
FORMELSAMLING VT14
MATEMATISK STATISTIK FÖR K, TMA073

Sannolikheteorins grunder

- Följande gäller för sannolikheter
 - * $0 \leq P(A) \leq 1$
 - * $P(\Omega) = 1$
 - * $P(A \cup B) = P(A) + P(B)$ om händelserna A och B är oförenliga (disjunkta)
- Additionssatsen för två händelser: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Betingad sannolikhet: $P(B|A) = \frac{P(A \cap B)}{P(A)}$
- Bayes sats: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Satsen om total sannolikhet: $P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$, där händelserna $B_1, \dots, B_n$ är parvis oförenliga händelser och $\bigcup_{i=1}^n B_i = \Omega$
- A och B är oberoende $\Leftrightarrow P(A \cap B) = P(A)P(B)$

Endimensionella stokastiska variabler

- Fördelningsfunktionen för X : $F_X(x) = P(X \leq x)$
- Sannolikhetsfunktionen för en diskret s.v. X : $p_X(k) = P(X = k)$
- Täthetsfunktionen för en kontinuerlig s.v. X : $f_X(x) = \frac{dF_X(x)}{dx}$
- $P(a < X \leq b) = F_X(b) - F_X(a) = \begin{cases} \sum_{k=a+1}^b p_X(k) & \text{om } X \text{ är diskret och } a \text{ och } b \text{ är heltal} \\ \int_a^b f_X(x)dx & \text{om } X \text{ är kontinuerlig} \end{cases}$

Flerdimensionella stokastiska variabler

- Simultan fördelningsfunktion:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \begin{cases} \sum_{i \leq x, j \leq y} p_{X,Y}(i,j), & \text{(diskret s.v.)} \\ \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(t,u) dt du, & \text{(kontinuerlig s.v.)} \end{cases}$$

- Marginell täthetsfunktion: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$
- $P(a \leq X \leq b \text{ och } c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y) dx dy$.
- X och Y är oberoende om
 - $p_{X,Y}(i,j) = p_X(i)p_Y(j)$ för alla i och j för diskreta s.v.
 - $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ för alla x och y för kontinuerliga s.v.

Väntevärden

- Väntevärdet av $g(X, Y)$:

$$E(g(X, Y)) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} g(i, j) p_{X,Y}(i, j), & (\text{diskret s.v.}) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy, & (\text{kontinuerlig s.v.}) \end{cases}$$

- Varians: $V(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$
- Standardavvikelse: $\sqrt{V(X)}$
- Kovarians: $C(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$
- Väntevärde av linjärkombination: $E(\sum_{i=1}^n a_i X_i + b) = \sum_{i=1}^n a_i E(X_i) + b$
- Varians av linjärkombination: $V(\sum_{i=1}^n a_i X_i + b) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j C(X_i, X_j)$
- Om $X_1, \dots, X_n$ är oberoende så är de okorrelerade, dvs $C(X_i, X_j) = 0, i \neq j$
- Korrelationskoefficient: $\rho(X, Y) = \frac{C(X, Y)}{\sqrt{V(X)V(Y)}}$

Vanliga fördelningar

Binomialfördelning $X \sim \text{Bin}(n, p)$

- $p(k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, \dots, n$
- $E(X) = np, V(X) = np(1-p)$
- $X = \sum_{i=1}^n Y_i$ där Y_i är oberoende och $\text{Bin}(1, p)$
- $X \sim \text{Bin}(n_1, p), Y \sim \text{Bin}(n_2, p)$ samt oberoende $\Rightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$
- $\text{Bin}(n, p)$ -fördelningen kan approximeras med $\text{Po}(np)$ -fördelningen om $p \leq 0.1$ och $n \geq 10$

Poissonfördelning $X \sim \text{Po}(\mu)$

- $p(k) = \frac{\mu^k}{k!} e^{-\mu}, k = 0, 1, \dots$
- $E(X) = V(X) = \mu$.
- $X \sim \text{Po}(\mu_1), Y \sim \text{Po}(\mu_2)$ samt oberoende $\Rightarrow X + Y \sim \text{Po}(\mu_1 + \mu_2)$

Geometrisk fördelning, $X \sim \text{Ge}(p)$

- $p(k) = p(1-p)^{k-1}, k = 1, 2, \dots$
- $E(X) = 1/p, V(X) = (1-p)/p^2$

Likformig fördelning $X \sim \text{U}(a, b)$

- $f(x) = \frac{1}{b-a}, a \leq x \leq b, f(x) = 0$ för övrigt.
- $E(X) = \frac{a+b}{2}, V(X) = \frac{(b-a)^2}{12}$

Exponentialfördelning $X \sim \text{Exp}(\lambda)$

- $f(x) = \frac{1}{\lambda} e^{-x/\lambda}, x \geq 0, f(x) = 0$ för övrigt.
- $E(X) = \lambda, V(X) = \lambda^2$

Normalfördelning $X \sim N(\mu, \sigma^2)$

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty \leq x \leq \infty$
- $E(X) = \mu, V(X) = \sigma^2$
- $X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$
- $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ där $\Phi(\cdot)$ ges av tabell.
- $X_i \sim N(\mu_i, \sigma_i^2), i = 1, \dots, n$ oberoende $\Rightarrow \sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$

t-fördelning $X \sim t(\nu)$

- $f(x) = \frac{1}{\sqrt{\mu\pi}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, -\infty \leq x \leq \infty$
- $X \sim N(0, 1), Y \sim \chi^2(\nu)$ samt oberoende $\Rightarrow \frac{X}{\sqrt{Y/\nu}} \sim t(\nu)$

 χ^2 -fördelning $X \sim \chi^2(n)$

- $f(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} e^{-x/2} x^{n/2-1}, x \geq 0, f(x) = 0$ för övrigt.
- $E(X) = n, V(X) = 2n$
- $X_1, \dots, X_n$ oberoende och $N(0, 1) \Rightarrow \sum_{i=1}^n X_i^2 \sim \chi^2(n)$
- $X_1, \dots, X_n$ oberoende och $N(\mu, \sigma^2) \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$

F-fördelning $X \sim F(n, m)$

- $f(x) = \frac{\Gamma(\frac{n+m}{2}) n^{n/2} m^{m/2}}{\Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \cdot \frac{x^{(n-2)/2}}{(m+nx)^{(n+m)/2}}, x \geq 0, f(x) = 0$ för övrigt.
- $E(X) = \frac{m}{m-2}$ om $m > 2, V(X) = \frac{m^2(2m+2n-4)}{n(m-2)^2(m-4)}$ om $m > 4$
- $X \sim \chi^2(n), Y \sim \chi^2(m)$ samt oberoende $\Rightarrow \frac{X/n}{Y/m} \sim F(n, m)$
- $F_{1-\alpha}(n, m) = 1/F_\alpha(m, n)$

Centrala gränsvärdessatsen

Om $X_1, \dots, X_n$ är oberoende och likafördelade med $E(X_i) = \mu_i$ och $V(X_i) = \sigma^2$ så gäller att $\sum_{i=1}^n X_i$ är approximativt $N(n\mu, n\sigma^2)$ -fördelad om n är stort nog. Från detta följer att

- $\text{Po}(\mu) \approx N(\mu, \mu)$ om $\mu \geq 15$
- $\text{Bin}(n, p) \approx N(np, np(1-p))$ om $np(1-p) \geq 10$

Statistik och punktskattningar

Beskrivning av data:

- Stickprovsmedelvärde: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Stickprovsvarians: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} [\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2]$
- Stickprovskovarians: $c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- Korrelationskoefficient: $r_{xy} = \frac{c_{xy}}{s_x s_y}$

Låt $x_1, \dots, x_n$ vara observationer av oberoende och likafördelade s.v. $X_1, \dots, X_n$ med väntevärde μ och varians σ^2 , då är stickprovsmedelvärdet en väntevärdesriktig skattning av μ och stickprovsvariansen en väntevärdesriktig skattning av σ^2 .

Intervallskattningar

Samtliga intervall nedan är tvåsidiga med $100(1 - \alpha)\%$ konfidensgrad

- μ då $X_i \sim N(\mu, \sigma^2)$ och σ är känd

$$I_\mu = \left(\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

- μ då $X_i \sim N(\mu, \sigma^2)$ och σ är okänd

$$I_\mu = \left(\bar{x} \pm t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right)$$

- σ^2 då $X_i \sim N(\mu, \sigma^2)$ och μ är okänd

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} \right)$$

- $\mu_1 - \mu_2$ då $X_i \sim N(\mu_1, \sigma_1^2), i = 1, \dots, n_1$ och $Y_i \sim N(\mu_2, \sigma_2^2), i = 1, \dots, n_2$

* då σ_1 och σ_2 är kända

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

* då $\sigma_1 = \sigma_2 = \sigma$ där σ är okänd

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} \pm t_{\alpha/2}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right), \text{ där } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

* då $\sigma_1 \neq \sigma_2$ är okända (approximativt)

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} \pm t_{\alpha/2}(f) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right), \text{ där } f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- σ_1^2/σ_2^2 då $X_i \sim N(\mu_1, \sigma_1^2), i = 1, \dots, n_1$ och $Y_i \sim N(\mu_2, \sigma_2^2), i = 1, \dots, n_2$ där μ_1 och μ_2 okända

$$I_{\sigma_1^2/\sigma_2^2} = \left(\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right)$$

- Δ då $Z_i = X_i - Y_i \sim \mathbf{N}(\Delta, \sigma^2), i = 1, \dots, n$ där σ okänd (stickprov i par)

$$I_{\Delta} = \left(\bar{z} \pm t_{\alpha/2}(n-1) \frac{s_z}{\sqrt{n}} \right)$$

- p då $X \sim \text{Bin}(n, p)$ (approximativt då $np(1-p) \geq 10$)

$$I_p = \left(p^* \pm z_{\alpha/2} \sqrt{\frac{p^*(1-p^*)}{n}} \right), \text{ där } p^* = \frac{x}{n}$$

- $p_1 - p_2$ då $X_1 \sim \text{Bin}(n_1, p_1)$ och $X_2 \sim \text{Bin}(n_2, p_2)$ (approximativt då $n_i p_i(1-p_i) \geq 10$)

$$I_{p_1-p_2} = \left(p_1^* - p_2^* \pm z_{\alpha/2} \sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}} \right), \text{ där } p_i^* = \frac{x_i}{n_i}$$

- μ då $X \sim \text{Po}(\mu)$ (approximativt då $\mu \geq 15$)

$$I_{\mu} = \left(x \pm z_{\alpha/2} \sqrt{x} \right)$$

Regression

Modellen är $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$ där $\varepsilon_i \sim \mathbf{N}(0, \sigma^2)$ är oberoende.

- Minsta-kvadratskattningar

$$\beta_1^* = \frac{S_{xy}}{S_{xx}} \sim \mathbf{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\beta_0^* = \bar{y} - \beta_1^* \bar{x} \sim \mathbf{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$s^2 = \frac{Q_0}{n-2} \text{ där } Q_0 = \sum_{i=1}^n (y_i - \beta_0^* - \beta_1^* x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Tvåsidigt konfidensintervall för $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$

$$I_{\mu_Y(x_0)} = \left(\beta_0^* + \beta_1^* x_0 \pm t_{\alpha/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

- Tvåsidigt prediktionsintervall för $Y(x_0)$

$$I_{Y(x_0)} = \left(\beta_0^* + \beta_1^* x_0 \pm t_{\alpha/2}(n-2) s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$