FORMELSAMLING HT-17
MATEMATISK STATISTIK FÖR B, K, N, BME OCH KEMISTER; FMSF70 & MASB02

Sannolikhetsteori

- Följande gäller för sannolikheter:
  - $0 \leq \mathsf{P}(\mathrm{A}) \leq 1$
  - $\mathsf{P}(\Omega) = 1$
  - $\mathsf{P}(\mathrm{A} \cup \mathrm{B}) = \mathsf{P}(\mathrm{A}) + \mathsf{P}(\mathrm{B})$, om händelserna A och B är oförenliga (disjunkta).
- Additionssatsen för två händelser: $\mathsf{P}(\mathrm{A} \cup \mathrm{B}) = \mathsf{P}(\mathrm{A}) + \mathsf{P}(\mathrm{B}) - \mathsf{P}(\mathrm{A} \cap \mathrm{B})$.
- Betingad sannolikhet: $\mathsf{P}(\mathrm{B} \mid \mathrm{A}) = \dfrac{\mathsf{P}(\mathrm{A} \cap \mathrm{B})}{\mathsf{P}(\mathrm{A})}$.
- "Satsen om total sannolikhet": $\mathsf{P}(\mathrm{A}) = \sum_{i=1}^{n} \mathsf{P}(\mathrm{A} \mid \mathrm{H}_i)\,\mathsf{P}(\mathrm{H}_i)$,

  där händelserna $\mathrm{H}_1, \ldots, \mathrm{H}_n$ är parvis oförenliga (disjunkta) händelser och $\bigcup_{i=1}^{n} \mathrm{H}_i = \Omega$.
- A och B är oberoende $\iff \mathsf{P}(\mathrm{A} \cap \mathrm{B}) = \mathsf{P}(\mathrm{A})\,\mathsf{P}(\mathrm{B})$.

Beskrivning av data

- Medelvärde: $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$
- Varians: $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \dfrac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right]$
- Variationskoefficient: $\dfrac{s}{\bar{x}}$
- Kovarians: $c_{xy} = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{1}{n-1} \left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right]$
- Korrelationskoefficient: $r_{xy} = \dfrac{c_{xy}}{s_x s_y}$

Läges-, spridnings- och beroendemått

- Väntevärdet av $g(X)$:

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

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

Fördelningar

Vanliga fördelningar

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

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

Additionsformler

Om X och Y oberoende så gäller:

$$X \in Bin(n_1,\, p)\,,\ Y \in Bin(n_2,\, p) \quad \Rightarrow \quad X + Y \in Bin(n_1 + n_2,\, p)\,.$$

$$X \in Po(\mu_1)\,,\ Y \in Po(\mu_2) \quad \Rightarrow \quad X + Y \in Po(\mu_1 + \mu_2)\,.$$

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

Normalfördelning

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

Centrala gränsvärdessatsen

- $X_1, X_2, \ldots$ oberoende och likafördelade med $\mathsf{E}(X_i) = \mu$, $\mathsf{V}(X_i) = \sigma^2 \Rightarrow$
  $\displaystyle\sum_{i=1}^{n} X_i \underset{\sim}{\in} N(n\mu,\, n\sigma^2)$ om n är stort nog
- Med utnyttjande av, bland annat, CGS gäller följande approximationer:

$Bin(n,\, p)$	$\to$	$Po(np)$	om	$p \le 0.1$ och $n \ge 10$.
$Bin(n,\, p)$	$\to$	$N(np,\, np(1-p))$	om	$np(1-p) \ge 10$.
$Po(\mu)$	$\to$	$N(\mu,\, \mu)$	om	$\mu \ge 15$.

Gauss approximationsformler:

Med $\mu = \mathsf{E}(X)$ gäller att

$$\mathsf{E}\left[g(X)\right] \approx g(\mu),$$
$$\mathsf{V}\left[g(X)\right] \approx \left[g'(\mu)\right]^2 \cdot \mathsf{V}(X).$$

Med $\mu_i = \mathsf{E}(X_i)$ och $c_i = g_i'(\mu_1, \ldots, \mu_k)$ gäller att

$$\mathsf{E}\left[g(X_1, \ldots, X_n)\right] \approx g(\mu_1, \ldots, \mu_k),$$
$$\mathsf{V}\left[g(X_1, \ldots, X_n)\right] \approx \sum_{i=1}^{n} c_i^2\, \mathsf{V}(X_i) + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} c_i\, c_j\, \mathsf{C}(X_i,\, X_j).$$

Obs: $X_1, \ldots, X_n$ oberoende $\Rightarrow X_1, \ldots, X_n$ okorrelerade, dvs $\mathsf{C}(X_i, X_j) = 0,\ i \ne j$.

Fördelningar besläktade med normalfördelningar

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

Konfidensintervall

- Konfidensintervall med konfidensgrad $1 - \alpha$ för väntevärdet av en normalfördelad skattning:
  Om $\vartheta^* \in N\big(\vartheta,\ \mathsf{D}(\vartheta^*)^2\big)$ så

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

  $I_\vartheta = (\vartheta^* \pm \lambda_{\alpha/2} \cdot \mathsf{d}(\vartheta^*))$, om $\mathsf{D}(\vartheta^*)$ skattas med $\mathsf{d}(\vartheta^*)$, eller $\vartheta^* \underset{\sim}{\in} N$ enl. CGS.

  $I_\vartheta = (\vartheta^* \pm t_{\alpha/2}(f) \cdot \mathsf{d}(\vartheta^*))$, om $\mathsf{D}(\vartheta^*) = c \cdot \sigma$ där σ okänd och skattad med $(\sigma^2)^* = s^2 = \dfrac{Q}{f}$ med $\dfrac{Q}{\sigma^2} \in \chi^2(f)$

  Intervallen är approximativa vid normalapproximation av skattaren, $\vartheta^* \underset{\sim}{\in} N\big(\vartheta,\ \mathsf{D}(\vartheta^*)^2\big)$.

- Konfidensintervall med konfidensgrad $1 - \alpha$ för variansen i en normalfördelning:
  Om $X_1, \ldots, X_n \in N\big(\mu,\ \sigma^2\big)$ med $(\sigma^2)^* = s^2 = \dfrac{Q}{f}$ och $\dfrac{Q}{\sigma^2} \in \chi^2(f)$ så

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

- Konfidensintervall med konfidensgrad $1 - \alpha$ för kvoten mellan varianserna i två normalfördelningar:
  Om $X_1, \ldots, X_{n_1} \in N\big(\mu_1,\ \sigma_1^2\big)$ och $Y_1, \ldots, Y_{n_2} \in N\big(\mu_2,\ \sigma_2^2\big)$ 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 $X_i \in N(\mu,\ \sigma^2)$, $i = 1, \ldots, n$ är oberoende och μ okänd skattas variansen med

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

- Poolade variansskattning vid k stickprov:

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

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

Vanliga medelfel

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

Hypotestest

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

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

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

Regression

Enkel linjär regression:

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

$$\beta^* = \frac{S_{xy}}{S_{xx}} \in N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \qquad \alpha^* = \bar{y} - \beta^*\bar{x} \in N\left(\alpha,\ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

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

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

- 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 $x_0 = \dfrac{y_0 - \alpha}{\beta}$ ges av

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

Multipel linjär regression:

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

$$\beta^* = (X^T X)^{-1} X^T Y \qquad \mathsf{V}(\beta^*) = \sigma^2 (X^T X)^{-1}$$

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

- Konfidensintervall för β_i:

$$I_{\beta_i} = \big(\beta_i^* \pm t_{\alpha/2}\,\big(n - p - 1\big) \cdot \mathsf{d}(\beta_i^*)\big) \qquad \text{där} \qquad \mathsf{d}(\beta_i^*) = s\sqrt{\text{element}(i+1, i+1) \text{ i } (X^T X)^{-1}}$$

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

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

Faktorförsök

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 $\mathsf{d}(\text{effekt}) = \frac{s}{\sqrt{2^k n}}$, 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 $\mathsf{E}((\text{effekt})^2) = \sigma^2/2^k$.

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 $\mathsf{d}(\text{effekt}) = \frac{s}{\sqrt{2^{k-1}}}$.