Matematisk statistik
KTH

Formel- och tabellsamling i matematisk statistik

Höstterminen 1998

1. Stokastiska variabler

$V(X) = E(X^2) - (E(X))^2$

$C(X,Y) = E\big((X - E(X))(Y - E(Y))\big) = E(XY) - E(X)E(Y)$

$\rho(X,Y) = \dfrac{C(X,Y)}{D(X)D(Y)}$

Gauss' approximationsformler

Med beteckningar och förutsättningar enligt läroboken gäller

a) $E(g(X)) \approx g(E(X))$

$V(g(X)) \approx V(X)[g'(E(X))]^2$

b) $E(g(X_1, \ldots, X_n)) \approx g(m_1, \ldots, m_n)$

$V(g(X_1, \ldots, X_n)) \approx \sum_{i=1}^{n} V(X_i)\Big(\dfrac{\partial g}{\partial m_i}\Big)^2 + 2\sum_{i<j} C(X_i, X_j) \cdot \dfrac{\partial g}{\partial m_i}\dfrac{\partial g}{\partial m_j}$

2. Diskreta fördelningar

Binomialfördelningen

X är $\mathrm{Bin}(n,p)$ om $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$, där $0 < p < 1$

$E(X) = np$, $V(X) = np(1-p)$

"För-första-gången"-fördelningen

X är $\mathrm{ffg}(p)$ om $p_X(k) = p(1-p)^{k-1}$, $k = 1, 2, 3, \ldots$, där $0 < p < 1$

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

Hypergeometriska fördelningen

X är $\mathrm{Hyp}(N,n,p)$ om $p_X(k) = \dfrac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$, där N, Np och n är positiva heltal samt $0 < p < 1$

$E(X) = np$, $V(X) = \dfrac{N-n}{N-1} \cdot np(1-p)$

Poissonfördelningen

X är $\mathrm{Po}(m)$ där $m > 0$ om $p_X(k) = \dfrac{m^k}{k!} \cdot e^{-m}$, $k = 0, 1, 2, \ldots$

$E(X) = m$, $V(X) = m$

3. Kontinuerliga fördelningar

Likformig fördelning (Rektangelfördelning)

X är $R(a,b)$ där $a<b$ om $f_X(x)=\begin{cases}\dfrac{1}{b-a} & \text{för } a\le x\le b\\ 0 & \text{annars}\end{cases}$

$E(X)=\dfrac{a+b}{2}$, $V(X)=\dfrac{(b-a)^2}{12}$

Exponentialfördelningen

X är $\mathrm{Exp}(m)$ där $m>0$ om $f_X(x)=\begin{cases}\dfrac{1}{m}\cdot e^{-x/m} & \text{för } x>0\\ 0 & \text{annars}\end{cases}$

$E(X)=m$, $V(X)=m^2$

Normalfördelningen

X är $N(m,\sigma)$ om $f_X(x)=\dfrac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$, $-\infty<x<\infty$, $\sigma>0$

$E(X)=m$, $V(X)=\sigma^2$

Antag att $Z=\dfrac{X-m}{\sigma}$. Då gäller:
X är $N(m,\sigma)$ om och endast om Z är $N(0,1)$.

Om Z är $N(0,1)$ så har Z fördelningsfunktionen $\Phi(x)$ enligt tabell 1 och täthetsfunktionen $\varphi(x)=\dfrac{1}{\sqrt{2\pi}}\cdot e^{-x^2/2}$, $-\infty<x<\infty$

En linjär sammansättning $\sum a_iX_i$ av oberoende, normalfördelade stokastiska variabler är normalfördelad.

4. Centrala gränsvärdessatsen

Om $X_1,X_2,\ldots,X_n$ är oberoende likafördelade stokastiska variabler med väntevärde m och standardavvikelse $\sigma>0$ så är $Y_n=X_1+\cdots+X_n$ approximativt $N(nm,\sigma\sqrt{n})$ om n är stort.

5. Approximation

$\mathrm{Hyp}(N,n,p)\sim\mathrm{Bin}(n,p)$ om $\dfrac{n}{N}\le 0.1$

$\mathrm{Bin}(n,p)\sim\mathrm{Po}(np)$ om $p\le 0.1$

$\mathrm{Bin}(n,p)\sim N\big(np,\sqrt{np(1-p)}\big)$ om $np(1-p)\ge 10$

$\mathrm{Po}(m)\sim N(m,\sqrt{m})$ om $m\ge 15$

6. Tjebysjovs olikhet

Om $E(X) = m$ och $D(X) = \sigma > 0$ så gäller för varje $k > 0$ att

$$P(|X - m| > k\sigma) \leq \frac{1}{k^2}$$

7. Statistiskt material

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n} \Big(\sum_{j=1}^{n} x_j \Big)^2 \right]$$

$$z_j = (x_j - m_0)/k \Rightarrow \begin{cases} \overline{x} \ = m_0 + k\overline{z} \\ s_x = |k| \cdot s_z \end{cases}$$

8. Punktskattningar

8.1 Maximum-likelihood-metoden

Låt x_i vara en observation på X_i, $i = 1, 2, \ldots, n$, där fördelningen för X_i beror på en okänd parameter θ, och antag att $X_1, X_2, \ldots, X_n$ är oberoende.

Det värde θ^* som maximerar likelihoodfunktionen

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} p_{X_i}(x_i; \theta) \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta) \end{cases}$$

kallas *maximum-likelihood-skattningen (ML-skattningen)* av θ.

8.2 Minsta-kvadrat-metoden

Låt x_i vara en observation på X_i, $i = 1, 2, \ldots, n$, och antag att

$E(X_i) = m_i(\theta_1, \theta_2, \ldots, \theta_k)$ och $V(X_i) = \sigma^2$,

där $\theta_1, \theta_2, \ldots, \theta_k$ är okända parametrar och $X_1, X_2, \ldots, X_n$ är oberoende.

Minsta-kvadrat-skattningarna (MK-skattningarna) av $\theta_1, \theta_2, \ldots, \theta_k$

är de värden $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$ som minimerar kvadratsumman

$$Q = Q(\theta_1, \theta_2, \ldots, \theta_k) = \sum_{i=1}^{n} \big(x_i - m_i(\theta_1, \theta_2, \ldots, \theta_k) \big)^2.$$

8.3 Medelfel

Medelfelet till en skattning θ^* är en skattning av standardavvikelsen för θ^*.

9. Några vanliga fördelningar i statistiken

χ^2-fördelningen

Om $X_1, X_2, \ldots, X_f$ är oberoende och $N(0,1)$ så gäller att
$\sum_{k=1}^{f} X_k^2$ är $\chi^2(f)$-fördelad.

t-fördelningen

Om X är $N(0,1)$ och Y är $\chi^2(f)$ samt om X och Y är oberoende så gäller att $\dfrac{X}{\sqrt{Y/f}}$ är $t(f)$-fördelad.

10. Stickprovsvariablernas fördelningar vid normalfördelade stickprov

10.1 Låt $X_1, \ldots, X_n$ vara oberoende stokastiska variabler som alla är $N(m, \sigma)$.

Då gäller:

a) $\overline{X}$ är $N\Big(m, \dfrac{\sigma}{\sqrt{n}}\Big)$

b) $(n-1)\dfrac{s^2}{\sigma^2}$ är $\chi^2(n-1)$

c) $\overline{X}$ och s^2 är oberoende

d) $\dfrac{\overline{X} - m}{s/\sqrt{n}}$ är $t(n-1)$

10.2 Låt $X_1, \ldots, X_{n_1}$ vara $N(m_1, \sigma)$ och $Y_1, \ldots, Y_{n_2}$ vara $N(m_2, \sigma)$ och samtliga stokastiska variabler antas oberoende. Då gäller:

a) $\overline{X} - \overline{Y}$ är $N\Big(m_1 - m_2, \sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\Big)$

b) $\dfrac{(n_1 + n_2 - 2)s^2}{\sigma^2}$ är $\chi^2(n_1 + n_2 - 2)$ där $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

(s_1^2 och s_2^2 är respektive stickprovsvarians)

c) $\overline{X} - \overline{Y}$ och s^2 är oberoende

d) $\dfrac{\overline{X} - \overline{Y} - (m_1 - m_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ är $t(n_1 + n_2 - 2)$

10.3 Låt $X_1, \ldots, X_{n_1}$ vara $N(m_1, \sigma_1)$ och $Y_1, \ldots, Y_{n_2}$ vara $N(m_2, \sigma_2)$ och samtliga stokastiska variabler antas oberoende. Då gäller:

$\overline{X} - \overline{Y}$ är $N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

11. Konfidensintervall

11.1 λ-metoden

Låt θ^* vara $N(\theta, D)$ där D är känd och θ okänd. Då är

$\theta^* \pm D \cdot \lambda_{\alpha/2}$

ett konfidensintervall för θ med konfidensgraden $1 - \alpha$.

11.2 t-metoden

Låt θ^* vara $N(\theta, D)$ där D och θ är okända och D inte beror på θ.

Låt d vara en skattning av D sådan att $\frac{\theta^* - \theta}{d}$ är $t(f)$. Då är

$\theta^* \pm d \cdot t_{\alpha/2}(f)$

ett konfidensintervall för θ med konfidensgraden $1 - \alpha$.

11.3 Approximativa metoden

Låt θ^* vara approximativt $N(\theta, D)$.

Antag att d är en lämplig skattning av D. Då är

$\theta^* \pm d \cdot \lambda_{\alpha/2}$

ett konfidensintervall för θ med den *approximativa* konfidensgraden $1 - \alpha$.

11.4 χ^2-metoden

Låt θ^* vara en skattning av en parameter θ sådan att $f \cdot \left(\frac{\theta^*}{\theta}\right)^2$ är $\chi^2(f)$.

Då är

$\left(\frac{f \cdot (\theta^*)^2}{\chi^2_{\alpha/2}(f)}\ ,\ \frac{f \cdot (\theta^*)^2}{\chi^2_{1-\alpha/2}(f)}\right)$

ett konfidensintervall för θ^2 med konfidensgraden $1 - \alpha$.

12. Linjär regression

12.1 Fördelningar

Låt Y_i vara $N(\alpha+\beta(x_i-\overline{x}),\sigma)$, $i=1,2,\ldots,n$, och oberoende. Då gäller:

a) $\alpha^*=\overline{Y}$ är $N\Big(\alpha,\dfrac{\sigma}{\sqrt{n}}\Big)$

b) $\beta^*=\dfrac{\sum_{i=1}^n(x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^n(x_i-\overline{x})^2}$ är $N\left(\beta,\dfrac{\sigma}{\sqrt{\sum_{i=1}^n(x_i-\overline{x})^2}}\right)$

c) $\dfrac{(n-2)s^2}{\sigma^2}$ är $\chi^2(n-2)$ där $s^2=\dfrac{1}{n-2}\displaystyle\sum_{i=1}^n\big(Y_i-\alpha^*-\beta^*(x_i-\overline{x})\big)^2$

d) α^*, β^* och s^2 är oberoende

12.2 Konfidensintervall

$$I_\alpha:\ \alpha^*\pm t_{p/2}(n-2)\frac{s}{\sqrt{n}}$$

$$I_\beta:\ \beta^*\pm t_{p/2}(n-2)\frac{s}{\sqrt{\sum_{i=1}^n(x_i-\overline{x})^2}}$$

$$I_{\alpha+\beta(x_0-\overline{x})}:\ \alpha^*+\beta^*(x_0-\overline{x})\pm t_{p/2}(n-2)\,s\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^n(x_i-\overline{x})^2}}$$

12.3 Beräkningsaspekter

$$\sum_{i=1}^n(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^n(x_i-\overline{x})y_i=\sum_{i=1}^n x_i(y_i-\overline{y})=\sum_{i=1}^n x_iy_i-n\overline{x}\,\overline{y}$$

$$\sum_{i=1}^n(x_i-\overline{x})^2=\sum_{i=1}^n x_i^2-n\overline{x}^2$$

$$(n-2)s^2=Q_0=\sum_{i=1}^n(y_i-\overline{y})^2-\beta^{*2}\sum_{i=1}^n(x_i-\overline{x})^2$$

13. Hypotesprövning

13.1 Definitioner

Signifikansnivån (felrisken) α är (det maximala värdet av) $P(\text{förkasta } H_0)$ då hypotesen H_0 är sann.

Styrkefunktionen $h(\theta)=P(\text{förkasta } H_0)$ då θ är rätt parametervärde.

13.2 Konfidensmetoden

Förkasta $H_0:\theta=\theta_0$ om θ_0 ej faller inom ett lämpligt valt konfidensintervall med motsvarande konfidensgrad.

13.3 χ^2-test

Man gör n oberoende upprepningar av ett försök som ger något av resultaten $A_1, A_2, \ldots, A_r$ med respektive sannolikheter $p_1, p_2, \ldots, p_r$. Låt för $j = 1, 2, \ldots, r$ den stokastiska variabeln X_j beteckna antalet försök som ger resultatet A_j. Då blir

$Q = \sum_{j=1}^{r} \frac{(X_j - np_j)^2}{np_j}$ approximativt $\chi^2(r-1)$-fördelad

då $p_1, p_2, \ldots, p_r$ är kända och $np_j \geq 5,\ j = 1, 2, \ldots, r$.

Om vi skattar k parametrar ur data för att skatta $p_1, p_2, \ldots, p_r$ med $p_1^*, p_2^*, \ldots, p_r^*$ så är

$Q' = \sum_{j=1}^{r} \frac{(X_j - np_j^*)^2}{np_j^*}$ approximativt $\chi^2(r-k-1)$-fördelad.

Homogenitetstest: Man vill testa om sannolikheterna för resultaten $A_1, A_2, \ldots, A_r$ är desamma i s försöksserier. Inför beteckningar enligt nedanstående tabell:

Serie	Antal observationer av					Antal försök
	A_1	A_2	A_3	$\ldots$	A_r	
1	x_{11}	x_{12}	x_{13}	$\ldots$	x_{1r}	n_1
2	x_{21}	x_{22}	x_{23}	$\ldots$	x_{2r}	n_2
$\vdots$	$\vdots$					$\vdots$
s	x_{s1}	x_{s2}	x_{s3}	$\ldots$	x_{sr}	n_s

Bilda $Q = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*}$

där $p_j^* = \frac{1}{N} \sum_{i=1}^{s} x_{ij},\ \ j = 1, 2, \ldots, r,\ \ N = \sum_{i=1}^{s} n_i$

Q är approximativt $\chi^2((r-1)(s-1))$-fördelad.

Beräkningsaspekt: $Q = \sum_{j=1}^{r} \frac{x_j^2}{np_j} - n,\ \ Q' = \sum_{j=1}^{r} \frac{x_j^2}{np_j^*} - n$