Formuleblad Statistiek (2024-2025)

Statistiek deel 1

Steekproefgemiddelde (gegeven een steekproef met n uitkomsten $x_1, x_2, \ldots, x_n$)

$$\overline{x} = \frac{\sum_i x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Steekproefvariantie en steekproefstandaardafwijking:

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}$$
$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$

Rekenregels kansrekening:

$$P(A \text{ of } B) = P(A) + P(B) - P(A \text{ en } B) \qquad \text{(optelregel)}$$
$$P(B) = 1 - P(\text{niet } B) \qquad \text{(complementregel)}$$
$$P(A \mid B) = \frac{P(A \text{ en } B)}{P(B)} \qquad \text{(conditionele kansen)}$$

Discrete en continue kansverdelingen:

	Discrete kansvariabelen	**Continue kansvariabelen**
Uitkomstenruimte:	Eindig / aftelbaar oneindig	Overaftelbaar oneindig
Toepassingen:	Tellen / categoriseren	Meten
Kansbegrip:	Kansfunctie $p(k) = P(X = k)$	Kansdichtheidsfunctie $f(x)$
CDF:	$F(k) = P(X \leq k) = \sum_{\ell:\ell \leq k} p(\ell)$	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy$
Verwachtingswaarde:	$E[X] = \sum_k k \cdot P(X = k)$	$E[X] = \int x \cdot f(x)\, dx$
Variantie:	$\text{Var}(X) = \sum_k (k - E[X])^2 \cdot P(X = k)$	$\text{Var}(X) = \int (x - E[X])^2 \cdot f(x)\, dx$
Standaardafwijking:	$\sigma(X) = \sqrt{\text{Var}(X)}$	$\sigma(X) = \sqrt{\text{Var}(X)}$

Speciale kansverdelingen:

- $X \sim$ Binomiaal(n, p): tellen van aantal successen bij onafhankelijke kansexperimenten met twee uitkomsten (Bernoulli-experimenten): succes / mislukking.

  Parameters: het aantal Bernoulli-experimenten n en de succeskans per experiment p.

- $X \sim$ Poisson$(\lambda \cdot t)$: tellen van aantal "gebeurtenissen" in een "interval" van tijd / ruimte.

  Parameters: het gemiddelde aantal gebeurtenissen λ per meeteenheid (tijd / ruimte) en het aantal meeteenheden t.

  → *Voorbeeld: bij de meeteenheid van een dag bestaat een week uit $t = 7$ meeteenheden.*

- $T \sim$ Exponentieel(λ): meten van de tijd / ruimte tot de volgende gebeurtenis.

  Parameter: het gemiddelde aantal gebeurtenissen λ per meeteenheid (tijd / ruimte).

Verwachtingswaarde en variantie van veelgebruikte kansverdelingen:

Verdeling	Kans(dichtheids)functie	CDF	$E(X)$	$\text{Var}(X)$
		Discreet		
Uniform(a,b)	$p(k)=\frac{1}{b-a+1}$ $(k=a,a+1,\ldots,b)$	$F(k)=\begin{cases}0 & x<a\\ \frac{k-a+1}{b-a+1} & a\le k<b\\ 1 & k\ge b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomiaal(n,p)	$p(k)=\binom{n}{k}p^k(1-p)^{n-k}$	$F(k)=\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}$	np	$np(1-p)$
Poisson(λ)	$p(k)=e^{-\lambda}\cdot\frac{\lambda^k}{k!}$	$F(k)=\sum_{i=0}^{k}e^{-\lambda}\cdot\frac{\lambda^i}{i!}$	λ	λ
		Continuous		
Uniform(a,b)	$f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & \text{elders.}\end{cases}$	$F(x)=\begin{cases}0, & x<a\\ \frac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentieel(λ)	$f(x)=\lambda e^{-\lambda x}$, $x\ge 0$	$F(x)=\begin{cases}1-e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Veelgebruikte functies op de grafische rekenmachine

Type vraag	TI-84 Plus	Casio
	Continue kansverdeling (willekeurig)	
$P(a\le X\le b)$	$\int_a^b f(x)\,dx$	$\int_a^b f(x)\,dx$
	$X\sim$ **Binomiaal**(n,p)	
$P(X=k)$	binompdf(n,p,k)	BinomialPD(k,n,p)
$P(X\le k)$	binomcdf(n,p,k)	BinomialCD(k,n,p)
	$X\sim N(\mu,\sigma)$	
$P(a\le X\le b)$	normalcdf(a,b,μ,σ)	NormalCD(a,b,σ,μ)
Grenswaarde g zodat $P(X\le g)=p$?	invNorm(p,μ,σ)	InvNormCD(tail=left$,p,\sigma,\mu)$
	$X\sim$ **Poisson**(λ)	
$P(X=k)$	poissonpdf(λ,k)	PoissonPD(k,λ)
$P(X\le k)$	poissoncdf(λ,k)	PoissonCD(k,λ)

z-score:

$$z=\frac{x-\mu}{\sigma}$$

Centrale limietstelling: Gegeven n kansvariabelen $X_1,X_2,\ldots,X_n$ die onderling onafhankelijk zijn en dezelfde kansverdeling hebben met een verwachtingswaarde μ en standaardafwijking σ, dan geldt (bij benadering) dat

- de som $\sum X = X_1+X_2+\ldots+X_n$ normaal verdeeld is met verwachtingswaarde $n\cdot\mu$ en standaardafwijking $\sqrt{n}\cdot\sigma$.
- het gemiddelde $\overline{X}=\frac{X_1+X_2+\ldots+X_n}{n}$ normaal verdeeld is met verwachtingswaarde μ en standaardafwijking $\frac{\sigma}{\sqrt{n}}$.

Statistiek deel 2:

Betrouwbaarheidsintervallen voor het gemiddelde μ (σ bekend)

- **$100 \cdot (1-\alpha)\%$-betrouwbaarheidsinterval (BI) voor μ:**

$$z_{\alpha/2} = \text{InvNorm}(\text{opp} = 1 - \alpha/2; \mu = 0; \sigma = 1)$$

$$[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$$

- **Minimale steekproefomvang voor $100 \cdot (1-\alpha)\%$-BI als μ maximaal $\pm a$ mag afwijken:**

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{a}\right)^2$$

Betrouwbaarheidsintervallen voor het gemiddelde μ (σ onbekend)

- **$100 \cdot (1-\alpha)\%$-betrouwbaarheidsinterval (BI) voor μ:**

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 1)$$

$$[\overline{x} - t \cdot \frac{s}{\sqrt{n}}; \overline{x} + t \cdot \frac{s}{\sqrt{n}}]$$

- **Minimale steekproefomvang voor $100 \cdot (1-\alpha)\%$-BI als μ maximaal $\pm a$ mag afwijken:**

$$\text{GR tabel (voor verschillende } n\text{):} \quad \frac{s}{\sqrt{n}} \cdot \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 1) \leq a$$

NB: zodra $n \geq 30$, vallen de normale en de t-verdeling nagenoeg samen. Je mag dan rekenen met de schatting s in plaats van de daadwerkelijke (onbekende) σ.

- **Onderscheidend vermogen (toets met $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, en gegeven $\mu = \mu_1$)**

$$1 - \beta = P(\overline{X} \text{ neemt waarde aan in het kritieke gebied} \mid \mu = \mu_1)$$

Betrouwbaarheidsintervallen voor de binomiale succeskans p

Betrouwbaarheidsinterval voor p (Clopper-Pearson): Gegeven een binomiale verdeling met n Bernoulli-experimenten en onbekende p, en uitkomst k.

1. Bereken de succeskans p_1 zodat geldt $P(X \leq k) = \text{binomcdf}(n; p; k) = \alpha/2$
2. Bereken de succeskans p_2 zodat geldt $P(X \geq k) = 1 - \text{binomcdf}(n; p; k-1) = \alpha/2$
3. De berekende waarden voor p_1 en p_2 zijn de grenzen van het Clopper-Pearson interval.

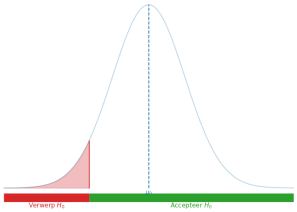
Hypothesetoetsen

Stappenplan hypothesetoetsen

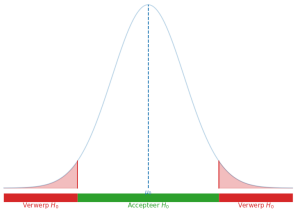
1. Definieer de nulhypothese H_0 en de alternatieve hypothese H_1.
2. Bepaal het significantieniveau α (kans op verwerpen van H_0 terwijl H_0 waar is $\rightarrow$ type-I fout)
3. Verzamel data voor de toetsingsgrootheid
4. Bereken de toetsingsgrootheid
    - Uitgaande van de nulhypothese H_0 maken we aannames over de kansverdeling van de toetsingsgrootheid!
5. Geef een conclusie (met behulp van het kritieke gebied / p-waarde) en vertaal deze terug naar de originele probleemcontext.

Drie typen hypothesetoetsen: linkszijdig, tweezijdig, rechtszijdig

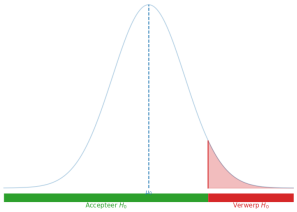
Linkszijdige toets
Kritiek gebied:
$(-\infty; g]$



Tweezijdige toets
Kritiek gebied:
$(-\infty; g_1]$ en $[g_2; \infty)$



Rechtszijdige toets
Kritiek gebied:
$[g; \infty)$



Kansverdeling (onder H_0)	**Linkszijdig**	**Tweezijdig**	**Rechtszijdig**
$N(\mu;\sigma)$	$g = \text{InvNorm}(\alpha;\mu;\sigma)$	$g_1 = \text{InvNorm}(\text{opp} = \frac{\alpha}{2};\mu;\sigma)$ $g_2 = \text{InvNorm}(\text{opp} = 1-\frac{\alpha}{2};\mu;\sigma)$	$g = \text{InvNorm}(1-\alpha;\mu;\sigma)$
$t(\text{df})$	$g = \text{InvT}(\alpha;\text{df})$	$g_1 = \text{InvT}(\text{opp} = \frac{\alpha}{2};\text{df})$ $g_2 = \text{InvT}(\text{opp} = 1-\frac{\alpha}{2};\text{df})$	$g = \text{InvT}(1-\alpha;\text{df})$
Grenzen die met de solver functie moeten worden opgelost:			
$\chi^2(\text{df})$ (chikwadraat)	$\chi^2\text{cdf}(0;g;\text{df}) = \alpha$	$\chi^2\text{cdf}(0;g_1;\text{df}) = \frac{\alpha}{2}$ $\chi^2\text{cdf}(g_2;10^{99};\text{df}) = \frac{\alpha}{2}$	$\chi^2\text{cdf}(g;10^{99};\text{df}) = \alpha$
$F(\text{df}_A;\text{df}_B)$	$\text{Fcdf}(0;g;\text{df}_A;\text{df}_B) = \alpha$	$\text{Fcdf}(0;g_1;\text{df}_A;\text{df}_B) = \frac{\alpha}{2}$ $\text{Fcdf}(g_2;10^{99};\text{df}_A;\text{df}_B) = \frac{\alpha}{2}$	$\text{Fcdf}(g;10^{99};\text{df}_A;\text{df}_B) = \alpha$

p-waardes uitrekenen (gegeven een theoretische en geobserveerde toetsingsgrootheid T en t)

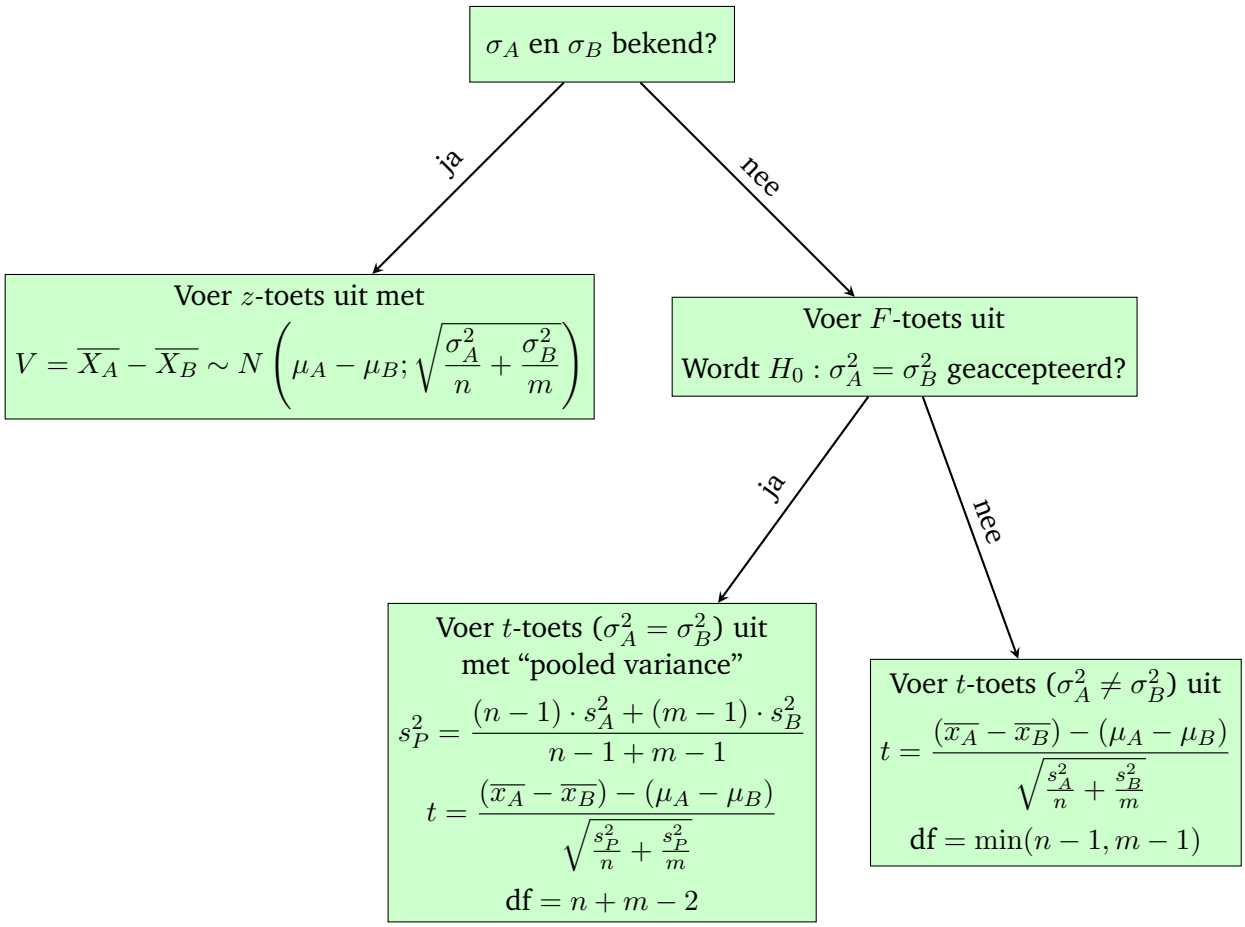
Kansverdeling (onder H_0)	**Linkszijdig ($P(T \leq t)$)**	**Rechtszijdig ($P(T \geq t)$)**
$N(\mu;\sigma)$	$p = \text{normalcdf}(-10^{99};t;\mu;\sigma)$	$p = \text{normalcdf}(t;10^{99};\mu;\sigma)$
$t(\text{df})$	$p = \text{tcdf}(-10^{99};t;\text{df})$	$p = \text{tcdf}(t;10^{99};\text{df})$
$\chi^2(\text{df})$	$p = \chi^2\text{cdf}(0;t;\text{df})$	$p = \chi^2\text{cdf}(t;10^{99};\text{df})$
$F(\text{df}_A;\text{df}_B)$	$p = \text{Fcdf}(0;t;\text{df}_A;\text{df}_B)$	$p = \text{Fcdf}(t;10^{99};\text{df}_A;\text{df}_B)$

NB: Om met de p-waarde een conclusie te trekken uit een hypothesetoets vergelijken we de p-waarde met het significantieniveau α. Let op: bij tweezijdige toetsen neem je het minimum van de linkszijdige en rechtszijdige p-waarde en vergelijk je deze met $\alpha/2$!

Soorten toetsen

Soort toets	Toetsingsgrootheid	Kansverdeling (onder H_0)
Toetsen voor het gemiddelde $\mu \leq \mu_0$ of $\mu = \mu_0$ of $\mu \geq \mu_0$		
z-toets (σ bekend)	$\overline{X}$	$N(\mu_0; \frac{\sigma}{\sqrt{n}})$
t-toets (σ onbekend)	$T = \frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$	$t(\text{df} = n-1)$
Chikwadraattoetsen (χ^2)		
Onafhankelijkheid	$X^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$\chi^2(\text{df} = (\#\text{rijen-1}) \cdot (\#\text{kolommen-1}))$
Aanpassing (goodness-of-fit)	$X^2 = \sum_i \frac{(O_i-E_i)^2}{E_i}$	$\chi^2(\text{df} = (\#\text{categorieën-1}))$
Verschiltoetsen (op basis van twee populaties A en B)		
F-toets: $\sigma_A^2 = \sigma_B^2$	$F = \frac{S_A^2}{S_B^2}$	$F(\text{df}_A, \text{df}_B)$
z-toets	$V = \overline{X_A} - \overline{X_B}$	$N\left(\mu_A - \mu_B; \sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}\right)$
t-toets ($\sigma_A^2 = \sigma_B^2$)	$T = \frac{(\overline{X_A}-\overline{X_B})-(\mu_A-\mu_B)}{\sqrt{\frac{S_P^2}{n}+\frac{S_P^2}{m}}}$	$t(\text{df} = n+m-2)$
t-toets ($\sigma_A^2 \neq \sigma_B^2$)	$T = \frac{(\overline{X_A}-\overline{X_B})-(\mu_A-\mu_B)}{\sqrt{\frac{S_A^2}{n}+\frac{S_B^2}{m}}}$	$t(\text{df} = \min(n-1; m-1))$

Beslisboom verschiltoetsen



Correlatie en regressie

Correlatiecoëfficiënt van Pearson:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}$$

Correlatiecoëfficiënt van Spearman:

$$r_s = 1 - \frac{6 \cdot \sum_i d_i^2}{n^3 - n}$$

Coëfficiënten van de lineaire regressielijn $Y = a + b \cdot X$:

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$
$$a = \overline{y} - b \cdot \overline{x}$$

Schatting van de variantie van de storingsterm ε:

$$s_\varepsilon^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum \left(y_i - (a + b \cdot x_i)\right)^2}{n-2} = \frac{n}{n-2} \cdot \left(\overline{y^2} - a \cdot \overline{y} - b \cdot \overline{xy}\right)$$

$100 \cdot (1-\alpha)\%$-betrouwbaarheidsinterval voor de gemiddelde Y bij een gegeven $X = x_0$:

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2)$$
$$s_\mu = s_\varepsilon \cdot \sqrt{\frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$
$$[a + b \cdot x_0 - t \cdot s_\mu; a + b \cdot x_0 + t \cdot s_\mu]$$

$100 \cdot (1-\alpha)\%$-betrouwbaarheidsinterval voor Y bij een gegeven $X = x_0$:

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2)$$
$$s_f = s_\varepsilon \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$
$$[a + b \cdot x_0 - t \cdot s_f; a + b \cdot x_0 + t \cdot s_f]$$