Samenvatting

Inleiding Statistiek - Collegejaar 2012-2013

Mathematical Statistics and Data Analysis, 3-rd edition. John A. Rice

Hoofdstuk 7 paragraaf 1, 2, 3 en 5. Hoofdstuk 8 en paragraaf 1, 2 en 3 van Hoofdstuk 9.

Inleiding. Er worden veel Engelse en Nederlandse termen door elkaar gebruikt. Voor de duidelijkheid zijn de Engelse termen *schuin gedrukt*. De (eventueel relevante) Nederlandse vertaling staat er dan tussen haakjes achter.

1. Probability

§1.2 De **uitkomstenruimte** (sample space) Ω is de verzameling van alle mogelijke uitkomsten ω van een experiment.

§1.3 De **kansmaat** (probability measure) van Ω is een functie P van deelverzamelingen van Ω naar $\mathbb{R}$ die aan de volgende axioma's voldoet:

- $P(\Omega) = 1$.
- Als $A \subset \Omega$, dan $P(A) \geq 0$.
- Als A_1 enA_2 disjunct zijn, dan $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

Een logisch gevolg hiervan is dat $P(A^c) = 1 - P(A)$, $P(\emptyset) = 0$ en als $A \subset B$, dan $P(A) \leq P(B)$. Ook geldt dat $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

§1.4 Als de kans uniform verdeeld is, en er N elementen zijn, dan geldt $P(A) = n/N$. Bij n uitkomsten bij 1 experiment, en m bij het andere, zijn er totaal mn uitkomsten. Dit zet zich uiteraard voort bij meer dan 2 experimenten.

Permutaties: bij een verzameling van n elementen en een steekproef met r elementen zijn er n^r verschillend geordende mogelijkheden als we mogen hergebruiken ("met terugleggen") en $n(n-1)\cdots(n-(r+1))$ verschillenden als we niet mogen hergebruiken. n elementen kunnen op $n!$ manieren herschikt worden.

Combinaties: bij combinaties doet de volgorde er niet toe. Als we r elementen kiezen uit een verzameling van n elementen, niet hergebruiken en de volgorde er niet toe doet, hebben we $\binom{n}{r} = \dfrac{n!}{(n-r)!r!}$ mogelijkheden. Dit zijn tevens de binomiaalcoëfficiënten. Als we een verzameling van n elementen willen ordenen in r groepen met k_i de i-de groep, dan is het aantal mogelijkheden: $\binom{n}{k_1 k_2 \cdots k_r}$. Deze getallen staan bekend als de multinoniaalcoëfficiënten.

§1.5 Laat A en B twee gebeurtenissen met $P(B) \neq 0$, dan de kans dat A gebeurt, gegeven dat B gebeurt, is $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$. Dit komt overeen met $P(A \cap B) = P(A|B)P(B)$.

Wet van de totale kans: laat $B_1, \ldots, B_n$ zodanig zijn dat $\bigcup_{i=1}^n B_i = \Omega$ en $B_i \cap B_j = \emptyset$ voor $i \neq j$ en $P(B_i) > 0$ voor alle i. Dan geldt voor alle A :
$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$. Een gevolg hiervan is:

- Als A en C disjunct: $P(A \cup C|B) = P(A|B) + P(C|B)$.
- $P(A|B) = 1 - P(A^c|B)$.

Regel van Bayes: laat A en $B_1, \ldots, B_n$ gebeurtenissen met de B_i disjunct, $\bigcup_{i=1}^n B_i = \Omega$, en $P(B_i) > 0$ voor alle i. Dan $P(B_j|A) = \dfrac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$.
In een simpele situatie geldt dus: $P(X|Y) = \dfrac{P(Y|X)P(X)}{P(Y|X)P(X) + P(Y|X^c)P(X^c)}$ en $P(X|Y) = \dfrac{P(Y|X)P(X)}{P(Y)}$.

§1.6 **Onafhankelijkheid** betekent intuïtief dat $P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$. Dit herschrijven we naar onze definitie van onafhankelijkheid: A en B zijn onafhankelijk als $P(A \cap B) = P(A)P(B)$. Gebeurtenissen zijn **onderling onafhankelijk** (mutually independent) als voor een verzameling gebeurtenissen $A_1, \ldots, A_n$ voor elke deelverzameling $A_1, \ldots, A_m$ geldt dat $P(A_1 \cap \cdots \cap A_m) = P(A_1) \cdots P(A_m)$.

2. Random Variables

§2.1 We noemen X een **discrete stochast** (discrete random variable), als X of een eindig, of hoogstens aftelbaar oneindig waarden aan kan nemen. De functie p waarvoor geldt dat $p(x_i) = P(X = x_i)$ en $\sum_i p(x_i) = 1$ noemen we de **kansfunctie** (probability mass function), of **frequentiefunctie** (frequency function). Het kan ook handig zijn om de **cumulatieve verdelingsfunctie** (cumulative distribution function - cdf), te gebruiken, waarvoor geldt $F(x) = P(X \leq x), -\infty < x < \infty$. Verder geldt $\lim_{x \to -\infty} F(x) = 0$ en $\lim_{x \to \infty} F(x) = 1$. Twee discrete stochasten X en Y zijn onafhankelijk als voor alle i en j: $P(X = x_i \text{ en } Y = y_j) = P(X = x_i)P(Y = y_j)$.

§2.1.1 Een **Bernoulli-stochast** (Bernoulli random variable) neemt alleen de waarden 0 en 1 aan, met kans $1 - p$ en p, dus voor de frequentiefunctie geldt: $p(1) = p, p(0) = 1 - p$ en $p(x) = 0$ voor $x \neq 0$ en $x \neq 1$. Als A een gebeurtenis is, dan neemt de **stochastindicator** (indicator random variable) I_A de waarde 1 aan als A gebeurt, en 0 als A niet gebeurt, dus $I_A(\omega) = 1$ als $\omega \in A$ en anders $I_A(\omega) = 0$. I_A is ook een Bernoulli-stochast.

§2.1.2 Als we n experimenten uitvoeren (n vast) met kans op succes p en kans op falen $1 - p$ dan is het totaal aantal successen X **binomiaal verdeeld**. De kans dat $X = k$ is $p(k) = \binom{n}{k} p^k (1 - p)^{n-k}$. Als $X_1, \ldots, X_n$ onafhankelijke Bernoulli-stochasten zijn met $p(X_i = 1) = p$, dan $Y = X_1 + \ldots + X_n$ is een binomiale stochast.

§2.1.3 De **geometrische verdeling** (geometric distribution) is ook gebaseerd op een serie onafhankelijke Bernoulli-experimenten, maar nu oneindig veel. De kans op succes is p en X is het aantal keren tot en met de eerste keer succes (dus als $X = k$, dan is er $k - 1$ keer gefaald). Vanwege onafhankelijkheid volgt $p(k) = P(X = k) = (1-p)^{k-1}p$. Ook geldt $\sum_{k=1}^{\infty} (1-p)^{k-1}p = p \sum_{j=0}^{\infty} (1-p)^j = 1$.

De **negatieve binomiaalverdeling** (negative binomial distribution) is een generalisatie van de geometrische verdeling. Als we een serie uitvoeren met succes p totdat er r keer succes is, dan is X het totaal aantal beurten. Er geldt $P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}$. Merk op dat de negatieve binomiaalverdeling een som is van r onafhankelijke geometrische verdelingen: het aantal pogingen tot de eerste keer succes + het aantal pogingen tot de tweede keer succes + ... + het aantal pogingen tot de r-de keer succes.

§2.1.4 De **hypergeometrische verdeling** (hypergeometric distribution) is bekend van het vaasmodel. Stel dan een vaas n ballen bevat, waarvan r zwart en $n - r$ wit. Noem X het aantal zwarte ballen dat we krijgen wanneer we zonder terugleggen m ballen trekken uit de vaas, dan $P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$.

§2.1.5 De **Poisson-frequentiefunctie** (Poisson frequency function) met parameter $\lambda > 0$ is $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$. De **Poisson-verdeling** (Poisson distribution)

is $p(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$ (en dit is dus de Poisson-frequentiefunctie). De Poisson-verdeling kan gebruikt worden om de binomiaalverdeling te benaderen voor grote n en kleine p.

§2.2 Bij een **continue stochast** (continue random variable) wordt de rol van de frequenctiefunctie vervult door de **dichtheidsfunctie** (density function) $f(x)$ met $f(x) \geq 0$ en $\int_{-\infty}^{\infty} f(x)dx = 1$. Als X een stochast is met dichtheidsfunctie f, dan voor alle $a < b$ geldt dat de kans dat X in het interval (a, b) ligt gelijk is aan $P(a < X < b) = \int_a^b f(x)dx$. Een logisch gevolg hiervan is dat voor een constante c geldt dat $P(c) = 0$. Zo ook geldt $P(a < X < b) = P(a \leq X < b) = P(a < X \leq B)$ (wat niet waar is voor een discrete stochast). De **cumulatieve distributiefunctie** met $F(x) = P(X \leq x)$ is net zo gedefinieerd als bij de discrete stochasten, dus $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$. Voor het bepalen van de **mediaan** kijken we naar $x = F^{-1}(y)$ als $y = F(x)$. Het p-de **kwantiel** (quantile) van F is de waarde x_p z.d.d. $F(x_p) = p$ of $P(X \leq x_p) = p$. Speciale gevallen hiervan zijn de mediaan met $p = 1/2$ en de **kwartielen** (quartiles) met $p = 1/4$ en $p = 3/4$.

§2.2.1 De **exponentiële dichtheidsfunctie** (exponential density function) wordt gegeven door $f(x) = \lambda e^{-\lambda x}$ als $x \geq 0$ en 0 als $x < 0$. Hieruit volgt direct dat voor $x \geq 0$ volgt dat $F(x) = 1 - e^{-\lambda x}$. De exponentiële verdeling is geheugenloos, wat wil zeggen $P(Y > y|Y > x) = P(Y > y)$ of $P(X \geq x + y) = P(x)P(y)$(zie evt. blz. 51). Dit is karakteristiek voor de exponentiële verdeling.

§2.2.3 (merk op: we slaan §2.2.2 over) De **normale verdeling** (normal distribution) hangt af van twee parameters, μ en σ, en er geldt $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ met $-\infty < x < \infty$. μ heet het gemiddelde en σ de standaardafwijking. X is normaal verdeeld met parameters μ en σ noteren we ook als $X \sim N(\mu, \sigma^2)$.

§2.3 Het enige belangrijke hier is een viertal proposities:

- Als $X \sim N(\mu, \sigma^2)$ en $Y = aX + b$, dan $Y \sim N(a\mu + b, a^2\sigma^2)$.
- Laat X een continue stochast met dichtheid $f(x)$ zijn en laat $Y = g(X)$ met g differentieerbaar en strikt monotoon op een interval I. Neem aan $f(x) = 0$ als $x \notin I$. Dan heeft Y dichtheidsfunctie $f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy} g^{-1}(y)\right|$ met $y = g(x)$ voor bepaalde x en $f_Y = 0$ als $y \neq g(x)$ voor x in I.
- Laat $Z = F(X)$, dan heeft Z een uniforme verdeling op $[0, 1]$.
- Laat U uniform verdeeld zijn op $[0, 1]$ en $X = F^{-1}(U)$, dan de cdf van X is F. Dit betekent dus dat als we willekeurige getallen met cdf F willen maken, we alleen F^{-1} op deze getallen hoeven toe te passen.

3. Joint Distributions

§3.2 (§3.1 is alleen de inleiding) Neem aan dat X en Y discrete stochasten zijn, gedefinieerd op dezelfde uitkomstenruimte. Dan is hun **gezamenlijke frequentiefunctie** (joint frequency funtion) of **gezamenlijke kansfunctie** (joint probability mass function) $p(x, y)$ is $p(x_i, y_j) = P(X = x_i, Y = y_i)$. Als we hieruit de frequenctiefunctie van X willen krijgen, noemen we dit de **marginale frequentiefunctie** van X (marginal frequency function) en er geldt dat dit gelijk is aan $p_X(x) = \sum_i p(x, y_i)$.

§3.3 Neem aan dat X en Y continue stochasten zijn met gezamenlijke cdf $F(x, y)$. Hun **gezamenlijke dichtheidsfunctie** (joint density function) is continu en een functie van x en y, $f(x, y)$. Deze $f(x, y)$ is niet negatief en

$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$. Zo ook volgt dat $f(x,y) = \dfrac{\partial^2}{\partial x \partial y}F(x,y)$.
De **marginale cdf** (marginal cdf) van X, F_X, is $F_X(x) = P(X \leq x)$
$= \int_{-\infty}^{x}(\int_{-\infty}^{\infty} f(u,y)dy)du$. Hieruit volgt dat de **marginale dichtheid** (marginal density) van $X = f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$. Dit principe zet zich ook voort bij dimensies groter dan twee.

§3.4 Voor zowel discrete en continue stochasten X en Y geldt dat zij onafhankelijk zijn als hun gezamenlijke cdf ontbonden kan worden in het product van de marginale cdf's: $F(x_1,\ldots,x_n) = F_{X1}(x_1)\cdots F_{Xn}(x_n)$. Voor discrete stochasten is het equivalent om te zeggen dat de gezamenlijk frequentiefunctie ontbonden kan worden, terwijl het bij continue stochasten equivalent is met het ontbinden van de dichtheidsfuntie.

§3.5 Als discrete stochasten X en Y gezamenlijk verdeeld zijn, dan is de conditionele kans dat $X = x_i$, gegeven dat $Y = Y_j$, als $p_Y(y_j) > 0, P(X = x_j|Y = y_j) = \dfrac{P(X = x_j, Y = y_j)}{P(Y = y_j)} = \dfrac{p_{XY}(x_i,y_j)}{p_Y(y_j)}$. Indien $p_Y(y_j) = 0$, dan is de conditionele kans ook 0. De conditionele kans noteren we als $p_{Y|X}(y|x)$. In het discrete geval is de vergelijking ook te herschrijven als $p_X(x) = \sum_y p_{X|Y}(x|y)p_y(y)$.
Als X en Y gezamenlijke, continue stochasten zijn, wordt de conditionele dichtheid van Y, gegeven X, gegeven door $f_{Y|X}(y|x) = \dfrac{f_{XY}(x,y)}{f_X(x)}$ indien $0 < f_X(x) < \infty$ en anders is deze 0.

4. Expected Values:

§4.1 Als X een discrete stochast is met frequentiefunctie $p(x)$ dan is de **verwachtingswaarde** (expected value) van X, genoteerd door $E(X)$, is $E(X) = \sum_i x_i p(x_i)$ indien de som kleiner is dan oneindig. Als de som divergeert, is $E(X)$ niet gedefinieerd. $E(X)$ wordt ook wel het gemiddelde van X genoemd. Als X een continue stochast is met dichtheid $f(x)$ dan is de verwachtingswaarde $E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$ indien de integraal kleiner is dan oneindig.

Als X Poisson-verdeeld is, is $E(X) = \lambda$. Als X normaal verdeeld is, is $E(X) = \mu$. Markovs ongelijkheid: als X een stochast is met $P(X \geq 0) = 1$ waarvoor $E(X)$ bestaat, dan $P(X \geq t) \leq E(X)/t$. Simpel gezegd: de kans dat X veel groter is dan $E(X)$, is klein.

§4.1.1 We nemen aan dat $Y = g(X)$ en dat onderstaande som en integraal kleiner dan oneindig zijn.

- Als X discreet met frequentiefunctie $p(x)$, dan $E(X) = \sum_x g(x)p(x)$.
- Als X continu met dichtheidsfuntie $f(x)$ dan $E(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Verder is het belangrijk om op te merken dat $E(g(X)) \neq g(E(X))$.

§4.1.2 Verwachtingswaarden zijn lineair. Hieruit volgt dat de regenregels voor verwachtingswaarden de volgende zijn:

- $E(X+Y) = E(X) + E(Y)$.
- $E(cX) = cE(X)$.
- $E(c) = c$.
- $E(XY) = E(X)E(Y)$ als X en Y onafhankelijk zijn.

Algemener: als $X_1, \ldots, X_n$ gezamenlijk verdeelde stochasten zijn met verwachtingswaarden $E(X_i)$ en als Y een lineaire functie van de X_i is, dus $Y = a + \sum_{i=1}^{n} b_i X_i$, dan $E(Y) = a + \sum_{i=1}^{n} b_i E(X_i)$.

§4.2 Als X een stochast is met verwachtingswaarde $E(X)$, dan is de **variantie** (variance) van $X = \mathrm{Var}(X) = E((X - E(X)^2)$. De **standaardafwijking** (standard deviation) van X is de wortel uit de variantie. De variantie wordt vaak genoteerd met σ^2 en de standaardafwijking met σ.

Als rekenregel geldt dat $\mathrm{Var}(X) = E(X^2) - E^2(X)$. Hieruit volgt:

- Als $\mathrm{Var}(X)$ bestaat en $Y = a + bX$, dan $\mathrm{Var}(Y) = b^2\mathrm{Var}(X)$. Zo ook geldt voor de standaardafwijking dat $\sigma_Y = |b|\sigma_X$.
- Als X en Y onafhankelijk zijn, dan $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
- Als $c \in \mathbb{R}$ dan $\mathrm{Var}(c) = 0$.

§4.3 Als X en Y gezamenlijk verdeelde stochasten zijn met verwachtingen μ_X en μ_Y, dan is de **covariantie** (covariance) van X en $Y = \mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$. De covariantie is een maat voor de de samenhang van X en Y. Als rekenregel geldt dat $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$. Als X en Y onafhankelijk zijn, dan $E(XY) = E(X)E(Y) \Rightarrow \mathrm{Cov}(X, Y) = 0$ (maar $\Leftarrow$ geldt niet!).
De volgende rekenregels voor de covariantie gelden:

- $\mathrm{Cov}(a + X, Y) = \mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$.

Als X en Y gezamenlijk verdeelde stochasten zijn, en de varianties en covarianties van X en Y bestaan en zijn niet 0, dan is de **correlatie** (correlation) van X en Y, genoteerd door ρ, is $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$.

Dit was de stof van de tussentoets. Het volgende is de stof die daarna behandeld is.

§4.5 De **momentgenerende functie** (moment-generating function, mgf) van een stochast X is $M(t) = E(e^{tX})$. In het discrete geval is dit $M(t) = \sum_x e^{tx} p(x)$ en in het continue geval is dit $\int_{-\infty}^{\infty} e^{tx} f(x)\mathrm{d}x$. Als de momentgenererende functie bestaat voor alle t in een open interval rond 0, dan bepaalt het de kansverdeling uniek. Het r-de moment van een stochast is $E(X^r)$. Het r-de centrale moment wordt gegeven door $E((X - E(X))^r)$. De variantie is dus het tweede centrale moment. Het derde centrale moment staat bekend als de **scheefheid** (skewness) en is een maat van asymmetrie van de verdeling t.o.v. zijn μ.

De momentgenererende functie heeft de volgende eigenschappen:

- Als de momentgenererende functie bestaat, dan $M^{(r)}(0) = E(X^r)$.
- Als X de mgf $M_X(t)$ heeft, en $Y = a + bX$, dan heeft Y de mgf $M_Y(t) = e^{at}M_x(bt)$.
- Als X en Y onafhankelijke stochasten zijn met mgf's $M_X(t)$ en $M_Y(t)$, en $Z = X + Y$, dan $M_Z(t) = M_X(t)M_Y(t)$. Dit zet zich ook voort voor sommen van > 2 stochasten.

5. Limit Theorems

§5.2 **Wet van de grote getallen** (law of large numbers): laat $X_1, X_2, \ldots, X_i, \ldots$ een rij zijn van onafhankelijke stochasten met $E(X_i) = \mu$ en $\text{Var}(X_i) = \sigma^2$. Laat $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dan geldt $\forall \epsilon > 0$ dat $P(|\overline{X}_n - \mu| > \epsilon) \to 0$ indien $n \to \infty$. Indien voor een rij Z_n geldt dat $P(|Z_n - \alpha| > \epsilon)$ gaat naar 0 als n gaat naar ∞ en $\alpha \in \mathbb{R}$, dan zeggen we dat Z_n in kans convergeert naar α. We zeggen dat Z_n bijna zeker naar α convergeert als $\forall \epsilon > 0$ $|Z_n - \alpha| > \epsilon$ voor een eindig aantal keer met kans 1. Het verschil is dus vanaf een bepaald moment kleiner dan ϵ, maar we kunnen niet zeggen vanaf welk punt.

Een vergelijking die vaak opduikt bij bewijzen met de wet van de grote getallen is de **ongelijkheid van Chebyshev**. Laat X een stochast zijn met gemiddelde μ en variantie σ^2. Dan geldt voor alle $t > 0$ dat $P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$.

§5.3 Laat $X_1, X_2, \ldots$ een rij stochasten met cdf $F_1, F_2, \ldots$ en laat X een stochast met verdelingsfunctie F. We zeggen dat X_n in verdeling convergeert naar X als $\lim_{n\to\infty} F_n(x) = F(x)$ op elk punt waar F contin is.
Laat F_n een rij cumulatieve verdelingsfuncties met bijbehorende momentgenererende functie M_n. Laat F een cdf zijn met mgf M. Als $M_n(t) \to M(t)$ voor alle t in een open interval rond 0, dan $F_n(x) \to F(x)$ op alle continue punten van F.

We standaardiseren een verdeling S_n naar Z_n door $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. De centrale limietstelling zegt dat Z_n convergeert naar de standaard normale verdeling.

Centrale limietstelling (central limit theorem): laat $X_1, X_2, \ldots$ een rij van onafhankelijke stochasten met gemiddelde 0 en variantie σ^2 met een gezamenlijke verdelingsfunctie F en momentgenererende functie M gedefinieerd rond 0. Noem $S_n = \sum_{i=1}^{n} X_i$. Dan $\lim_{n\to\infty} P(\frac{S_n}{\sigma\sqrt{n}} \leq x) = \Phi(x)$.

6. Distributions Derived from the Normal Distribution

§6.2 Als Z een standaard normaal verdeelde stochast is, dan is de verdeling $U = Z^2$ de χ^2-verdeling met 1 **vrijheidsgraad** (degree of freedom). We noteren dit als χ^2_1. Als $U_1, U_2, \ldots, U_n$ onafhankelijke χ^2-verdelingen zijn met 1 vrijheidsgraad, dan is de verdeling van $V = U_1 + U_2 + \ldots + U_n$ de χ^2-verdeling met n vrijheidsgraden, genoteerd als $V \sim \chi^2_n$.

Als $Z \sim N(0,1)$ en $U \sim \chi^2_n$ en Z en U zijn onafhankelijk, dan is de verdeling van $\frac{Z}{\sqrt{U/n}}$ de t-verdeling met n vrijheidsgraden.

Laat U en V onafhankelijke χ^2-stochasten zijn met respectievelijk m en n vrijheidsgraden. De verdeling van $\frac{U/m}{V/n}$ heet de F-verdeling met m en n vrijheidsgraden en wordt genoteerd als $F_{m,n}$.

§6.3 Laat $X_1, \ldots X_n$ onafhankelijke $N(\mu, \sigma^2)$ stochasten zijn. We refereren hier ook wel naar als een **steekproef** (sample) van een normale verdeling. Verder noemen we $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ en $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
Nu geldt er:

- De stochast $\overline{X}$ en de vector van de stochasten $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ zijn onafhankelijk.
- $\overline{X}$ en S^2 zijn onafhankelijke verdeeld.
- De verdeling van $(n-1)S^2/\sigma^2$ is de χ^2-verdeling met $n-1$ vrijheidsgraden.
- $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

7. Survey Sampling

Sample surveys (steekproeven) worden gebruikt om informatie over een (zeer) grote populatie te vergaren, door een steekproef/sample te nemen en deze te onderzoeken. De informatie over de steekproef wordt dan gebruikt om uitspraken te doen over de hele populatie.

7.1. **Populatieparameters.** Neem aan dat je een populatie hebt van grootte N en dat elk lid van de populatie gerepresenteerd kan worden met een nummer $x_1, x_2, \ldots, x_n$. De *population mean* (populatieverwachting) is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

en de *population total* (de totale populatie) wordt gegeven door

$$\tau = \sum_{i=1}^{N} x_i = N\mu.$$

Verder is de *population variance* (de populatievariantie) gelijk aan

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2.$$

en dan is natuurlijk de *population standard deviation* (standaardafwijking van de populatie) de wortel uit de *population variance*.

7.2. **Simple Random Sampling.** *Simple random sampling* (eenvoudige aselecte steekproef) is de meest elementaire vorm van *sampling*. Elke steekproef van dezelfde grootte n heeft ook dezelfde kans om voor te komen, oftewel: elke van de $\binom{N}{n}$ mogelijke steekproeven van grootte n heeft dezelfde waarschijnlijkheid om voor te komen. De steekproeven worden zonder teruglegging gedaan zodat ieder lid van de populatie ten hoogste één keer voorkomt in een steekproef. De fractie $\frac{n}{N}$ heet de *sampling fraction*.
Noem de waarden van de leden van de populatie $X_1, X_2, \ldots, X_n$. Iedere X_i is een *random variable* (stochast) en dus niet hetzelfde als x_i, de waarde van het i-de lid van de populatie (die ligt dus vast). De *sample mean* is gelijk aan

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

en de *population total* is

$$T = N\overline{X}$$

Merk op dat de sample mean ook een *random variable* is. De kansverdeling van de *sample mean* (en andere *statistics*[1]) heet de *sampling distribution*. De *sampling distribution* van $\overline{X}$ legt vast hoe nauwkeurig $\overline{X}$ μ benadert: Hoe 'strakker' de *sampling distribution* om μ gecentreerd ligt, hoe beter de benadering.

Lemma 1. *Beschouw een populatie van grootte N. Noem de verschillende waarden die de populatieleden aannemen $x_1, x_2, \ldots, x_n$ en noem het aantal leden die zo'n*

[1]Een statistic is niets anders dan een enkele meetwaarde van een eigenschap van de steekproef, zie ook `http://en.wikipedia.org/wiki/Statistic`

bepaalde waarde x_i hebben $n_i, i = 1, \ldots, m$. Dan is X_i een discrete stochast met probability mass function

$$P(X_i = x_j) = \frac{n_j}{N},$$

en

$$\mathrm{E}(X_i) = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2$$

Stelling 1. *Bij simple random sampling geldt*

$$\mathrm{E}(\overline{X}) = \mu.$$

Hieruit volgt

$$\mathrm{E}(T) = \tau.$$

Als we een populatieparameter, zeg θ, willen benaderen met een functie $\hat{\theta}$ van een *sample* $X_1, X_2, \ldots, X_n$, en we vinden $E(\hat{\theta}) = \theta$, dan noemen we $\hat{\theta}$ *unbiased* (zuiver). Als een functie zuiver is, geldt ook dat de *mean squared error* gelijk is aan de variantie.

Lemma 2. *Voor simple random sampling zonder teruglegging geldt*

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \qquad \textit{alsi} \neq j$$

Stelling 2. *Voor simple random sampling geldt*

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Het verschil tussen variantie bij *simple random sampling* zonder terugleggen verschilt van *simple random sampling* met teruglegging door de factor $\left(1 - \frac{n-1}{N-1}\right)$, de *finite population correction*. Vaak is de *sampling fraction* heel klein, zodat we voor $\overline{X}$ kunnen schrijven

$$\sigma_{\overline{X}} \approx \frac{\sigma}{\sqrt{n}}$$

De spreiding van de *sampling distribution* (en dus de precisie van $\overline{X}$) wordt dus bepaald door de steekproefgrootte n en niet de populatiegrootte N. De andere factor die de precisie van je steekproefgemiddelde beïnvloedt is de standaard deviatie, σ (en dat is ook logisch, want als σ klein is, zijn de populatiewaarden niet zo ver verspreid en is je gemiddelde accurater dan wanneer de σ groot is en de populatiewaarden dus ver uit elkaar liggen).

Gevolg 1. *Voor simple random sampling geldt*

$$\mathrm{Var}(T) = N^2\left(\frac{\sigma^2}{n}\right)\frac{N-n}{N-1}$$

7.2.1. *Estimation of the population variance.* We gaan nu kijken naar hoe de populatievariantie benaderd kan worden uit de steekproef van de populatie. De populatievariantie is de gemiddelde standaardafwijking van het populatiegemiddelde in het kwadraat, dus we kunnen hem ook zo benaderen:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

en deze is onzuiver:

Stelling 3. *Met simple random sampling,*

$$E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)\frac{N}{N-1}$$

Gevolg 2. *Een zuivere schatter voor* $\mathrm{Var}(\overline{X})$ *is*

$$s^2_{\overline{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

met

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

Een zuivere schatter van $\mathrm{Var}(\hat{p})$ *is*

$$s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$$

en een zuivere schatter voor de $\mathrm{Var}(T)$ *is*

$$s^2_T = N^2 s^2_{\overline{X}}$$

$s_{\overline{X}}, s_T$ en $s_{\hat{p}}$ heten *estimated standard errors.*

Een samenvatting van het voorgaande kan gegeven worden door de volgende tabel:

Population parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$s^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$s^2_{\overline{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$
p	$\hat{p}$ =sample proportion	$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$
τ	$T = N\overline{X}$	$\sigma^2_T = N^2\sigma^2_{\overline{X}}$	$s^2_T = N^2 s^2_{\overline{X}}$
σ^2	$\left(1 - \frac{n}{N}\right)s^2$		

met $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. De wortel uit de *variance of estimate* heet de standaard error, en de wortel uit de *estimated variance* heet de *estimated standard error.* NB: Let goed op het verschil tussen een breuk en een subscript-$\overline{X}$.

7.2.2. *Normal approximation to the sampling distribution of* $\overline{X}$*.* Beschouw een reeks i.i.d. random variables $X_1, X_2, \ldots, X_2$, met verwachting μ en variantie σ^2. We weten al dat de *sample mean* van $X_1, X_2, \ldots, X_n$ gelijk is aan $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, met $\mathrm{E}(\overline{X}_n) = \mu$ en $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$. Dan geeft de Centrale Limietstelling dat voor een vaste z

$$P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \to \Phi(z) \qquad \text{wanneer} \qquad n \to \infty$$

met Φ de cdf van de standaard normale verdeling.

Nu kunnen we een *confidence interval* (betrouwbaarheidsinterval) voor het populatieverwachting, μ maken. Een betrouwbaarheidsinterval voor een populatieparameter, zeg θ, is een random interval dat θ bevat met een gespecificeerde waarschijnlijkheid. Voor $0 \leq \alpha \leq 1$, zij $z(\alpha)$ het getal zodat de oppervlakte van de standaard normale kansdichtheidsfunctie rechts van $z(\alpha)$ gelijk is aan α. Als Z standaard normaal verdeeld is, volgt

$$P(-z(\alpha/2) \leq Z \leq z(\alpha/2)) = 1 - \alpha$$

Merk op dat, vanwege de symmetrie van de standaard normale kansdichtheidsfunctie volgt dat $z(1-\alpha) = -z(\alpha)$. Uit de Centrale Limietstelling volgt dan

$$P\left(-z(\alpha/2) \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq z(\alpha/2)\right) \approx 1 - \alpha, \text{oftewel}$$

$$P\left(\overline{X} - z(\alpha/2)\sigma_{\overline{X}} \leq \mu \leq \overline{X} + z(\alpha/2)\sigma_{\overline{X}}\right) \approx 1 - \alpha$$

dus de kans dat μ in het interval $\overline{X} \pm z(\alpha/2)\sigma_{\overline{X}}$ ligt is ongeveer $1 - \alpha$. Het interval heet dus een $100(1-\alpha)\%$ *confidence interval.*

7.3. **Stratified Random Sampling.** Bij *stratified random sampling* (gestratificeerde aselecte steekproeven) is de populatie verdeeld in subpopulaties, *strata*, die dan onafhankelijk van elkaar worden gebruikt voor de steekproeven. De resultaten hiervan worden dan gebruikt om uitspraken te doen over de totale populatie.

Beschouw een populatie met L strata. Het aantal populatie-elementen in stratum l wordt genoteerd met N_l. In totaal zijn er $N = N_1 + N_2 + \ldots + N_L$ populatie-elementen. De populatieverwachting en -variantie worden per stratum genoteerd als resp. μ_l en σ_l^2. Dan is de totale populatieverwachting

$$\mu = \sum_{l=1}^{L} W_l \mu_l,$$

waarbij $W_l = N_l/N$. De *sample mean* van elk stratum is

$$\overline{X}_l = \frac{1}{n_l} \sum_{i=1}^{L} X_{il},$$

met X_{il} de i-de steekproef uit het l-de stratum. Uit Stelling 1 volgt nu dat $\mathrm{E}(\overline{X}_l) = \mu_l$. Analoog met de eerdere relatie tussen de het gemiddelde van de totale populatie en de verwachtingen van de strata van de populatie volgt nu de schatter voor μ:

$$\overline{X}_s = \sum_{l=1}^{L} W_l \overline{X}_l$$

Stelling 4. *De gestratificeerde schatting, $\overline{X}_s$, van het populatiegemiddelde is zuiver*

Stelling 5. *De variantie van het gemiddelde van de gestratificeerde steekproef wordt gegeven door*

$$\mathrm{Var}(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l}\right) \left(1 - \frac{n_l - 1}{N_l - 1}\right) \sigma_l^2$$

Gevolg 3. *De verwachting en variantie van de gestratificeerde verwachting van het populatietotaal zijn*

$$\mathrm{E}(T_s) = \tau$$

en

$$\mathrm{Var}(T_s) = N^2 \mathrm{Var}(\overline{X}_s)$$

Om de standaard errors van $\overline{X}_s$ en T_s te schatten, moeten de varianties van de afzonderlijke strata apart geschat worden en in de voorgaande formules gesubstitueerd worden. De schatting voor σ_l^2 wordt gegeven door

$$s_l^2 = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (X_{il} - \overline{X}_l)^2$$

en de schatting voor $\mathrm{Var}(\overline{X}_s)$

$$s_{\overline{X}}^2 = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l}\right) \left(1 - \frac{n_l}{N_l}\right) s_l^2$$

Als je maar n trekkingen kunt doen uit de populatie, kun je je afvragen hoe je de $n_i, i = 1, \ldots, L$ kunt kiezen zodat $\mathrm{Var}(\overline{X}_s)$ geminimaliseerd wordt met behoud van de voorwaarde $n_1 + \ldots + n_L = n$. De volgende stelling is daar handig voor

Stelling 6. *De steekproefgroottes* $n_1, \ldots, n_L$ *die* $\mathrm{Var}(\overline{X}_s)$ *minimaliseren met behoud van* $n_1 + \ldots + n_L = n$ *worden gegeven door*

$$n_l = n \frac{W_l \sigma_l}{\sum_{k=l}^{L} W_k \sigma_k},$$

met $l = 1, \ldots, L$.

Intuïtief betekent dit dat strata met hoge $W_l \sigma_l$ grotere steekproeven krijgen (en dat is logisch, want als W_l groot is, bevat het stratum een grote fractie van de populatie, en als σ_l groot is, zijn de waarden van de populatie-elementen erg verdeeld. Dan wil je dus ook een grote steekproef uit dat stratum nemen om een goed beeld te krijgen van het gemiddelde van het stratum (ook wel *Neyman allocatie* genoemd)).

Gevolg 4. *Zij* $\overline{X}_{so}$ *de gestratificeerde schatting die gevonden is met de voorgaande stelling en negeer de eindige populatiecorrectie. Dan volgt*

$$\mathrm{Var}(\overline{X}_{so}) = \frac{\left(\sum_{l=1}^{L} W_l \sigma_l\right)^2}{n}$$

Deze optimale allocatie hangt af van de varianties van de individuele strata. Maar deze weet je over het algemeen niet. Een alternatief is het gebruik van *proportionele allocatie*. Hiervoor gebruik je dezelfde *sampling* fractie in elk stratum: $\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}$ wanneer $n_l = n\frac{N_l}{N} = nW_l$ voor $l = 1, \ldots, L$. De schatting voor het gemiddelde van de populatie is dan

$$\overline{X}_{sp} = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{il}$$

Stelling 7. *Als we gebruik maken van gestratificeerde trekking gebaseerd op proportionele allocatie en we de eindige populatiecorrectie negeren vinden we*

$$\mathrm{Var}(\overline{X}_{sp}) = \frac{1}{n} \sum_{l=1}^{L} W_l \sigma_l^2$$

Stelling 8. *Als we gebruik maken van gestratificeerde aselecte trekkingen, is het verschil tussen de variantie van de schatting van het populatiegemiddelde gebaseerd op proportionele allocatie, en de variantie van die schatting op basis van optimale allocatie (en we weer de eindige populatiecorrectie negeren)*

$$\mathrm{Var}(\overline{X}_{sp}) - \mathrm{Var}(\overline{X}_{so}) = \frac{1}{n} \sum_{l=1}^{L} W_l (\sigma_l - \overline{\sigma})^2,$$

met

$$\overline{\sigma} = \sum_{l=1}^{L} W_l \sigma_l$$

Intuïtief betekent dit dat als de varianties van de afzonderlijke strata gelijk zijn, proportionele allocatie hetzelfde resultaat geeft als optimale allocatie (hoe meer de varianties van de afzonderlijke strata verschillen, hoe beter het is om optimale allocatie te gebruiken).

Stelling 9. *Het verschil tussen de variantie van het gemiddelde van een simple random sample en de variantie van het gemiddelde van een stratified random sample*

gebaseerd op proportionele allocatie is (de eindige populatiecorrectie negerend)

$$\mathrm{Var}(\overline{X}) - \mathrm{Var}(\overline{X}_{sp}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\mu_l - \mu)^2$$

Stratified random sampling met proportionele allocatie geeft dus altijd een kleinere variantie dan *simple random sampling* (wanneer je de eindige populatiecorrectie negeert). *Stratified random sampling* met proportionele allocatie is beter dan *simple random sampling* wanneer de gemiddelden van de strata ver uiteenlopen (en *stratified random sampling* met optimale allocatie is nog beter).

8. Estimation of Parameters and Fitting of Probability Distribution

8.1. **Method of Moments.** Het k'de moment van de *probability law* is gedefinieerd als $\mu_k = \mathrm{E}(X^k)$, waarbij X een stochast is die de *probability law* volgt. Als $X_1, X_2, \ldots X_n$ i.i.d. zijn, dan is het k'de *sample moment* gedefinieerd als

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Definitie 1. *Zij $\hat{\theta}$ een schatting van een parameter θ gebaseerd op een steekproef van grootte n. Dan heet $\hat{\theta}_n$ consistent in kans als $\hat{\theta}_n$ in kans convergeert naar θ als n naar oneindig gaat. Dus $\forall\ \varepsilon > 0$,*

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0, \qquad \textit{als} \qquad n \to \infty.$$

8.2. **Method of Maximum Likelihood.** Neem aan dat stochasten $X_1, X_2, \ldots, X_n$ een *joint density function* of *joint frequency function* $f(x_1, x_2, \ldots, x_n|\theta)$ hebben. Zij $X_i = x_i, i = 1, 2, \ldots, n$ de geobserveerde data. Dan heet de *likelihood function* van θ als functie van $x_1, x_2, \ldots, x_n$

$$\mathrm{lik}(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$$

De *maximum likelihood estimate* van θ is die waarde van θ die de *likelihood* maximaliseert (de data het meest waarschijnlijk maakt). Als de X_i i.i.d. , dan is de *joint density* het product van de marginale dichtheden en de *likelihood* is

$$\mathrm{lik} = \prod_{i=1}^{n} f(X_i|\theta)$$

In plaats van dat we de *likelihood* maximaliseren, maximaliseren we zijn natuurlijk logaritme. Dan wordt de *log-likelihood*

$$l(\theta) = \sum_{i=1}^{n} \log\left[f(X_i|\theta\right]$$

Stelling 10. *Onder bepaalde gladheidsvoorwaarden op f is de mle van een i.i.d. steekproef consistent.*

Lemma 3. *Definieer $I(\theta)$ door*

$$I(\theta) = \mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2$$

Onder bepaalde gladheidsvoorwaarden op f kan $I(\theta)$ uitgedrukt worden als

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

De verdeling van een *maximum likelihood estimate* met een grote steekproef is bij benadering normaal met verwachting θ_0 en variantie $1/[nI(\theta_0)]$. Omdat dit alleen werkt met een zeer grote steekproef zeggen we dat de mle *asymptotisch zuiver* is, en de variantie noemen we dan de *asymptotische variantie van de mle.*

Stelling 11. *Onder bepaalde gladheidsvoorwaarden op f benadert de kansverdeling van $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ de standaard normale verdeling.*

8.3. **The Bayesian Approach to Parameter Estimation.** Bij de Bayesiaanse benadering beschouwen we θ als stochast met *prior distribution* (a priori verdeling) $f_\Theta(\theta)$ (dit representeert wat we weten over de parameter voor we data observeren). Dan wordt de *posterior distribution* (a posteriori verdeling) gegeven door

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

En dit geeft ons

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_{X|\Theta}(x|\theta) \times f_\Theta(\theta) \\ \text{Posterior density} &\propto \text{Likelihood} \times \text{Prior density} \end{aligned}$$

De *posterior mean* is het gemiddelde van de a posteriori verdeling. De *posterior mode* is de waarde van de maximum likelihoodfunctie.

Conjugate priors zijn a priori verdelingen waarvoor geldt dat als deze tot een familie verdelingen behoort, de a posteriori verdeling tot *dezelfde* verdeling behoort. Oftewel: Als de a priori verdeling tot een familie G behoort, en de data volgen een verdeling H, dan is G geconjugeerd met H als de a posteriori verdeling ook tot de familie G behoort.
Improper priors zijn a priori verdelingen die geen kansverdelingen zijn.

8.4. **Efficiency and the Cramér-Rao Lower Bound.** De efficiëntie van schatters $\hat{\theta}$ ten opzichte van schatter $\tilde{\theta}$ van een parameter θ is

$$\text{eff} = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}$$

Stelling 12. *(Cramér-Rao ongelijkheid)*
Zij $X_1, X_2, \ldots, X_n$ i.i.d. met kansdichtheid $f(x|\theta)$. Zij $T = t(X_1, X_2, \ldots, X_n)$ een zuivere schatter voor θ. Dan geldt, onder bepaalde gladheidsvoorwaaren op $f(x|\theta)$,

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

8.5. **Sufficiency.**

Definitie 2. *Een statistiek $T(X_1, \ldots, X_n$ heet voldoend voor θ als de voorwaardelijke verdeling van $X_1, \ldots, X_n$ gegeven $T = t$ niet afhankelijk is van θ voor alle waarden van t.*

Stelling 13. *Een noodzakelijke en voldoende voorwaarde voor $T(X_1, \ldots, X_n)$, die voldoend is voor parameter θ, is dat de gezamenlijke kansdichtheid te schrijven is als*

$$f(x_1, \ldots, x_n|\theta) = g\left[T(x_1, \ldots, x_n), \theta\right] h(x_1, \ldots, x_n)$$

Leden van de exponentiële familie (met één variabele) hebben een kansdichtheid (of *frequency function*)van de vorm

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

Als je de kansdichtheid (of *frequency function*) van een bepaalde verdeling kunt schrijven in deze vorm, is hij ook lid van de exponentiële familie.

Gevolg 5. *Als T voldoend is voor θ, is de mle een functie van T.*

Stelling 14. *(Rao-Blackwell stelling)*
Laat $\hat{\theta}$ een schatter van θ is met $\mathrm{E}(\hat{\theta}^2) < \infty$ voor alle θ. Neem aan dat T voldoend is voor θ en zij $\tilde{\theta} = \mathrm{E}(\hat{\theta})|T)$. Dan geldt voor alle θ

$$\mathrm{E}(\tilde{\theta} - \theta)^2 \leq \mathrm{E}(\hat{\theta} - \theta)^2$$

De ongelijkheid is strict behalve wanneer $\hat{\theta} = \tilde{\theta}$.

9. Testing Hypotheses and Assessing Goodness of Fit

Beschouw een stochast X. We specificeren twee hypothesen, H_0 en H_1. We ontwikkelen een Bayesiaanse methode om één van de hypothesen aan te nemen. Hiervoor moet je de a priori kansen $P(H_0)$ en $P(H_1)$ specificeren voordat je de data gaat observeren. Na de observatie worden je a posteriorikansen $P(H_0|x)$ en $P(H_1|x)$. Dus $P(H_0|x) = \frac{P(x|H_0)P(H_0)}{P(x)}$. De ratio

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \frac{P(x|H_0)}{P(x|H_1)}$$

is het product van de ratio tussen de a priori kansen en de *likelihood* ratio.
Dan neem je H_0 aan als

$$\frac{P(x|H_0)}{P(x|H_1)} > c$$

waarbij c afhangt van je a priori kansen.

9.1. Neyman-Pearson Paradigm.

Neyman en Pearson geven een methode om hypothesen te testen als beslissingsprobleem (waar je geen gebruik maakt van het specificeren van a priori kansen). Je stelt weer twee hypothesen op, de nulhypothese (H_0) en de alternatieve hypothese (H_1 of H_A). Er staan twee fouten centraal:
Als je de nulhypothese weerlegt terwijl hij waar is, is er sprake van een *type I fout*.
Wanneer je de nulhypothese accepteert terwijl hij onwaar is, is er sprake van een *type II fout*. De waarschijnlijkheid van de type I fout (significantieniveau) wordt aangeduid met α en de waarshijnlijkheid van de type II fout wordt aangeduid met een β.
De kans dat de nulhypothese verworpen wordt terwijl hij onwaar is heet de *power* (onderscheidingsvermogen) van de test en is gelijk aan $1 - \beta$.
De *likelihood ratio* wordt ook wel de *test statistic* genoemd. De verzameling waarden van deze *test statistic* dat tot verwerping van de nulhypothese leidt heet de *rejection region*(kritiek gebied) en de verzameling waarden dat leidt tot aanname van de nulhypothese heet de*acceptance region* (aanvaardingsgebied). Ten slotte heet de kansverdeling van de *test statistic* wanneer de nulhypothese waar is de *null distribution.*

Hypothesen heten *simpel* wanneer ze de kansverdeling helemaal specificeren.

Lemma 4. *De likelihood ratio test die H_0 verwerpt ten gunste van H_1 (met H_0 en H_1 simpel), waarvoor geldt dat de ratio test kleiner is dan c en het significantieniveau kleiner of gelijk is aan α is de test met het hoogste onderscheidingsvermogen ten opzichte van andere testen met een significantieniveau kleiner of gelijk α.*

Vaak wordt voor α een waarde tussen 0.01 en 0.05 aangehouden. De waarde van de kleinste significantie waarop de nulhypothese verworpen wordt heet ook wel de p-waarde. Het is verder gebruik om je nulhypothese strategisch te kiezen (omdat er in de Neyman-Pearson methode een zekere asymmetrie tussen de twee voorkomt). Als één van de twee hypothesen simpeler is, kies je die als nulhypothese. Als de

consequenties van het verkeerd afwijzen van de ene hypothese zwaarder zijn dan die van de andere, kies je de eerste als nulhypothese.

Als de alternatieve hypothese niet simpel is, is een test die het hoogste onderscheidingsvermogen heeft van alle simpele alternatieve hypothesen, *uniformly most powerful* (die heeft uniform het hoogste onderscheidingsvermogen).

9.2. **The Duality of Confidence Intervals and Hypothesis Tests.** Zij θ een parameter van een familie van kansverdelingen, en zij Θ de verzameling van alle mogelijke waarden van θ. Schrijf de stochastische variabelen die de data leveren als $\mathbf{X}$.

Stelling 15. *Neem aan dat er voor iedere waarde θ_0 in Θ een test op niveau α bestaat van de hypothese $H_0 : \theta = \theta_0$. Schrijf het aanvaardingsgebied van de test als $A(\theta_0)$. Dan is*

$$C(\mathbf{X}) = \{\theta : \mathbf{X} \in A(\theta)\}$$

een $100(1-\alpha)\%$ confidence interval voor θ.

In woorden: Een $100(1-\alpha)\%$ *confidence interval* voor θ bevat alle waarden van θ_0 waarvoor de nulhypothese (zoals in de stelling gedefinieerd) niet afgewezen wordt bij significantieniveau α.

Stelling 16. *Neem aan dat $C(\mathbf{X})$ een $100(1-\alpha)\%$ confidence interval is voor θ. Dus, voor alle θ_0*

$$P\left[\theta_0 \in C(\mathbf{X}) | \theta = \theta_0\right] = 1 - \alpha.$$

Dan is een aanvaardingsgebied voor een test met niveau α van de nulhypothese $H_0(\theta = \theta_0)$

$$A(\theta_0) = \{\mathbf{X} | \theta_0 \in C(\mathbf{X})\}$$

In woorden: De nulhypothese wordt geaccepteerd wanneer θ_0 in de *confidence region* ligt.

Deze twee stellingen geven de dualiteit waarmee je *confidence intervals* voor parameters van kansverdelingen kunt maken, waarna je deze intervallen gebruikt om hypothesen te testen over de waarden van de parameters.

9.3. **Generalized Likelihood Ratio Tests.** De 'gewone' Likelihood Ratio Test kan alleen gebruikt worden wanneer zowel de nulhypothese als de alternatieve hypothese enkelvoudig zijn. Nu ontwikkelen we een Likelihood Ratio Test die die eis niet heeft. Deze test is echter meestal niet optimaal.

Neem aan dat je observaties $\mathbf{X} = (X_1, \ldots, X_n)$ hebt met *joint density function* $f(\mathbf{x}|\theta)$. H_0 specificeert dat $\theta \in \omega_0$, waar ω_0 een deelverzameling is van alle mogelijke waarden van θ. Laat H_1 specificeren dat $\theta \in \omega_1$, met ω_1 disjunct met ω_0. Zij $\Omega = \omega_1 \cap \omega_0$. Als de hypothesen samengesteld zijn, wordt de likelihood geëvalueerd in de waarde van θ die deze likelihood maximaliseert:

$$\Lambda^* = \frac{\max_{\theta \in \omega_0} [(\theta)]}{\max_{\theta \in \omega_1} [(\theta)]}$$

Vaker wordt deze GLR gebruikt:

$$\Lambda = \frac{\max_{\theta \in \omega_0} [(\theta)]}{\max_{\theta \in \Omega} [(\theta)]}$$

Het verwerpingsgebied voor een Likelihood Ratio Test bestaat uit kleine waarden van Λ. De drempelwaarde λ_0 is gekozen zodat $P(\Lambda \leq \lambda_0 | H_0) = \alpha$, het significantieniveau van de test.

Stelling 17. *Onder bepaalde gladheidsvoorwaarden op de kansverdeling gaat de nulverdeling van* $-2\log\Lambda$ *naar een Chi-kwadraatverdeling met het aantal vrijheidsgraden gelijk aan* $\dim\Omega - \dim\omega_0$ *als de steekproefgrootte naar oneindig gaat.*

9.4. **Likelihood Ratio Tests voor de Multinomiale verdeling.** We zullen een gegeneraliseerde Likelihood Ratio Test ontwikkelen om de *goodness of fit* van een multinomiaal model te bepalen. Zij $p = (p_1, \ldots, p_m)$ de observatievector met tellingen (m is het aantal 'categorieën'). Zij $X = (X_1, \ldots, X_m)$ de steekproeven. De kansfunctie van deze observaties is

$$\left(\frac{n!}{x_1!x_m!}\right) p_1(\theta)^{x_1}$$