Samenvatting

Inleiding Statistiek - Collegejaar 2015-2016

Overzicht van colleges

College 1

Korte herhaling kansrekening, survey sampling (aselect, met en zonder teruglegging); H7, par 1, 2, 3.1 en 3.2

Wat basis kansrekening wordt hier uitgelegd. (Verwachting, variantie ….)

Survey Sampling:
Populatie van grootte $N, x_1, \dots, x_N$, populatie parameters:
Populatiegemiddelde: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$.
Populatievariantie: $\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Steekproef nemen, dus n trekkingen (**met** teruglegging): $X_1, \dots, X_n$
Steekproefgemiddelde: $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
$E(\overline{X_n}) = \mu, Var(\overline{X_n}) = \frac{\sigma^2}{n}$

Steekproef nemen, dus n trekkingen (**zonder** teruglegging): $X_1, \dots, X_n$
Steekproefgemiddelde: $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
$E(\overline{X_n}) = \mu, Var(\overline{X_n}) = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$

$\overline{X_n}$ is een zuivere schatter voor μ, want $E(\overline{X_n}) = \mu$.

College 2

Gestratificeerd steekproeftrekken, Neyman allocatie; H7, par 5
Zuivere schatter voor σ^2 van College 1:
Met teruglegging: $\frac{n}{n-1}S_n^2$
Zonder teruglegging: $\frac{n}{n-1}\left(1 - \frac{1}{N}\right)S_n^2$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

Betrouwbaarheidsinterval:
Wegens de centrale limietstelling:
$-1.96 \leq \sqrt{n}\frac{\overline{X_n} - \mu}{S_n} \leq 1.96$ met kans ~0.95

$$\overline{X_n} - \frac{1.96}{\sqrt{n}}S_n \leq \mu \leq \overline{X_n} + \frac{1.96}{\sqrt{n}}S_n$$

Neyman allocatie (hierover is ook een extra bestand lectures\slides op blackboard). Ik kan me alleen niet herinneren dat dit belangrijk is, dus hoeven we denk ik niet te leren.

College 3

Parameterschatten, het probleem. Momentenmethode. Monte Carlo en Bootstrap voorbeeld; H8, par 1, 2, 3 en 4

Momentenmethode: methode om een parameter te schatten.
Bepaalde verdeling. Bereken het eerste moment (de verwachting) en het steekproefgemiddelde. De schatter is gelijk aan de schatter, waarvoor het eerste moment en het steekproefgemiddelde gelijk zijn aan elkaar.

$$E(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Stel er zijn twee onbekende parameters (bijvoorbeeld de Gamma verdeling (α, λ)), gebruik dan ook het tweede moment.

$$E_{\alpha,\lambda}(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$E_{\alpha,\lambda}(X^2) = \frac{1}{N}\sum_{i=1}^{N} X_i^2$$

Dit kan dus ook voor een k aan parameters.

Bootstrap (Monte Carlo): Begin met een waarde voor de parameter die je wil schatten. Neem nu een steekproef op basis van deze parameter (op de slides neemt hij $n = 227$, misschien is dit de standaard grootte van de steekproef?). Schat nu de parameter op basis van deze steekproef. Voor met deze nieuwe schatting nog een keer de steekproef uit en schat de parameter weer opnieuw….. (Op de slides heeft hij het over dit 10000 keer doen). Het gemiddelde van al deze schattingen is dan je uiteindelijke schatting.

College 4

Maximum Likelihood schatten, algemeen idee en voorbeelden; H8, par 5, 5.1

Maximum Likelihood methode: Belangrijk om te zeggen dat het i.i.d. is. Stel we willen een θ schatten. Dan $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$ is de likelihood en $l(\theta) = \sum_{i=1}^{n} \log(f(X_1|\theta))$ is de log likelihood, log is het natuurlijk logaritme. Deze functie nu maximaliseren over θ, dus $L(\theta)$ of $l(\theta)$ differentiëren en gelijk stellen aan nul en dan checken of het een maximum is. Hiervoor kan er nog een keer gedifferentieerd worden en dan kijken of dat negatief is. We gebruiken meeste log likelihood.

Schatten van parameters:
Schatting: data afhankelijk element van de parameterruimte
Schatter: stochastisch element van de parameterruimte

College 5

Asymptotiek van de ML schatter, Betrouwbaarheidsgebieden, heel kort Bayesiaans schatten; H8, par 5.2, 5.3,

Fischer Informatie voor θ in één waarneming uit verdeling:
$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log\big(f(X|\theta)\big)\right]$ (De laatste gelijkheid alleen als f glad is.)

Assymptotische verdeling: $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0)\sim N(0,1) \leftrightarrow \sqrt{n}(\hat{\theta}-\theta_0)\sim N\left(0,\frac{1}{I(\theta_0)}\right)$
Dus assymptotisch normaal verdeeld met verwachting θ_0 en variantie $\frac{1}{nI(\theta_0)}$.

Momentenschatter en ML schatter zijn schatters met een zekere nauwkeurigheid; die nauwkeurigheid wordt niet in de schattingen zelf gerepresenteerd.
Betrouwbaarheidsgebieden geven niet alleen een schatting van de parameter, maar meteen een hele verzameling van plausibele parameters (een nauwkeurige schatter zal tot een kleinere verzameling leiden, een onnauwkeurige schatter tot een groter interval)

College 6
Bayesiaans schatten en relatieve efficientie; H8, par 6, 6.1

Bij Bayes: de waarde van de parameter wordt gemodelleerd als stochastisch, met a priori verdeling op de parameterruimte; de data en het model worden gebruikt om de a priori verdeling om te zetten in een a posteriori verdeling.

$$f_\Theta(\theta)$$
$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_\Theta(\theta)$$
$$f_X(x) = \int f_{X,\Theta}(x,\theta)d\theta = \int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta$$
$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta}$$

De eerste heet de a priori verdeling. De tweede heet de gezamelijke verdeling van X en Θ. De derde heet de marginale verdeling van X. De vierde is de posterior verdeling.

Posterior mean: $E(\Theta) = \int_\infty^\infty \theta f_{\Theta|X}(\theta|x)d\theta$.
Posterior mode: Maximum Likelihood schatter van θ. Vinden met $f_{\Theta|X}(\theta|x)$.

Geconjugeerde prior klassen, als de a priori verdeling gelijk is aan de posterior verdeling.

Mean squared error (MSE) $MSE(\hat{\theta}) = E_\theta(\hat{\theta}-\theta)^2 = Var_\theta(\hat{\theta}) + (E_\theta(\hat{\theta})-\theta)^2$
$(E_\theta(\hat{\theta})-\theta)^2$ is $(onzuiverheid)^2$ en is gelijk aan 0 voor zuivere schatters.
Relatieve efficiëntie van schatter $\hat{\theta}$ en $\tilde{\theta}$: $eff(\hat{\theta},\tilde{\theta}) = \frac{MSE(\hat{\theta})}{MSE(\tilde{\theta})}$
Als beide schatters zuiver, dan: $eff(\hat{\theta},\tilde{\theta}) = \frac{Var(\hat{\theta})}{Var(\tilde{\theta})}$

College 7
Cramér-Rao, voldoendheid, exponentiele familie, Rao Blackwell; H8, par 7, 8

Stelling Cramér-Rao ongelijkheid:
$X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} f(\cdot, \theta)$. Zij $T = \tau(X_1, \dots, X_n)$ een zuivere schatter voor θ. Er geldt onder gladheidrestricties op f, dat $Var(T) \geq \frac{1}{nI(\theta)}$.

Een (mogelijk vectorwaardige) grootheid $T = \tau(X_1, \dots, X_n)$ is **voldoende** voor θ als voor iedere t, de voorwaardelijke verdeling van $X_1, \dots, X_n$ gegeven $T = t$ niet afhangt van θ.

Factoriseringscriterium van Neyman:
$T = \tau(X_1, \dots, X_n)$ is voldoende voor $\theta \leftrightarrow$ de kansdichtheid van $X_1, \dots, X_n$ geschreven kan worden als: $f(x_1, \dots, x_n|\theta) = g(T; \theta)h(x_1, \dots, x_n)$.

Exponentiële families:
Er zijn functies c, T, d, S en een verzameling A zodat de kansdichtheid te schrijven is als:

$$f(x|\theta) = \begin{cases} e^{c(\theta)T(x)+d(\theta)+S(x)}, x \in A \\ 0, x \notin A \end{cases}$$

Stelling Rao-Blackwell:
Stel $\hat{\theta}$ is een zuivere schatter voor θ en T is voldoende voor θ. Dan is $\tilde{\theta} = E_\theta(\hat{\theta}|T)$ ook een zuivere schatter voor θ. Bovendien geldt:

$$E_\theta(\hat{\theta} - \theta)^2 = Var_\theta(\hat{\theta}) \geq Var\left(E_\theta(\hat{\theta}|T)\right) \geq Var_\theta(\tilde{\theta}) = E_\theta(\tilde{\theta} - \theta)^2$$

Voor alle θ en $E_\theta(\hat{\theta} - \theta)^2 = E_\theta(\tilde{\theta} - \theta)^2 \leftrightarrow \tilde{\theta} = \hat{\theta}$.

College 8

Toetsen van hypothesen, significantieniveau, foutentabel, onderscheidend vermogen, Neyman Pearson lemma, relatie toets-betrouwbaarheidsinterval; H9, par 2, 2.2, 2.3 en 3

Type 1 fout: H_0 verwerpen als H_0 waar is.
Type 2 fout: H_0 niet verwerpen als H_0 niet waar is.

Situatie: $X \sim f$ Kans(dichtheids)functie.
Enkelvoudige nulhypothese: $H_0: f = f_0$
Enkelvoudige alternatieve: $H_1: f = f_1$
Likelihood ratio toets: verwerp H_0 ten gunste van H_1 als:

$$\frac{f_0(X)}{f_1(X)} < c$$

Significantieniveau van deze toets is:

$$\alpha = P_0\left(\frac{f_0(X)}{f_1(X)} < c\right)$$

Powerfunctie: $\beta(\theta) = 1 - P(H_0\ verwerpen|\theta)$
Onderscheidend vermogen: $H_1: \theta = \theta_1, \beta(\theta_1) = 1 - P(H_0\ verwerpen|\theta_1)$
Stelling Neyman-Pearson lemma:
Onder alle toetsen voor deze hypothesen met significantieniveau kleiner dan of gelijk aan α, heeft de beschreven likelihood ratio toets het hoogste onderscheidingsvermogen.

Wat hieronder staat, staat in de slides, maar is volgens mij niet heel erg belangrijk.
Acceptatiegebied: $A_\alpha(\theta_0) = \{x|H_0: \theta = \theta_0 \; wordt \; o.b.v. x \; niet \; verworpen \; op \; niveau \; \alpha\}$
Betrouwbaarheidsgebied: $C_\alpha(x) \subset \Theta$: $\theta \in C_\alpha(x) \leftrightarrow x \in A_\alpha(\theta)$

$$P(\theta \in C_\alpha(X)|\theta) = P(X \in A_\alpha(\theta)|\theta) = 1 - \alpha$$

Dus $C_\alpha(x)$ geeft een $(1-\alpha)\cdot 100\%$ betrouwbaarheidsinterval voor θ.

College 9

p-waarde, Gegeneraliseerde Likelihood Ratio toets; H9, par 2.1 en 4

Voor welke waarden van α wordt H_0 verworpen? De p-waarde:
$p < \alpha \rightarrow$ verwerpen
$p > \alpha \rightarrow$ niet verwerpen

Samengestelde hypothese als het een verzameling van waarden is:
$X \sim f$ Kans(dichtheids)functie, $f = \{f(\cdot|\theta): \theta \in \Omega\}$
Zij $\omega_0 \subset \Omega, \omega_1 \subset \Omega \backslash \omega_0$ en $H_0: \theta \in \omega_0$ en $H_1: \theta \in \omega_1$. Als allebei de verzameling maar uit één punt bestaan, dan enkelvoudig en is dus de likelihood ratio toets meest onderscheidend (Neyman-Pearson).

Gegeneraliseerde likelihood ratio toetsingsgrootheid:

$$\Lambda^* = \frac{\sup\limits_{\theta \in \omega_0} f(X|\theta)}{\sup\limits_{\theta \in \omega_1} f(X|\theta)} \; of \; \Lambda = \frac{\sup\limits_{\theta \in \omega_0} f(X|\theta)}{\sup\limits_{\theta \in \omega_0 \cup \omega_1} f(X|\theta)}$$

Asymptotisch resultaat (i.i.d. data): Als de nulhypothese geldt ($\theta \in \omega_0$), dan:

$$\forall t > 0, P(-2\log(\Lambda) \leq t|H_0) \overset{n \rightarrow \infty}{\rightarrow} P(V \leq t)$$
$$V \sim \chi^2_d, d = Dim(\omega_0 \cup \omega_1) - Dim(\omega_0)$$

Voor uniform meest onderscheidend mag de alternatieve hypothese niet tweezijdig zijn ($H_1: \theta \neq \theta_0$), dus moet éénzijdig zijn ($H_1: \theta < \theta_0$ en $H_1: \theta > \theta_0$). Verder moet gelden dat het onderscheidend vermogen niet van de alternatieve hypothese afhangt. M.a.w. $\forall \theta \in \omega_1$ is het onderscheidend vermogen gelijk.

De Gegeneraliseerde likelihood ratio toetsingsgrootheid is i.h.a. niet uniform meest onderscheidend.

Berekening Λ:

1. Bepaal Maximum Likelihood schatter voor de noemer.
2. Als deze schatter in ω_0 ligt, dan is $\Lambda = 1$
3. Zo niet, maximaliseer de likelihood dan ook over ω_0.

College 10

GLR toetsen (vb multionomiaal) en Bayesiaans toetsen; H9, par 5, 6 en 1
Situatie: $X \sim f$ kans(dichtheids)functie

Enkelvoudige nulhypothese: $H_0: f = f_0$
Enkelvoudige alternatieve hypothese: $H_1: f = f_1$
A priori kansen op H_0 en H_1: $P(H_0)$ en $P(H_1)$.
Geobserveerde data: x.

$$\frac{P(H_0|X=x)}{P(H_1|X=x)} = \frac{P(X=x|H_0)P(H_0)}{P(X=x|H_1)P(H_1)} = \frac{f_0(x)}{f_1(x)} \cdot \frac{P(H_0)}{P(H_1)}$$

Verwerpen als dit kleiner is dan 1.

Gegeven dataset: $\{x_1, \dots, x_n\}$.
i.i.d. $X_1, \dots, X_n$ met $P(X_i \leq x) = F(x)$
Empirische verdelingsfunctie: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,\infty)}(x) = \frac{\#\{i:X_i \leq x\}}{n}$
$E\big(F_n(x)\big) = F(x), Var\big(F_n(x)\big) = \frac{F(x)(1-F(x))}{n}$
Consistentie: $\forall \varepsilon > 0, P(|F_n(x) - F(x)| > \varepsilon) \to 0 \; als \; n \to \infty$.
Asymptotische verdeling: $\forall x: 0 < F(x) < 1, als \; n \to \infty \; P\left(\sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)(1-F(x)}} \leq t\right) \to \Phi(t), t \in R$.

Er bestaat iets als de Hazard functie (heeft met survival gedoe te maken), dit is dus in de slides van College 10 te vinden. Het is volgens mij geen tentamenstof, dus ik ga het niet helemaal overschrijven, maar het is misschien wel slim om het een keer door te lezen.

College 11

Samenvatten van data (grafisch en numeriek); H10 geheel en goodness of fit; H9 par 9 (uitgebreid met Kolmogorov Smirnov)

Histogram

Kernschatter ('continue versie van histogram'): kernfunctie w, gladde dichtheid symmetrisch om 0, die geen relatie heeft met de dichtheid f. Bandbreedte $h > 0$. Herschaalde kernfunctie:

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i)$$

Gegeven data set: $\{x_1, \dots, x_n\}$
Kentallen voor locatie:

- Gemiddelde: $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Mediaan: $\begin{cases} x_{\frac{n}{2}+\frac{1}{2}}, n \; oneven \\ \frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}, n \; even \end{cases}$ Robuust (extreme waarde erbij veranderd de mediaan niet zeer)
- Afgeknot gemiddelde: $\frac{\sum_{i=k+1}^{n-k} x_i}{n-2k}$
- M-schatter: minimaliseert $\sum_{i=1}^{n} \Psi(x_i - \mu)$, Ψ symmetrische, convexe functie.

Kentallen voor spreiding:

- Steekproefstandaardafwijking: $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- IQR: $x_{\frac{3n}{4}} - x_{\frac{n}{4}}$ Robuust (extreme waarde erbij veranderd de IQR niet zeer)
- MAD: mediaan van de absolute afwijkingen t.o.v. de mediaan:

$$Med(\{|x_i - Med(x_1, \dots, x_n)| : 1 \leq i \leq n\})$$

Boxplot

Gegeven data set: $\{(x_1, y_1), \dots, (x_n, y_n)\}$
Scatterplots

Kental voor samenhang:
Steekproefcorrelatie$\in [-1,1]$: $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$
Rangcorrelatie: Geen idee hoe je dit zou moeten berekenen, maar het komt waarschijnlijk ook niet op het tentamen.

Goodness of fit (grafisch): we hebben een steekproef uit een onbekende verdeling en willen weten of een bepaald model op de data past:
Zij F de verdelingsfunctie van het model (continu en strikt stijgend), dan is $U_1, \dots, U_n$ met $U_i = F(X_i)$ een steekproef uit de $U[0,1]$ verdeling (uniforme verdeling op het interval [0,1]).
$E(U_i) = \frac{i}{n+1}$.
Probability plot: $\left\{\left(\frac{i}{n+1}, F(x_i)\right) : 1 \leq i \leq n\right\}$
QQ-plot: $\left\{\left(F^{-1}\left(\frac{i}{n+1}\right), x_i\right) : 1 \leq i \leq n\right\}$
Als deze plots rond $y = x$ liggen dan zal het model passen op de data.

Goodness of fit (via toets):
Model: $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$
Stel we denken dat het normaal verdeeld is, dan $H_0: F \in \{N(\mu, \sigma^2) : \mu \in R, \sigma^2 > 0\}, F_0 = N(\mu, \sigma^2)$.
Kolmogorev-Smirnov toets: $T = \sup_{x \in R} |F_n(x) - \mathrm{F}_0(x)| = \sup_{x \in R} |F_n(x) - \Phi\left(\frac{x-\overline{X_n}}{S_n}\right)|$,
waarbij $F_n(x)$ de empirische verdelingsfunctie. En hier dus voor het normaal verdeeld ingevuld.
We verwerpen voor grote waarden van T.

Dan nu uitrekenen $P(T > \sup_{x \in R}|F_n(x) - \mathrm{F}_0(x)| \ |H_0)$ geeft een p-waarde.

College 12

Twee steekproeven probleem; onafhankelijke steekproeven H11, par 1, 2, 2.1, 2.2 en 2.3

Twee-steekproeven probleem (onafh.) algemeen:

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} F, Y_1, \dots, Y_n \overset{i.i.d.}{\sim} G$$

Hypothese binnen model, $F = G$
Bijvoorbeeld $F = N(\mu_X, \sigma^2)$ en $G = (\mu_Y, \sigma^2)$, dan $H_0: \mu_X = \mu_Y$ en $H_1: \mu_X \neq \mu_Y$.

Dan hoop gereken op de slides… en dan komt er wat over de t-test. (Student) t-verdeeld met n vrijheidsgraden.

Toetsen zonder aanname van normaliteit:
Wilcoxon-Mann-Whitney toets
Data: $x_1, \dots, x_n$ en $y_1, \dots, y_m$

- Pool de data, dus zet de x-en en y-en in goede volgorde.
- Bepaal de rangen van de 1e data set en sommeer die; $r = \sum_{i=1}^{n} s_i$

We verwerpen voor grote/kleine waarde van r.

Er kan met veel steekproeven en de empirische verdelingsfunctie een kans worden bepaald en dus of er moet worden verworpen.

College 13

Twee steekproevenrprobleem (gepaard) H11, par 3, 3.1, 3.2 en 3.3
en ANOVA; H12, par 1, 2 en 2.1

Twee-steekproeven probleem (gepaard) (gepaard als de steekproeven niet onafh. zijn)
Data: $(x_1, y_1), \dots, (x_n, y_n)$ x_i en y_i zijn niet onafh. maar alle paren wel. Definieer $D_i = X_i - Y_i$
Normaal model: $D_1, \dots, D_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. $H_0: \mu = 0$, $H_1: \mu \neq 0$.

Uiteindelijk t-toets hiermee uitvoeren.

Toetsen zonder aanname van normaliteit:
Model: $(X_1, Y_1), \dots, (X_n, Y_n)$ onafhankelijk
$D_i = X_i - Y_i$, $D_1, \dots, D_n \overset{i.i.d.}{\sim} g$, dichtheid (als er geen systematisch verschil is tussen X en Y, dan is g symmetrisch rond 0).
De tekentoets: Aantal positieve waarden van D.

Wilcoxon signed-rang toetsingsgrootheid:

1. Bepaal de absolute waarden van D en orden deze
2. Bereken de rangen van deze waarden
3. Geef de oorspronkelijke tekens toe (positief/negatief)
4. Sommeer de rangen van de oorspronkelijke positieve waarden.

Onder H_0 zal de som ongeveer gelijk zijn aan $\frac{n^2}{4}$.

1-factor variantieanalyse:
Variantiemodel:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2), \sum \alpha_i = 0$$

ANOVA tabel te vinden op pagina 483.

College 14

Regressie en correlatie, enkelvoudige lineaire regressie; H14, par 1, 2, 2.1, 2.2, 2.3, 3

Regressie analyse:
Het lineaire regressie model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, \dots, n$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \Leftrightarrow Y = X\beta + e$$

Kleinste kwadratenschatter: $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Correlatie:

$$s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2, s_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_n})^2, s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(y_i - \overline{y_n})$$

Correlatiecoëfficient: $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$.

College 15

De Fisher Exact toets, H13, par 2 en terugblik op de stof.
Geen nieuwe stof, maar wel overzicht van wat we moeten kunnen:

• Rekenen met verwachtingen en varianties (iha en in kader van survey sampling)
• Eigenschappen van normale verdelingen kennen
• Bepalen van momenten- maximum likelihood en Bayesiaanse schatters in specifieke modellen
• Zuiverheid van schatter bepalen
• Variabiliteit in schatter
• Fisher informatie uitrekenen in specifiek model; relatie tot asymptotische verdeling ML schatter en Cramer Rao
• Bepalen of stochastische grootheid voldoende is voor een parameter in een model; Rao Blackwell
• Kunnen bepalen of klasse van verdelingen een exponentiele familie vormt en consequenties hiervan kennen
• Werken met de Neyman-Pearson opzet van het toetsen van hypothesen; [NP-lemma, significantieniveau, fout type 1 en 2, onderscheidend vermogen van een toets (U)MP]; pwaarde
• Afleiden van (G)LR toets in specifieke modellen
• Weten waar goodness of fit toetsen voor zijn; Kolmogorov Smirnov als voorbeeld
• Interpreteren van QQ-plot, construeren van histogram, empirische verdelingsfunctie
• Tweesteekproevenprobleem: (on)gepaard; onder normaliteit en zonder die aanname
• Variantieanalysemodel kennen, interpretatie van tabel
• Lineair regressiemodel kennen, kleinste kwadratenschatter