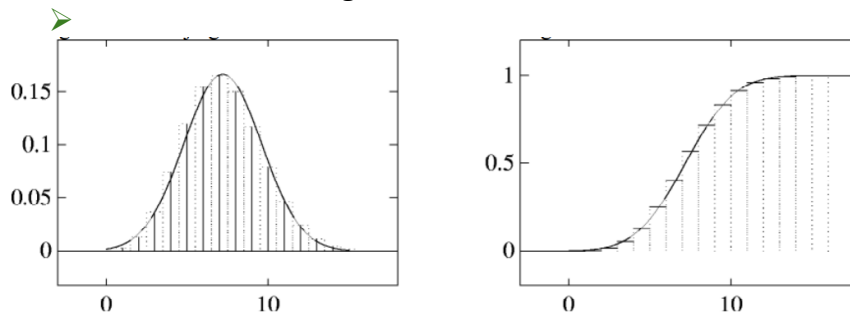
$$\lim_{n \rightarrow \infty} Z_n = \lim_{n \rightarrow \infty} \frac{Y_n - \mu_n}{\sigma_n} = W \quad \text{met } W \sim \mathcal{N}(0; 1)$$

standaard normaal verdeelde variabele:

Op grond van de centrale limietstelling en de eigenschap $P(\lambda + \mu) = P(\lambda) + P(\mu)$ voor onafhankelijke Poisson verdelingen, weten we dat de Poisson verdeling zelf voor grote waarden van λ naar de normale verdeling convergeert. Als $X_\lambda \sim P(\lambda)$, dan geldt $E[X_\lambda] = \lambda$ en $\text{Var}[X_\lambda] = \lambda$ zodat:

$$\lim_{\lambda \rightarrow \infty} \frac{X_\lambda - \lambda}{\sqrt{\lambda}} \sim \mathcal{N}(0; 1) \quad \text{ofwel } P(\lambda) \approx \mathcal{N}(\lambda; \lambda) \text{ als } \lambda \text{ voldoende groot.}$$

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



In het algemeen moeten we dus voor $X \sim B(n;p)$ en $Y \sim \mathcal{N}(np; np(1-p))$ de volgende benadering gebruiken:

$$P(j \leq X \leq k) \approx P(j-0.5 < Y < k+0.5)$$

$$P(X \leq 0) \approx P(Y \leq 0.5) \text{ en } P(X \geq n) \approx P(Y \geq n-0.5)$$

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.

Een **statistiek** is een stochastische variabele die alleen een functie is van de steekproef en niet van onbekende parameters (zoals μ en σ).

Een **schatter** is een statistiek die gebruikt wordt om een onbekende parameter te benaderen. Een schatting is de getalwaarde van de schatter in een concreet experiment.

Een schatter T van een parameter λ heet **zuiver** (Eng. unbiased), als $E[T] = \lambda$, dus als de verwachtingswaarde van de schatter gelijk is aan de (gezochte) parameter.

De steekproefvariantie is een zuivere schatter van de populatievariantie.

Bewijs:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n ((X_i - \mu) - (\bar{X}_n - \mu))^2 \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - \sum_{i=1}^n (\bar{X}_n - \mu)^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - E[X_i])^2 - n(\bar{X}_n - E[\bar{X}_n])^2 \right] \end{aligned}$$

Dus:

$$\begin{aligned} E[S_n^2] &= \frac{1}{n-1} \left[\sum_{i=1}^n E[(X_i - E[X_i])^2] - nE[(\bar{X}_n - E[\bar{X}_n])^2] \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n \text{Var}(X_i) - n\text{Var}(\bar{X}_n) \right] \\ &= \frac{1}{n-1} \left[n\sigma^2 - n\frac{\sigma^2}{n} \right] = \sigma^2 \end{aligned}$$

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.

De verwachtingswaarde van het steekproefgemiddelde is het populatiegemiddelde: $E[\bar{X}_n] = \mu$

Bewijs:

$$E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n \mu = \mu$$

De variantie van het steekproefgemiddelde is gelijk aan de populatievariantie gedeeld door de

steekproefgrootte: $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

Je neemt een eerste steekproef van grootte n uit een normaal verdeelde populatie met gemiddelde μ_1 en variantie σ_1^2 . Je neemt nadien een tweede steekproef van grootte m uit een andere normaal verdeelde populatie met gemiddelde μ_2 en een variantie σ_2^2 . Geef de verdeling van het verschil van de 2 steekproefgemiddelden. Hoe kun je deze verdeling gebruiken bij het toetsen van de gelijkheid van de twee populatiegemiddeldes?

Wanneer de onderliggende populaties normaal verdeeld zijn;

$$X \sim N(\mu_1, \sigma_1^2) \quad \text{en} \quad Y \sim N(\mu_2, \sigma_2^2)$$

$$\text{dan geldt: } \bar{Y}_{n_2} - \bar{X}_{n_1} \sim N\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\text{ofwel: } \frac{(\bar{Y}_{n_2} - \bar{X}_{n_1}) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Wanneer de steekproeven groot zijn (>30) dan mag je bovenstaande eigenschap ook gebruiken, zelfs als de onderliggende populaties niet normaal verdeeld zijn.

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?

Het 2-zijdig betrouwbaarheidsinterval (BI) voor μ op het onzekerheidsniveau α of het betrouwbaarheidsinterval met betrouwbaarheid $1-\alpha$ (confidence interval) is het interval:

$$\left[\bar{x}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}} \right]$$

De factoren zijn dus het gemiddelde, α , σ en het aantal metingen.

Als σ niet gekend is, wordt deze verandert met s_n (het is een zuivere schatter).

Leid een betrouwbaarheidsinterval af voor de populatievariantie als de populatie normaal verdeeld is.

Als een populatie normaal verdeeld is ($X \sim N(\mu, \sigma^2)$), dan is de stochastiek $\frac{(n-1)s_n^2}{\sigma^2}$ chi-kwadraat verdeeld is met $n-1$ vrijheidsgraden: $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

Het betrouwbaarheidsinterval voor σ^2 met betrouwbaarheid $1-\alpha$ is het interval:

$$\left[\frac{(n-1)s_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)s_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right]$$

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.

➤ t-toets

1. Opstellen theoretisch model: $X = \text{IQ-score van nieuwe eerstejaars}$; $E[X] = \mu$
2. De hypothesen:

Nulhypothese	$H_0: \mu = \dots$
Alternatieve hypothese	$H_A: \mu > \dots$
3. Keuze van de toetsingsgrootheid

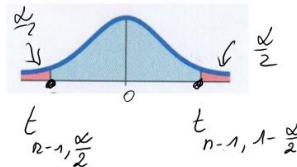
Toetsingsgrootheid $T = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}} \sim t(n-1) \text{ onder } H_0$ t-toets

(Populatie normaal verdeeld of $n > 30$)

- De tweezijdige toets

$H_0: \mu = \mu_0$
 $H_A: \mu \neq \mu_0$
 aanvaardingsgebied voor t $\rightarrow \left[-t_{n-1, 1-\frac{\alpha}{2}}, t_{n-1, 1-\frac{\alpha}{2}} \right]$

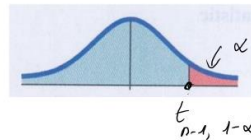
overschrijdingskans $p = 2 P(T \geq |t|)$



- De eenzijdige toets

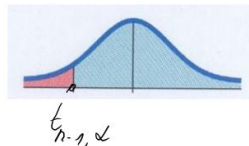
i) $H_0: \mu = \mu_0$
 $H_A: \mu > \mu_0$
 aanvaardingsgebied voor t $\rightarrow \left[-\infty, t_{n-1, 1-\alpha} \right]$

overschrijdingskans $p = P(T \geq t)$



ii) $H_0: \mu = \mu_0$
 $H_A: \mu < \mu_0$
 aanvaardingsgebied voor t $\rightarrow \left[t_{n-1, \alpha}, +\infty \right]$

overschrijdingskans $p = P(T \leq t)$



4. Beslissingscriterium:

Gebied waar t met kans 0.05 terecht komt ?

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

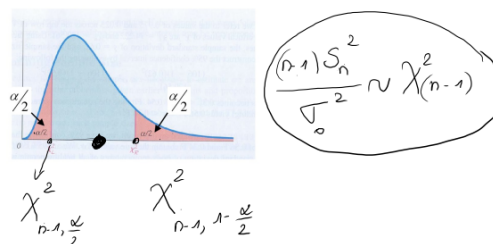


Toetsingsgrootheid $\frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2(n-1) \text{ onder } H_0$ χ^2 -toets

- De tweezijdige toets

$H_0: \sigma^2 = \sigma_0^2$
 $H_A: \sigma^2 \neq \sigma_0^2$
 aanvaardingsgebied voor $\chi^2 \rightarrow \left[\chi_{n-1, \frac{\alpha}{2}}^2, \chi_{n-1, 1-\frac{\alpha}{2}}^2 \right]$

overschrijdingskans $p = 2 \min\{P(Y \geq \chi), P(Y \leq \chi)\}$



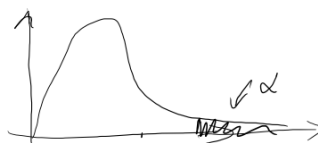
- De eenzijdige toets

i) $H_0: \sigma^2 = \sigma_0^2$
 $H_A: \sigma^2 > \sigma_0^2$
 aanvaardingsgebied voor $\chi^2 \rightarrow \left[0, \chi_{n-1, 1-\alpha}^2 \right]$

overschrijdingskans $p = P(Y \geq \chi)$

ii) $H_0: \sigma^2 = \sigma_0^2$
 $H_A: \sigma^2 < \sigma_0^2$
 aanvaardingsgebied voor $\chi^2 \rightarrow \left[\chi_{n-1, \alpha}^2, +\infty \right]$

overschrijdingskans $p = P(Y \leq \chi)$



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.

De **ongepaarde toets**: de t-toets, de **gepaarde toets** voor twee gemiddelden (gepaarde t-toets). Als er met gekoppelde data gewerkt wordt (en als beide steekproeven even groot zijn), dan kunnen we een gepaarde t-toets uitvoeren.

Bij het toetsen van hypothesen werden begrippen als kritisch punt, significantieniveau α 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.

Het punt dat de scheiding aangeeft tussen het aanvaardingsgebied en het verwerpingsgebied wordt het **kritisch punt** genoemd.

Het **significantieniveau α** wordt gedefinieerd als de kans om een type I fout te maken. Bij het uitvoeren van een toets zegt men dat “de nulhypothese (niet) verworpen wordt op significantieniveau α ”.

De **p-waarde** is de kans om, als de nulhypothese waar is, uitkomsten te zien die “even extreem zijn als of nog extremer dan” de observatie die we hebben waargenomen

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 ?

De **correlatiecoëfficiënt** van twee stochastische variabelen wordt gegeven door

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

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

Om te testen of we op basis van onze steekproef mogen besluiten, met een bepaald significantieniveau, dat er geen correlatie is tussen X en Y voeren we een **t-toets** uit.

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 $y=x$
Voordeel: eenvoudig te interpreteren, directe vergelijking van cumulatieve kansen.
Nadeel : je hebt een programma nodig om deze plot te tekenen, afhankelijk van de steekproefgrootte
Het al dan niet dicht bij de rechte liggen is subjectief
- **Q-Q-plot** (theoretische kwartielen uitgezet t.o.v. empirische waarde y_1)
Nadeel: je hebt een programma nodig om deze plot te tekenen. Het al dan niet dicht bij de rechte liggen is subjectief
Voordeel: breed toepasbaar
Nadeel: afhankelijk van de steekproefgrootte

- **Kolmogorovo-Smirnov toets**

Voordeel: kan gebruikt worden bij kleinere steekproeven

Nadeel: minder krachtig bij grote steekproeven

- **Chi-kwadraat toets**

Voordeel: breed toepasbaar, kan gebruikt worden om na te gaan of steekproef data verdeeld zijn volgens een bepaalde verdeling

Nadeel: strenge voorwaarden (data zijn frequenties, totaal aantal waarnemingen n moet groter zijn dan 20, verwachte frequentie in elke klasse moet groter zijn dan 5)

Kolmogorovo-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 χ^2 -toets minder is aangewezen.

Voordeel: significantieniveau kan zelf gekozen worden.

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

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.

De chi-kwadraat toets wordt gebruikt om geobserveerde frequenties te vergelijken met verwachte frequenties.

We kunnen dus vragen beantwoorden zoals:

- Hoe goed passen de geobserveerde gegevens bij een gepostuleerde verdelingsfunctie F ? Zijn de gegevens normaal verdeeld, uniform verdeeld, Poisson verdeeld, ...?
 - Zijn twee eigenschappen in een populatie onafhankelijk ?
-

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

Indien we over telgegevens beschikken, die volgens twee (kwalitatieve) variabelen zijn geclassificeerd, kunnen we deze gemakkelijk in een tabel weergeven. We noemen deze tabel een **kruistabel**.

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 grootte ? Schets dit.



	H_0 is waar	H_a is waar
H_0 wordt verworpen	Type I fout Kans α	Correcte uitspraak Kans $1-\beta$
H_0 wordt niet verworpen	Correcte uitspraak Kans $1-\alpha$	Type II fout Kans β

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

AF1:

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

$$T = \frac{(\bar{X}_n - \mu_0)}{S_n/\sqrt{n}} \sim t_{n-1}$$

$$BI : [t_{n-1,1-\alpha}, +\infty[$$

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

$$= 1 - F_{t_{n-1}}\left(\frac{\mu_0 - \mu_1 + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}}}{\frac{s_n}{\sqrt{n}}}\right)$$

$$= F_{t_{n-1}}\left(-t_{n-1,1-\alpha} + \frac{\mu_1 - \mu_0}{\frac{s_n}{\sqrt{n}}}\right)$$

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



Afleiding zie boek 7.4

$b_1 = \frac{Cov(x,y)}{S_x^2}$ (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.

- voor elke x_i is het residu een trekking uit een normaal verdeelde stochastiek met gemiddelde nul
- de residus zijn onafhankelijk
- de residus bij verschillende waarden van x_i hebben dezelfde variantie (homoscedasticiteit)

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

- residuële plots: scatter plot van ei in functie van yi of xi
 - de variantieanalyse (F-toets): analysis of variance (ANOVA) en de determinatiecoëfficiënt R²
 - t-toets op rico = 0 (betrouwbaarheidsintervallen)
-

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

De determinatiecoëfficiënt R² wordt nu gedefinieerd als de proportie van de totale variatie in de afhankelijke variabele y die door de regressie verklaard wordt:

$$R^2 = \frac{SS_{\text{Reg}}}{SS_T} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Leg volgende grafiek uit ? Hoe kun je aan de hand van de uitdrukkingen voor de BI's een goed experiment opzetten ? Geef de voor- en de nadelen.

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 ?

Outliers zijn aberante metingen, uitschieters. Ze hebben een 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, ...

Voorbij examenvragen:

Vraag 1

/

Vraag 2

- Leid de chi-kwadraat toets af.
- Toon aan dat een steekproef van grootte 100 uniform verdeeld is op het interval [0,4]. Leg gedetailleerd uit.
- Geef de andere methoden (ook grafische) om te testen of een steekproef een bepaalde verdeling volgt. Geef de voor- en nadelen ervan.

Vraag 3

/

Vraag 1

/

Vraag 2

/

Vraag 3

- Geef de voorwaarden voor de kleinste kwadraten methode.
 - Leid de regressiecoëfficiënten af met een duidelijke tekening waarop alle gebruikte symbolen worden aangeduid.
-

Vraag 1

Disciplines stat + teken

Vraag 2

Definitie kans + bewijs

Vraag 3

Chebyshev + bewijs

Stelling: formule van Chebyshev

Als X een stochastische variabele is met verwachtingswaarde $\alpha_1(X)$ en variantie $\mu_2(X)$, dan geldt voor elke $\lambda > 0$:

$$P(|X - \alpha_1(X)| \geq \lambda) \leq \frac{\mu_2(X)}{\lambda^2}$$

Bewijs:

$$\begin{aligned} \mu_2(X) &= \text{Var}[X] = \int_{-\infty}^{+\infty} (x - \alpha_1)^2 f_X(x) dx \\ &\geq \int_{-\infty}^{\alpha_1 - \lambda} \dots + \int_{\alpha_1 + \lambda}^{+\infty} \dots \\ &\geq \int_{-\infty}^{\alpha_1 - \lambda} \lambda^2 f_X(x) dx + \int_{\alpha_1 + \lambda}^{+\infty} \lambda^2 f_X(x) dx \\ &= \lambda^2 \left\{ \int_{-\infty}^{\alpha_1 - \lambda} f_X(x) dx + \int_{\alpha_1 + \lambda}^{+\infty} f_X(x) dx \right\} \\ &= \lambda^2 P(|X - \alpha_1| \geq \lambda) \\ \frac{\mu_2(X)}{\lambda^2} &\geq P(|X - \alpha_1| \geq \lambda) \end{aligned}$$

1^o integraal : $x \leq \alpha_1 - \lambda$
 $x - \alpha_1 \leq -\lambda$
 $(x - \alpha_1)^2 \geq \lambda^2$

2^o integraal : $x \geq \alpha_1 + \lambda$
 $x - \alpha_1 \geq \lambda$
 $(x - \alpha_1)^2 \geq \lambda^2$

Vraag 4

Chi-kwadraat wanneer gebruiken en 1 of 2 zijdig

➔ Afhankelijk van de alternatieve hypothese maken we een keuze tussen een eenzijdige of tweezijdige toets