
FORMELSAMLING HT-18
MATEMATISK STATISTIK FÖR B, K, N, BME OCH KEMISTER; FMSF70 & MASB02

Sannolighetsteori

- Följande gäller för sannolikheter:

- * $0 \leq \mathbf{P}(A) \leq 1$

- * $\mathbf{P}(\Omega) = 1$

- * $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$, om händelserna A och B är oförenliga (disjunkta).

- Additionsatsen för två händelser: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

- Betingad sannolikhet: $\mathbf{P}(B | A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)}$.

- "Satsen om total sannolikhet": $\mathbf{P}(A) = \sum_{i=1}^n \mathbf{P}(A | H_i) \mathbf{P}(H_i)$,

där händelserna H_1, \dots, H_n är parvis oförenliga (disjunkta) händelser och $\bigcup_{i=1}^n H_i = \Omega$.

- A och B är oberoende $\iff \mathbf{P}(A \cap B) = \mathbf{P}(A) \mathbf{P}(B)$.

Beskrivning av data

- Medelvärde: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

- Varians: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 \right]$

- Variationskoefficient: $\frac{s}{\bar{x}}$

- Kovarians: $c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right]$

- Korrelationskoefficient: $r_{xy} = \frac{c_{xy}}{s_x s_y}$

Läges-, spridnings- och beroendemått

- Väntevärdet av $g(X)$:

$$\mathbf{E}[g(X)] = \begin{cases} \sum_{k=-\infty}^{\infty} g(k) p_X(k) & \text{(diskreta s.v.)} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{(kontinuerliga s.v.)} \end{cases}$$

- Varians: $\mathbf{V}(X) = \mathbf{E}[(X - \mathbf{E}(X))^2] = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2$.
- Standardavvikelse: $\mathbf{D}(X) = \sqrt{\mathbf{V}(X)}$.
- Kovarians: $\mathbf{C}(X, Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))] = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$.
- Väntevärde av linjärkombination: $\mathbf{E}\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i \mathbf{E}(X_i) + b$
- Varians av linjärkombination: $\mathbf{V}\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i^2 \mathbf{V}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \mathbf{C}(X_i, X_j)$.
- X_1, \dots, X_n oberoende $\Rightarrow X_1, \dots, X_n$ okorrelerade, dvs $\mathbf{C}(X_i, X_j) = 0, i \neq j$.

Fördelningar

Vanliga fördelningar

Fördelning			Väntevärde	Varians
Binomialfördelning, $Bin(n, p)$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, 1, \dots, n$	np	$np(1-p)$
Poissonfördelning, $Po(\mu)$	$p(k) = e^{-\mu} \frac{\mu^k}{k!}$	$k = 0, 1, 2, \dots$	μ	μ
Rektangel- fördelning, $R(a, b)$	$f(x) = \frac{1}{b-a}$	$a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
Exponential- fördelning, $Exp(a)$	$f(x) = \frac{1}{a} e^{-x/a}$	$x \geq 0$	a	a^2
Normalfördelning ¹ , $N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$-\infty < x < \infty$	μ	σ^2
χ^2 -fördelning, $\chi^2(n)$	$f(x) = \frac{1}{2} \frac{e^{-x/2} (x/2)^{n/2-1}}{\Gamma(n/2)}$	$x \geq 0$	n	$2n$
t -fördelning, $t(n)$	$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	$-\infty < x < \infty$	$0, n > 1$	$\frac{n}{n-2}, n > 2$
F-fördelning, $F(n, m)$	$f(x) = \frac{\Gamma(\frac{n+m}{2}) n^{n/2} m^{m/2}}{\Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \cdot \frac{x^{(n-2)/2}}{(m+nx)^{(n+m)/2}}$	$x \geq 0$	$\frac{m}{m-2}$	$\frac{m^2(2m+2n-4)}{n(m-2)^2(m-4)}, m > 4$

¹I Olbjer och extentor $N(\mu, \sigma^2)$; i "Räkna med Variation" och MATLAB $N(\mu, \sigma)$.

Additionsformler

Om X och Y oberoende så gäller:

$$X \in \text{Bin}(n_1, p), Y \in \text{Bin}(n_2, p) \Rightarrow X + Y \in \text{Bin}(n_1 + n_2, p).$$

$$X \in \text{Po}(\mu_1), Y \in \text{Po}(\mu_2) \Rightarrow X + Y \in \text{Po}(\mu_1 + \mu_2).$$

$$X \in \chi^2(n), Y \in \chi^2(m) \Rightarrow X + Y \in \chi^2(n + m).$$

Normalfördelning

- $X \in N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \in N(0, 1)$
- $F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$ där $\Phi(\cdot)$ ges av tabell
- X_1, \dots, X_n oberoende och $N(\mu_1, \sigma_1^2), \dots, N(\mu_n, \sigma_n^2) \Rightarrow$
$$\sum_{i=1}^n a_i X_i \in N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Centrala gränsvärdesatsen

- X_1, X_2, \dots oberoende och likafördelade med $\mathbf{E}(X_i) = \mu, \mathbf{V}(X_i) = \sigma^2 \Rightarrow$
$$\sum_{i=1}^n X_i \in N(n\mu, n\sigma^2)$$
 om n är stort nog
- Med utnyttjande av, bland annat, CGS gäller följande approximationer:
$$\begin{array}{lll} \text{Bin}(n, p) & \rightarrow & \text{Po}(np) & \text{om } p \leq 0.1 \text{ och } n \geq 10. \\ \text{Bin}(n, p) & \rightarrow & N(np, np(1-p)) & \text{om } np(1-p) \geq 10. \\ \text{Po}(\mu) & \rightarrow & N(\mu, \mu) & \text{om } \mu \geq 15. \end{array}$$

Gauss approximationsformler:

Med $\mu = \mathbf{E}(X)$ gäller att

$$\mathbf{E}[g(X)] \approx g(\mu),$$

$$\mathbf{V}[g(X)] \approx [g'(\mu)]^2 \cdot \mathbf{V}(X).$$

Med $\mu_i = \mathbf{E}(X_i)$ och $c_i = g'_i(\mu_1, \dots, \mu_k)$ gäller att

$$\mathbf{E}[g(X_1, \dots, X_n)] \approx g(\mu_1, \dots, \mu_k),$$

$$\mathbf{V}[g(X_1, \dots, X_n)] \approx \sum_{i=1}^n c_i^2 \mathbf{V}(X_i) + 2 \sum_{i=1}^k \sum_{j=i+1}^k c_i c_j \mathbf{C}(X_i, X_j).$$

Obs: X_1, \dots, X_n oberoende $\Rightarrow X_1, \dots, X_n$ okorrelerade, dvs $\mathbf{C}(X_i, X_j) = 0, i \neq j$.

Fördelningar besläktade med normalfördelningar

- X_1, \dots, X_n oberoende och $N(0, 1) \Rightarrow \sum_{i=1}^n X_i^2 \in \chi^2(n)$
- X_1, \dots, X_n oberoende och $N(\mu, \sigma^2) \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \in \chi^2(n-1)$
- $X \in N(0, 1), Y \in \chi^2(n)$ samt oberoende $\Rightarrow \frac{X}{\sqrt{Y/n}} \in t(n)$
- $X \in \chi^2(n), Y \in \chi^2(m)$ samt oberoende $\Rightarrow \frac{X/n}{Y/m} \in F(n, m)$
- $F_{1-\alpha}(n, m) = 1/F_\alpha(m, n)$

Konfidensintervall

- Konfidensintervall med konfidensgrad $1 - \alpha$ för väntevärdet av en normalfördelad skattning:

Om $\vartheta^* \in N(\vartheta, \mathbf{D}(\vartheta^*))$ så

$$I_\vartheta = (\vartheta^* \pm \lambda_{\alpha/2} \cdot \mathbf{D}(\vartheta^*)), \quad \text{om } \mathbf{D}(\vartheta^*) \text{ är känd}$$

$$I_\vartheta = (\vartheta^* \pm \lambda_{\alpha/2} \cdot \mathbf{d}(\vartheta^*)), \quad \begin{array}{l} \text{om } \mathbf{D}(\vartheta^*) \text{ skattas med } \mathbf{d}(\vartheta^*), \\ \text{eller } \vartheta^* \lesssim N \text{ enl. CGS.} \end{array}$$

$$I_\vartheta = (\vartheta^* \pm t_{\alpha/2}(f) \cdot \mathbf{d}(\vartheta^*)), \quad \begin{array}{l} \text{om } \mathbf{D}(\vartheta^*) = c \cdot \sigma \text{ där } \sigma \text{ okänd och skattad med} \\ (\sigma^2)^* = s^2 = \frac{Q}{f} \text{ med } \frac{Q}{\sigma^2} \in \chi^2(f) \end{array}$$

Intervallen är approximativa vid normalapproximation av skattaren, $\vartheta^* \lesssim N(\vartheta, \mathbf{D}(\vartheta^*))$.

- Konfidensintervall med konfidensgrad $1 - \alpha$ för variansen i en normalfördelning:

Om $X_1, \dots, X_n \in N(\mu, \sigma^2)$ med $(\sigma^2)^* = s^2 = \frac{Q}{f}$ och $\frac{Q}{\sigma^2} \in \chi^2(f)$ så

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

- Konfidensintervall med konfidensgrad $1 - \alpha$ för kvoten mellan varianserna i två normalfördelningar:

Om $X_1, \dots, X_{n_1} \in N(\mu_1, \sigma_1^2)$ och $Y_1, \dots, Y_{n_2} \in N(\mu_2, \sigma_2^2)$ och μ_1, μ_2 är okända:

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

Skattning av σ^2

- Om $X_i \in N(\mu, \sigma^2)$, $i = 1, \dots, n$ är oberoende och μ okänd skattas variansen med

$$(\sigma^2)^* = s^2 = \frac{Q}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{och} \quad \frac{Q}{\sigma^2} \in \chi^2(n-1)$$

- Poolade variansskattning vid k stickprov:

$$(\sigma^2)^* = s_p^2 = \frac{Q}{f} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_k - 1)} \quad \text{och} \quad \frac{Q}{\sigma^2} \in \chi^2(f)$$

med $f = \sum n_i - k$ frihetsgrader.

Vanliga medelfel

Modell	Skattning	Medelfel
$X_i \in N(\mu, \sigma^2)$, $i = 1, \dots, n$	$\mu^* = \bar{x}$	$D(\mu^*) = \frac{\sigma}{\sqrt{n}}$
$X_i \in N(\mu_1, \sigma^2)$, $i = 1, \dots, n_1$ $Y_j \in N(\mu_2, \sigma^2)$, $j = 1, \dots, n_2$	$\mu_1^* = \bar{x}$ $\mu_2^* = \bar{y}$	$D(\mu_1^* - \mu_2^*) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$X \in \text{Bin}(n, p)$	$p^* = \frac{x}{n}$	$d(p^*) = \sqrt{\frac{p^*(1-p^*)}{n}}$
$X_1 \in \text{Bin}(n_1, p_1)$ $X_2 \in \text{Bin}(n_2, p_2)$	$p_i^* = \frac{x_i}{n_i}$	$d(p_1^* - p_2^*) = \sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}}$
$X \in \text{Po}(\mu)$	$\mu^* = x$	$d(\mu^*) = \sqrt{x}$

Hypotestest

- Direktmetoden: \mathbf{P} (Få det vi fått eller längre från $H_0 \parallel H_0$ sann), jmf. med signifikansnivån α .
- Teststorhet, om skattningen ϑ^* är (approximativt) normalfördelad,

$$T = \frac{\vartheta^* - \vartheta_0}{\mathbf{d}_{H_0}(\vartheta^*)}, \quad \text{jmf. med } \lambda \text{ eller } t(f)\text{-kvantil.}$$

- Styrkefunktion: $h(\vartheta) = \mathbf{P}(H_0 \text{ förkastas} \parallel \vartheta \text{ är det rätta parametervärdet})$
- Speciellt: Signifikansnivån, $\alpha = \mathbf{P}(H_0 \text{ förkastas} \parallel H_0 \text{ sann})$

Regression

Enkel linjär regression:

- Modell: $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \dots, n$, där $\varepsilon_i \in N(0, \sigma^2)$ är oberoende.
- Parameterskattningar:

$$\beta^* = \frac{S_{xy}}{S_{xx}} \in N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \quad \alpha^* = \bar{y} - \beta^* \bar{x} \in N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$
$$s^2 = \frac{Q_0}{n-2} \quad \text{där} \quad Q_0 = \sum_{i=1}^n (y_i - \alpha^* - \beta^* x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$
$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Ett tvåsidigt konfidensintervall med konfidensgrad $1 - p$ för $\mu_Y(x_0) = \alpha + \beta x_0$ ges av

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

- Ett prediktionsintervall för $y(x_0) = \alpha + \beta x_0 + \varepsilon_0$ ges av

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

- Ett kalibreringsintervall med konfidensgrad $1 - p$ för $x_0 = \frac{y_0 - \alpha}{\beta}$ ges av

$$I_{x_0} = \left(x_0^* \pm t_{p/2}(n-2) \cdot \frac{s}{|\beta^*|} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0^* - \bar{x})^2}{S_{xx}}} \right) \quad \text{där} \quad x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$$

Multipl linjär regression:

- Modell: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, där $\varepsilon_i \in N(0, \sigma^2)$ är oberoende.
- Med matrisrepresentation kan modellen skrivas som $Y = X\beta + E$.
- Parameterskattningar:

$$\beta^* = (X^T X)^{-1} X^T Y \quad \mathbf{V}(\beta^*) = \sigma^2 (X^T X)^{-1}$$
$$s^2 = \frac{Q_0}{n - (p + 1)} \quad \text{där} \quad Q_0 = \sum_{i=1}^n (y_i - \beta_0^* - \beta_1^* x_{1i} - \dots - \beta_p^* x_{pi})^2 = Y^T Y - \beta^{*T} X^T Y$$

- Konfidensintervall för β_j :

$$I_{\beta_j} = (\beta_j^* \pm t_{\alpha/2}(n-p-1) \cdot \mathbf{d}(\beta_j^*)) \quad \text{där} \quad \mathbf{d}(\beta_j^*) = s \sqrt{\text{element}(j+1, j+1) \text{ i } (X^T X)^{-1}}$$

- Konfidensintervall för $\mu_Y(\mathbf{x}^0) = \beta_0 + \beta_1 \mathbf{x}_1^0 + \dots + \beta_p \mathbf{x}_p^0$:

$$I_{\mu_Y(\mathbf{x}^0)} = \left(\mu_Y^*(\mathbf{x}^0) \pm t_{\alpha/2}(n-p-1) \cdot s \sqrt{\mathbf{x}^0 T (X^T X)^{-1} \mathbf{x}^0} \right)$$

Faktor försök

2^k -försök

Varje faktor kan anta låg (-) och hög (+) nivå. För t.ex. ett 2^3 -försök med n observationer per faktorkombination är modellen

$$y_{ijkl} = \mu \pm A \pm B \pm C(\pm)(\pm)AB(\pm)(\pm)AC(\pm)(\pm)BC(\pm)(\pm)(\pm)ABC + \varepsilon_{ijkl}$$

Effekten skattas med hjälp av ett teckenschema. Dividera med 2^3 (allmänt med 2^k)

Förs	Medelv	μ	A	B	C	AB	AC	BC	ABC
(1)	\bar{y}_{----}	+	-	-	-	+	+	+	-
(a)	\bar{y}_{+--}	+	+	-	-	-	-	+	+
(b)	\bar{y}_{-+-}	+	-	+	-	-	+	-	+
(ab)	\bar{y}_{++-}	+	+	+	-	+	-	-	-
(c)	\bar{y}_{--+}	+	-	-	+	+	-	-	+
(ac)	\bar{y}_{++-}	+	+	-	+	-	+	-	-
(bc)	\bar{y}_{-++}	+	-	+	+	-	-	+	-
(abc)	\bar{y}_{+++}	+	+	+	+	+	+	+	+

Medelfelet $\mathbf{d}(\text{effekt}) = \frac{s}{\sqrt{2^k n}}$, där s^2 är den poolade variansskattningen från de olika försökspunkterna om $n \geq 2$.

Om $n = 1$ kan en variansskattning erhållas från samspel av högre ordning. För dessa måste då antas $E((\text{effekt})^2) = \sigma^2 / 2^k$.

2^{k-1} -försök

Vanligen kopplas högsta samspelet till I . För $k = 4$, t.ex., blir kopplingen $I = \pm ABCD$.

Härur erhålles kopplingar mellan övriga effekter. Försökspunkterna fås genom att i teckenschemat för 2^k -försöket välja de rader som antingen har + eller - för högsta samspelet. Effekterna skattas med hjälp av det så erhållna halverade teckenschemat. Dividera med 2^{k-1} . Medelfelet $\mathbf{d}(\text{effekt}) = \frac{s}{\sqrt{2^{k-1}}}$.