Matematisk statistik
KTH

Formel- och tabellsamling i matematisk statistik

Vårterminen 2005

1. Kombinatorik

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Tolkning: $\binom{n}{k}$ = antalet delmängder av storlek k ur en mängd med n element.

2. Stokastiska variabler

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

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

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

3. 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 $0 < p < 1$.

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

Hypergeometriska fördelningen

X är $\mathrm{Hyp}(N,n,p)$ om $p_X(k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$, $0 \le k \le Np$,

$0 \le n-k \le N(1-p)$, där N, Np och n är positiva heltal samt $N \ge 2$,

$n < N$, $0 < p < 1$. $E(X) = np$, $V(X) = \frac{N-n}{N-1} \cdot np(1-p)$

Poissonfördelningen

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

$E(X) = \mu$, $V(X) = \mu$

4. Kontinuerliga fördelningar

Likformig fördelning

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

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

Exponentialfördelningen

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

$E(X) = \frac{1}{\lambda}, \ V(X) = \frac{1}{\lambda^2}$

Normalfördelningen

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

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

X är $N(\mu, \sigma)$ om och endast om $\frac{X - \mu}{\sigma}$ är $N(0, 1)$.

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

En linjär sammansättning $\sum a_i X_i + b$ av oberoende, normalfördelade stokastiska variabler är normalfördelad.

5. Centrala gränsvärdessatsen

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

6. Approximation

$\text{Hyp}(N, n, p) \sim \text{Bin}(n, p)$ om $\frac{n}{N} \leq 0.1$

$\text{Bin}(n, p) \sim \text{Po}(np)$ om $p \leq 0.1$

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

$\text{Po}(\mu) \sim N(\mu, \sqrt{\mu})$ om $\mu \geq 15$

7. Tjebysjovs olikhet

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

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

8. Statistiskt material

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

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

9. Punktskattningar

9.1 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. Det värde θ^*_{obs} som maximerar L-funktionen

$$L(\theta) = \begin{cases} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = (\text{om oberoende}) = p_{X_1}(x_1; \theta) \cdots p_{X_n}(x_n; \theta) \\ f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = (\text{om oberoende}) = f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta) \end{cases}$$

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

9.2 Minsta-kvadrat-metoden

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

$E(X_i) = \mu_i(\theta_1, \theta_2, \ldots, \theta_k)$ 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 $(\theta_1)^*_{\text{obs}}, (\theta_2)^*_{\text{obs}}, \ldots, (\theta_k)^*_{\text{obs}}$ som minimerar kvadratsumman

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

9.3 Medelfel

En skattning av $D(\theta^*)$ kallas *medelfelet* för θ^* och betecknas $d(\theta^*)$.

9.4 Felfortplantning

Med beteckningar och förutsättningar enligt läroboken gäller

a) $E(g(\theta^*)) \approx g(\theta^*_{\text{obs}})$

$D(g(\theta^*)) \approx \big|g'(\theta^*_{\text{obs}})\big| \cdot D(\theta^*)$

b) $E(g(\theta^*_1, \ldots, \theta^*_n)) \approx g\big((\theta_1)^*_{\text{obs}}, \ldots, (\theta_n)^*_{\text{obs}}\big)$

$$V(g(\theta^*_1, \ldots, \theta^*_n) \approx \sum_{i=1}^{n} \sum_{j=1}^{n} C(\theta^*_i, \theta^*_j) \cdot \left[\frac{\partial g}{\partial x_i} \cdot \frac{\partial g}{\partial x_j}\right]_{x_k = (\theta_k)^*_{\text{obs}}, k=1,\ldots,n}$$

10. 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.

11. Stickprovsvariablernas fördelningar vid normalfördelade stickprov

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

Då gäller:

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

b) $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2}$ är $\chi^2(n-1)$

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

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

11.2 Låt $X_1, \ldots, X_{n_1}$ vara $N(\mu_1, \sigma)$ och $Y_1, \ldots, Y_{n_2}$ vara $N(\mu_2, \sigma)$ och samtliga stokastiska variabler antas oberoende. Då gäller:

a) $\overline{X} - \overline{Y}$ är $N\Big(\mu_1 - \mu_2, \sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\Big)$

b) $\dfrac{(n_1 + n_2 - 2)S^2}{\sigma^2}$ är $\chi^2(n_1 + n_2 - 2)$ där $S^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$,

$S_1^2 = \dfrac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$ och $S_2^2 = \dfrac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$

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

d) $\dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ är $t(n_1 + n_2 - 2)$

11.3 Låt $X_1, \ldots, X_{n_1}$ vara $N(\mu_1, \sigma_1)$ och $Y_1, \ldots, Y_{n_2}$ vara $N(\mu_2, \sigma_2)$ och samtliga stokastiska variabler antas oberoende. Då gäller:

$\overline{X} - \overline{Y}$ är $N\left(\mu_1 - \mu_2, \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$

12. Konfidensintervall

12.1 λ-metoden

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

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

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

12.2 t-metoden

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

Låt D^*_{obs} vara en punktskattning av D sådan att $\dfrac{\theta^* - \theta}{D^*}$ är $t(f)$. Då är

$\theta^*_{\text{obs}} \pm D^*_{\text{obs}} \cdot t_{\alpha/2}(f)$

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

12.3 Approximativa metoden

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

Antag att D^*_{obs} är en lämplig punktskattning av D. Då är

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

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

12.4 Metod baserad på χ^2-fördelning

Låt θ^*_{obs} vara en punktskattning av en parameter θ sådan att

$f \cdot \left(\dfrac{\theta^*}{\theta}\right)^2$ är $\chi^2(f)$. Då är

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

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

13. Linjär regression

13.1 Fördelningar

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

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

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

c) $\alpha^* + \beta^* x_0$ är $N\left(\alpha + \beta x_0, \sigma \sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right)$

d) $\dfrac{(n-2)S^2}{\sigma^2}$ är $\chi^2(n-2)$ där $S^2 = \dfrac{1}{n-2} \displaystyle\sum_{i=1}^n (Y_i - \alpha^* - \beta^* x_i)^2$

e) S^2 är oberoende av α^* och β^*

13.2 Konfidensintervall

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

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

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

13.3 Beräkningsaspekter

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

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

$$S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2$$

$$(n-2)s^2 = S_{yy} - (\beta^*_{\text{obs}})^2 S_{xx} = S_{yy} - \beta^*_{\text{obs}} \cdot S_{xy} = \min_{\alpha,\beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

14. Hypotesprövning

14.1 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.

14.2 Konfidensmetoden

Förkasta $H_0 : \theta = \theta_0$ på nivån α om θ_0 ej faller inom ett lämpligt valt konfidensintervall med konfidensgraden $1 - \alpha$.

14.3 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 $P(A_1), P(A_2), \ldots, P(A_r)$. Låt för $j = 1, 2, \ldots, r$ den stokastiska variabeln X_j beteckna antalet försök som ger resultatet A_j.

Test av given fördelning: Vi vill testa $H_0 : P(A_1) = p_1, P(A_2) = p_2, \ldots,$ $P(A_r) = p_r$ för givna sannolikheter $p_1, p_2, \ldots, p_r$. Då blir $Q = \sum_{j=1}^{r} \frac{(x_j - np_j)^2}{np_j}$ ett utfall av en approximativt $\chi^2(r-1)$-fördelad stokastisk variabel om H_0 är sann och $np_j \geq 5,\ j = 1, 2, \ldots, r$.

Om vi skattar k parametrar ur data, $\theta = (\theta_1, \ldots, \theta_k)$, för att skatta $p_1, p_2, \ldots, p_r$ med $p_1(\theta^*_{\text{obs}}), p_2(\theta^*_{\text{obs}}), \ldots, p_r(\theta^*_{\text{obs}})$ så är $Q' = \sum_{j=1}^{r} \frac{\left(x_j - np_j(\theta^*_{\text{obs}})\right)^2}{np_j(\theta^*_{\text{obs}})}$ ett utfall av en approximativt $\chi^2(r-k-1)$-fördelad stokastisk variabel.

Beräkningsaspekt: $Q = \sum_{j=1}^{r} \frac{x_j^2}{np_j} - n,\ \ Q' = \sum_{j=1}^{r} \frac{x_j^2}{np_j(\theta^*_{\text{obs}})} - n$

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
Kolonnsumma	m_1	m_2	m_3	$\ldots$	m_r	N

Bilda $Q = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$.

Q är ett utfall av en approximativt $\chi^2((r-1)(s-1))$-fördelad stokastisk variabel.

Kontingenstabell (test av oberoende mellan rader och kolonner):

Samma teststorhet och fördelning som ovan.