Sannolikhetsteori

- Satsen om total sannolikhet: $\mathsf{P}(\mathrm{A}) = \sum_{i=1}^{n} \mathsf{P}(\mathrm{A} \mid \mathrm{H}_i)\, \mathsf{P}(\mathrm{H}_i)$ om $\mathrm{H}_i \cap \mathrm{H}_j = \emptyset$ då $i \neq j$ och $\bigcup_{i=1}^{n} \mathrm{H}_i = \Omega$.
- A och B är oberoende $\iff \mathsf{P}(\mathrm{A} \cap \mathrm{B}) = \mathsf{P}(\mathrm{A})\, \mathsf{P}(\mathrm{B})$.

Flerdimensionella stokastiska variabler

- Marginell täthetsfunktion för X: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$.

Summor av stokastiska variabler

- Om X och Y är oberoende, så gäller för $Z = X + Y$

$$p_Z(k) = \sum_{i=-\infty}^{\infty} p_X(i)\, p_Y(k-i),$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx.$$

- Om X_1 och X_2 är oberoende s.v. så gäller:

$$\begin{cases} X_1 \in Bin(n_1,\, p) \\ X_2 \in Bin(n_2,\, p) \end{cases} \Rightarrow \quad X_1 + X_2 \in Bin(n_1 + n_2,\, p)\,.$$

$$\begin{cases} X_1 \in Po(\mu_1) \\ X_2 \in Po(\mu_2) \end{cases} \Rightarrow \quad X_1 + X_2 \in Po(\mu_1 + \mu_2)\,.$$

$$\begin{cases} X_1 \in \Gamma(p_1, \lambda) \\ X_2 \in \Gamma(p_2, \lambda) \end{cases} \Rightarrow \quad X_1 + X_2 \in \Gamma(p_1 + p_2, \lambda).$$

$$\begin{cases} X_1 \in \chi^2(f_1) \\ X_2 \in \chi^2(f_2) \end{cases} \Rightarrow \quad X_1 + X_2 \in \chi^2(f_1 + f_2).$$

- Om $X_1 \in N(\mu_1,\, \sigma_1)$ och $X_2 \in N(\mu_2,\, \sigma_2)$ är oberoende och a och b konstanter är:

$$aX_1 + bX_2 \in N\left(a\mu_1 + b\mu_2,\, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$$

Väntevärde, Varians och Kovarians

- Varians: $\mathsf{V}(X) = \mathsf{E}\Big(\big(X - \mathsf{E}(X)\big)^2\Big) = \mathsf{E}(X^2) - \big(\mathsf{E}(X)\big)^2$.
- Kovarians: $\mathsf{C}(X, Y) = \mathsf{E}\Big(\big(X - \mathsf{E}(X)\big)\big(Y - \mathsf{E}(Y)\big)\Big) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$.
- X, Y oberoende $\Rightarrow X, Y$ okorrelerade, dvs $\mathsf{C}(X, Y) = 0$.
- Väntevärde och varians av linjärkombination:

$$\mathsf{E}\left(\sum_i a_i X_i + b\right) = \sum_i a_i \mathsf{E}(X_i) + b$$

$$\mathsf{V}\left(\sum_i a_i X_i + b\right) = \sum_i a_i^2 \mathsf{V}(X_i) + 2\sum_{i<j} a_i a_j \mathsf{C}(X_i, X_j).$$

Centrala gränsvärdessatsen

Om $X_1, X_2, \ldots$ är oberoende och likafördelade gäller approximationen $\sum_{i=1}^{n} X_i \underset{\sim}{\in} N$ om n är tillräckligt stort.

Följande approximationer gäller:

Binomial	$\rightarrow$	Poisson	om $p \le 0.1$ och $n \ge 10$.
Binomial	$\rightarrow$	Normal	om $np(1-p) \ge 10$.
Poisson	$\rightarrow$	Normal	om $\mu \ge 15$.

Gauss approximationsformler:

Med $\mu = \mathsf{E}(X)$ gäller att

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

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

$$\mathsf{E}\left[g(X_1, \ldots, X_n)\right] \approx g(\mu_1, \ldots, \mu_k),$$

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

Punktskattningar

Ett stickprov

Låt $x_1, \ldots, x_n$ vara observationer av oberoende och likafördelade s.v. $X_1, \ldots, X_n$ med (okänt) väntevärde μ och standardavvikelse σ. Väntevärdesriktiga skattningar av μ och σ^2 är då

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$(\sigma^2)^* = s^2 = \frac{Q}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Korrelation

Låt $x_1, \ldots, x_n$ och $y_1, \ldots, y_n$ vara två uppsättningar observationer, korrelationen mellan X och Y skattas med

$$\rho^* = r_{xy} = \frac{c_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Om x_i och y_i kommer från en bivariat normalfördelning med $\rho = 0$ så gäller följande:

$$t = r_{xy}\sqrt{\frac{n-2}{1 - r_{xy}^2}} \in t(n-2) \qquad \text{om } \rho = 0$$

Vanliga skattningsmetoder

Låt $x_1, \dots, x_n$ vara observationer av oberoende stokastiska variabler med täthets- (sannolikhets-)funktioner $f(x_i;\theta)$ $(p(x_i;\theta))$ eller väntevärde $\mathsf{E}(X_i) = \mu_i(\theta)$

- ML-skattning, θ^*_{ML}, är det θ-värde som maximerar

$$L(\theta;\, x_1,\dots,x_n) = \begin{cases} p(x_1;\theta)\cdot p(x_2;\theta)\cdot\dots\cdot p(x_n;\theta), \\ f(x_1;\theta)\cdot f(x_2;\theta)\cdot\dots\cdot f(x_n;\theta). \end{cases}$$

- MK-Skattningen, θ^*_{MK}, är det θ-värde som minimerar

$$Q(\theta;\, x_1,\dots,x_n) = \sum_{i=1}^{n}(x_i - \mu_i(\theta))^2.$$

Konfidensintervall för Normalfördelning

Om $\theta^* \in N(\theta,\, \mathsf{D}(\theta^*))$ så ges ett konfidensintervall med konfidensgrad $1-\alpha$ för väntevärdet av:

$I_\theta = (\theta^* \pm \lambda_{\alpha/2}\cdot \mathsf{D}(\theta^*))$,	om $\mathsf{D}(\theta^*)$ är känd
$I_\theta = (\theta^* \pm \lambda_{\alpha/2}\cdot \mathsf{d}(\theta^*))$,	om $\mathsf{D}(\theta^*)$ skattas med $\mathsf{d}(\theta^*)$, eller $\theta^* \underset{\sim}{\in} N$ enl. CGS.
$I_\theta = (\theta^* \pm t_{\alpha/2}(f)\cdot \mathsf{d}(\theta^*))$,	om $\mathsf{D}(\theta^*) = c\cdot\sigma$ där σ okänd och skattad med $s^2 = Q/f$.

Intervallen är approximativa om skattningen är approximativt normalfördelad (enligt CGS), $\theta^* \underset{\sim}{\in} N(\theta,\, \mathsf{D}(\theta^*))$.

Hypotestest

Direktmetoden: Jmf. p-värdet med signifikansnivån α:
p-värde $= \mathsf{P}\big(\text{Få det vi fått eller längre från } H_0 \,\|\, H_0 \text{ sann}\big)$,

Konfidensintervall: Konstruera ett lämpligt intervall (under H_0) och jmf. med θ_0.

Teststorhet: Om skattningen θ^* är (approx.) normalfördelad, $T = \dfrac{\theta^* - \theta_0}{\mathsf{d}_{H_0}(\theta^*)}$, jmf. med λ eller $t(f)$-kvantil.

Styrkefunktion: $h(\theta) = \mathsf{P}(H_0 \text{ förkastas} \,\|\, \theta \text{ är rätt värde})$

Signifikansnivå: $\alpha = \mathsf{P}(H_0 \text{ förkastas} \,\|\, H_0 \text{ sann})$

Regression

Enkel linjär regression

- Modell: $y_i = \alpha + \beta x_i + \varepsilon_i$, $\quad \varepsilon_i \in N(0,\, \sigma)$.
- Parameterskattningar:

$$\alpha^* = \bar{y} - \beta^*\bar{x}, \qquad \beta^* = \frac{S_{xy}}{S_{xx}},$$

$$(\sigma^2)^* = s^2 = \frac{Q_0}{n-2}, \qquad Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

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

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

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

$$\mu_0^* = \alpha^* + \beta^* x_0 \in N\left(\alpha + \beta x_0,\, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right).$$

$$y^* = \alpha^* + \beta^* x_0 + \varepsilon_0 \in N\left(\alpha + \beta x_0,\, \sigma\sqrt{1 + \tfrac{1}{n} + \tfrac{(x_0 - \bar{x})^2}{S_{xx}}}\right).$$

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

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

där $x_0^* = \dfrac{y_0 - \alpha^*}{\beta^*}$

Multipel linjär regression

- Modell: $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_p x_i^{(p)} + \varepsilon_i$, $\varepsilon_i \in N(0,\, \sigma)$.
- På matrisform blir modellen $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$ med

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1^{(1)} & \dots & x_1^{(p)} \\ 1 & x_2^{(1)} & \dots & x_2^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Parameterskattningar:

$$\boldsymbol{\beta}^* = (X^TX)^{-1}X^T\mathbf{y}, \qquad \mathsf{V}(\boldsymbol{\beta}^*) = \sigma^2(X^TX)^{-1},$$

$$(\sigma^2)^* = s^2 = \frac{Q_0}{n-(p+1)}, \qquad Q_0 = \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^{*T}X^T\mathbf{y},$$

$$\beta_i^* \in N\left(\beta_i,\, \sigma\sqrt{\text{element}(ii) \text{ i } (X^TX)^{-1}}\right),$$

$$\mu_0^* = \mathbf{x}_0\boldsymbol{\beta}^* \in N\left(\mu_0,\, \sigma\sqrt{\mathbf{x}_0(X^TX)^{-1}\mathbf{x}_0^T}\right)$$

$$y_0^* = \mathbf{x}_0\boldsymbol{\beta}^* + \varepsilon_0 \in N\left(\mu_0,\, \sigma\sqrt{1 + \mathbf{x}_0(X^TX)^{-1}\mathbf{x}_0^T}\right)$$

där $\mathbf{x}_0 = \begin{pmatrix} 1 & x_0^{(1)} & \dots & x_0^{(p)} \end{pmatrix}$

Tidsserier

Om $y_1, \ldots, y_n$ är en tidsserie är korrelationen

$$\rho_k^* = r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

AR(1)-process: $Y_t = \alpha \cdot Y_{t-1} + e_t$ där $-1 < \alpha < 1$ och slumpvariablerna, e_t, är oberoende med varians σ^2;

$$\rho_k = \alpha^k, \quad k = 0, 1, 2, \ldots$$

Miniräknare

Funktioner för fördelningar finns med ändelserna **cdf/CD** *cumulative distribution function* (fördelningsfunktion, $F_X(x) = P(X \leq x)$, **pdf/PD** *probability density function* (sannolikhets-, $p_X(k) = P(X = k)$, eller täthetsfunktion, $f_X(x)$), eller **inv** *inverse cdf* (dvs $x = F^{-1}(p)$). Notera att kvantil och inverse funktionen skiljer sig ($X > x_\alpha$ eller $X \leq x_p$):
kvantil: Finn x_α så att $\mathsf{P}(X > x_\alpha) = \alpha$.
inverse: $F^{-1}(p) = x_p$ eller finn x_p så att $\mathsf{P}(X \leq x_p) = p$.

Fördelningsfunktionerna beräknar ibland sannolikheter i ett intervall $a < X \leq b$, använd $-1 \cdot 10^{99}$ eller $-1 \cdot 10^{99}$ för värdena $\pm\infty$ (en mindre exponent kan eventuellt behövas men testa att den är "tillräckligt stor", dvs inte ändrar resultatet).

För att beräkna statistik och skattningar (μ^*, s^2, $\sum x_i^2$, etc) behöver värdena först sparas i listor.

	Texas instrument	**Casio**
$1 \cdot 10^x$	**2nd** $\rightarrow$ **EE** Ger 1Ex på räknaren.	**SHIFT** $\rightarrow$ **log** (10^x) Ger $1 * 10^x$ på räknaren.
$k!$ & $\binom{n}{k}$	**MATH** $\rightarrow$ **PROB** $\rightarrow$ **!** **MATH** $\rightarrow$ **PROB** $\rightarrow$ **nCr**	**OPTN** $\rightarrow$ **PROB** $\rightarrow$ **!** **OPTN** $\rightarrow$ **PROB** $\rightarrow$ **nCr**
Fördelningar	**DISTR (2nd VARS)** $\rightarrow$ välj lämplig fördelning **normalpdf(x, μ, σ)** eller **normalcd(a, b, μ, σ)** **binompdf(n, p, x)** eller **binomcdf(n, p, x)** **poissonpdf(λ, x)** eller **poissoncdf(λ, x)** **normalinv(1–α, μ, σ)** för normal-kvantiler, λ_α **tinv(1–α, f)** för t-kvantiler, $t_\alpha(f)$	**OPTN** $\rightarrow$ **STAT** $\rightarrow$ **DISTR** $\rightarrow$ välj lämplig fördelning **Normal PD(x, σ, μ)** eller **Normal CD(a, b, σ, μ)** **Binomial PD(x, n, p)** eller **Binomial CD(x, n, p)** **Poisson PD(x, λ)** eller **Poisson CD(x, λ)** **Inverse Normal** för normal-kvantiler, alternativet ***Tail*** anger vilket håll som inversen ska beräknas (left=inverse eller right=kvantil). **Inverse Student-t** för t-kvantiler.
Listor	**STAT** $\rightarrow$ **EDIT** För att lägga in värden listorna **2nd** $\rightarrow$ **1 – 6** Hämta namn på listor för använding. Det går att räkna med och tilldela listor: Exempel: **2*L1–1 (STO►) L2**	**MENU** $\rightarrow$ **STAT(2)** För att lägga in värden listorna **SHIFT** $\rightarrow$ **1(List)** $\rightarrow$ **1 – 6** Ange namn på listor för använding. Det går att räkna med och tilldela listor: Exempel: **2*List 1–1** $\rightarrow$ **List 2**
Statistik	**STAT** $\rightarrow$ **CALC** $\rightarrow$ **1-Var Stats** Beräkna sammanfatttande statistik för angiven lista. Exempel: **1-Var Stats L1** **STAT** $\rightarrow$ **CALC** $\rightarrow$ **2-Var Stats** Beräkna sammanfatttande statistik för två angivna listor (beräknar t.ex. $\sum x_i y_i$ som behövs för regression och korrelation). Exempel: **2-Var Stats L1,L2** **VARS** $\rightarrow$ **Statistics** $\rightarrow$ Hämta variabler som innehåller statistik som beräknats ovan.	**OPTN** $\rightarrow$ **CALC** $\rightarrow$ **1VAR** Beräkna sammanfatttande statistik för angiven lista. **OPTN** $\rightarrow$ **CALC** $\rightarrow$ **2VAR** Beräkna sammanfatttande statistik för två angivna listor (beräknar t.ex. $\sum x_i y_i$ som behövs för regression och korrelation). **VARS** $\rightarrow$ **STAT** $\rightarrow$ Hämta variabler som innehåller statistik som beräknats ovan.

Använding av miniräknare på tentan: När miniräknaren används för beräkningar på tentan måste det tydligt framgå vad som beräknats (formler), vad resultatet är och vilka slutsatser som kan dras. Att bara repetera resultat från miniräknaren utan kommentarer utgör **INTE** väl motiverade lösningar.

Fördelningar			Väntevärde	Varians
Binomialfördelning, $Bin(n, p)$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, 1, \ldots, n$	np	$np(1-p)$
Hypergeometrisk fördelning	$p(k) = \dfrac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$	$k \le Np$, $n-k \le N(1-p)$	np	$\dfrac{N-n}{N-1}\, np(1-p)$
Poissonfördelning, $Po(\mu)$	$p(k) = \frac{\mu^k}{k!} e^{-\mu}$	$k = 0, 1, 2, \ldots$	μ	μ
Geometrisk fördelning	$p(k) = p(1-p)^k$	$k = 0, 1, 2, \ldots$	$\dfrac{1-p}{p}$	$\dfrac{1-p}{p^2}$
ffg-fördelning	$p(k) = p(1-p)^{k-1}$	$k = 1, 2, 3, \ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Rektangelfördelning, $R(a, b)$, Uab	$f(x) = \dfrac{1}{b-a}$	$a \le x \le b$	$\dfrac{a+b}{2}$	$\dfrac{(a-b)^2}{12}$
Exponential-fördelning, $Exp(\lambda)$, $\Gamma(1, \lambda)$	$f(x) = \lambda e^{-\lambda x}$	$x \ge 0$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
$Exp(\mu)$, $\Gamma(1, \mu)$	$f(x) = \dfrac{1}{\mu} e^{-x/\mu}$	$x \ge 0$	μ	μ^2
Gamma-fördelning[1,2], $\Gamma(p, \lambda)$	$f(x) = \dfrac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x}$	$x \ge 0$	$\dfrac{p}{\lambda}$	$\dfrac{p}{\lambda^2}$
χ^2-fördelning[2], $\chi^2(n)$	$f(x) = \dfrac{1}{2\,\Gamma\left(\frac{n}{2}\right)} \left(\dfrac{x}{2}\right)^{n/2-1} e^{-x/2}$	$x \ge 0$	n	$2n$
Normalfördelning[3], $N(\mu, \sigma)$, $N(\mu, \sigma^2)$	$f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x \in \mathbb{R}$	μ	σ^2
t-fördelning[2], $t(n)$	$f(x) = \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \dfrac{x^2}{n}\right)^{-\frac{n+1}{1}}$	$x \in \mathbb{R}$	$0,\ n > 1$	$\dfrac{n}{n-2},\ n > 2$
Gumbel-fördelning[4]	$F(x) = e^{-e^{-(x-a)/b}}$ (OBS! fördelningsfunktion)	$x \in \mathbb{R}$, $b > 0$	$a + \gamma b$	$\dfrac{b^2\pi^2}{6}$
Weibull-fördelning[2]	$F(x) = 1 - e^{-\left(\frac{x-b}{a}\right)^c}$ (OBS! fördelningsfunktion)	$x \ge b$, $a, c > 0$	$b + a\Gamma\left(1 + \frac{1}{c}\right)$	$a^2\Gamma\left(1 + \frac{2}{c}\right) - a^2\Gamma^2\left(1 + \frac{1}{c}\right)$
Lognormal-fördelning $\ln X \in N(m, \sigma)$	$f(x) = \dfrac{1}{\sqrt{x^2\sigma^2 2\pi}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}$	$x \ge 0$	$e^{m+\sigma^2/2}$	$e^{2m+\sigma^2}\left(e^{\sigma^2} - 1\right)$

1. Gammafördelningen har två parametriseringarna, precis som exponentialfördelningen, för att få $\Gamma(p, \mu)$ sätt $\lambda = 1/\mu$.
2. Gamma-funktionen, $\Gamma(p)$, är en generaliserad fakultet:

$$\Gamma(p) = (p-1) \cdot \Gamma(p-1) \qquad \Gamma(p) = (p-1)! \text{ om } p \text{ heltal}$$

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x}\, dx,\ p > 0 \qquad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}$$

3. Båda beteckningarna förekommer för normalfördelningar, $N(\mu, \sigma)$ är vanligast och används i den här formelsamlingen.
4. Gumbelfördelningen kallas ibland för dubbel exponentialfördelning. Eulers konstant: $\gamma \approx 0.57722$.