

Matematisk statistik
KTH

Formelsamling i matematisk statistik

November 2019

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

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ C(X, Y) &= E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y) \\ \rho(X, Y) &= \frac{C(X, Y)}{D(X)D(Y)} \end{aligned}$$

3 Diskreta fördelningar

Binomialfördelningen

X är $\text{Bin}(n, p)$ om $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \dots, n$, där $0 < p < 1$.
 $E(X) = np$, $V(X) = np(1-p)$

”För-första-gången”-fördelningen

X är $\text{ffg}(p)$ om $p_X(k) = p(1-p)^{k-1}$, $k = 1, 2, 3, \dots$, där $0 < p < 1$.
 $E(X) = \frac{1}{p}$, $V(X) = \frac{1-p}{p^2}$

Hypergeometrisk fördelningen

X är $\text{Hyp}(N, n, p)$ om $p_X(k) = \frac{\binom{Np}{k} \binom{N(1-p)}{n-k}}{\binom{N}{n}}$, $0 \leq k \leq Np$,
 $0 \leq n-k \leq N(1-p)$, där N , Np och n är positiva heltal samt $N \geq 2$, $n < N$,
 $0 < p < 1$. $E(X) = np$, $V(X) = \frac{N-n}{N-1} np(1-p)$

Poissonfördelningen

X är $\text{Po}(\mu)$, där $\mu > 0$, om $p_X(k) = \frac{\mu^k}{k!} e^{-\mu}$, $k = 0, 1, 2, \dots$
 $E(X) = \mu$, $V(X) = \mu$

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 $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{för } x > 0 \\ 0 & \text{annars} \end{cases}$

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

Normalfördelningen

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

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

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 $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $-\infty < x < \infty$.

En linjärkombination $\sum_i a_i X_i + b$ av oberoende, normalfördelade stokastiska variabler är normalfördelad.

5 Centrala gränsvärdessatsen

Om $X_1, X_2, \dots, X_n$ är oberoende, likafördelade stokastiska variabler med väntevärde μ och standardavvikelse $\sigma > 0$, så är $Y_n = X_1 + \dots + X_n$ approximativt $N(\mu n, \sigma\sqrt{n})$ 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 $p \leq 0.1$

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

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

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

9 Punktskattningar

9.1 Maximum-likelihoodmetoden

Låt x_i vara en observation av X_i , $i = 1, 2, \dots, n$, där fördelningen för X_i beror på en okänd parameter θ . Det värde θ_{obs}^* som maximerar likelihoodfunktionen

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

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

9.2 Minsta-kvadratmetoden

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

$$E(X_i) = \mu_i(\theta_1, \theta_2, \dots, \theta_k) \text{ och } V(X_i) = \sigma^2,$$

där $\theta_1, \theta_2, \dots, \theta_k$ är okända parametrar och $X_1, X_2, \dots, X_k$ är oberoende.

Minsta-kvadratskattningarna (*MK-skattningarna*) av $\theta_1, \theta_2, \dots, \theta_k$ är de värden

$(\theta_1)_{\text{obs}}^*, (\theta_2)_{\text{obs}}^*, \dots, (\theta_k)_{\text{obs}}^*$ som minimerar kvadratsumman

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

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

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

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

$$\text{b) } E(g(\theta_1^*, \dots, \theta_n^*)) \approx g((\theta_1)_{\text{obs}}^*, \dots, (\theta_n)_{\text{obs}}^*)$$

$$V(g(\theta_1^*, \dots, \theta_n^*)) \approx \sum_{i=1}^n \sum_{j=1}^n C(\theta_i^*, \theta_j^*) \left[\frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} \right]_{x_k = (\theta_k)_{\text{obs}}^*, k=1, \dots, n}$$

10 Några vanliga fördelningar i statistiken

χ^2 -fördelningen

Om $X_1, X_2, \dots, X_f$ är oberoende $N(0, 1)$, så gäller det att

$$\sum_{k=1}^f X_k^2 \text{ är } \chi^2(f)\text{-fördelad.}$$

t -fördelningen

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

11 Stickprovsvariablernas fördelningar vid normalfördelade stickprov

11.1 Ett normalfördelat stickprov

Låt $X_1, \dots, X_n$ vara oberoende stokastiska variabler som alla är $N(\mu, \sigma)$. Då gäller:

- a) $\bar{X}$ är $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- b) $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$ är $\chi^2(n-1)$
- c) $\bar{X}$ och S^2 är oberoende
- d) $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ är $t(n-1)$

11.2 Två normalfördelade stickprov med samma varians

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

- a) $\bar{X} - \bar{Y}$ är $N\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$
- b) $\frac{(n_1 + n_2 - 2)S^2}{\sigma^2}$ är $\chi^2(n_1 + n_2 - 2)$ där $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$,
 $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$ och $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$
- c) $\bar{X} - \bar{Y}$ och S^2 är oberoende
- d) $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ är $t(n_1 + n_2 - 2)$

11.3 Två normalfördelade stickprov med olika varians

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

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

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 $\frac{\theta^* - \theta}{D^*}$ 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

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

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

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

13 Linjär regression

13.1 Fördelningar

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

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

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

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

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

e) S^2 är oberoende av α^* och β^*

13.2 Konfidensintervall

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

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

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

13.3 Beräkningsaspekter

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

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

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

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

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

Antag att n oberoende upprepningar av ett försök med de möjliga utfallen $A_1, A_2, \dots, A_r$ med respektive sannolikheter $P(A_1), P(A_2), \dots, P(A_r)$. Låt, för $j = 1, 2, \dots, r$, den stokastiska variabeln X_j beteckna antalet försök som ger resultatet A_j .

Test av given fördelning

Vi vill testa $H_0 : P(A_1) = p_1, P(A_2) = p_2, \dots, P(A_r) = p_r$ för givna sannolikheter $p_1, p_2, \dots, p_r$. Då blir

$$Q = \sum_{j=1}^r \frac{(x_j - np_j)^2}{np_j} \text{ ett utfall av en approximativt } \chi^2(r-1)\text{-fördelad stokastisk}$$

variabel om H_0 är sann och $np_j \geq 5, j = 1, 2, \dots, r$.

Om vi skattar k parametrar ur data, $\theta = (\theta_1, \dots, \theta_k)$ för att skatta $p_1, p_2, \dots, p_r$ med $p_1(\theta_{\text{obs}}^*), p_2(\theta_{\text{obs}}^*), \dots, p_r(\theta_{\text{obs}}^*)$, så är

$$Q' = \sum_{j=1}^r \frac{(x_j - np_j(\theta_{\text{obs}}^*))^2}{np_j(\theta_{\text{obs}}^*)} \text{ ett utfall av en approximativt } \chi^2(r-k-1)\text{-fördelad}$$

stokastisk variabel.

Homogenitetstest

Vi vill testa om sannolikheterna för utfallen $A_1, A_2, \dots, A_r$ är desamma i s försöksserier. Inför beteckningar enligt nedanstående tabell:

Serie	Antal observationer av					Antal försök
	A_1	A_2	A_3	$\dots$	A_r	
1	x_{11}	x_{12}	x_{13}	$\dots$	x_{1r}	n_1
2	x_{21}	x_{22}	x_{23}	$\dots$	x_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s	x_{s1}	x_{s2}	x_{s3}	$\dots$	x_{sr}	n_s
Kolonnsumma	m_1	m_2	m_3	$\dots$	m_r	N

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

Q är ett utfall av en approximativt $\chi^2((r-1)(s-1))$ -fördelad stokastisk variabel om $n_i m_j / N \geq 5$, för alla $i = 1, 2, \dots, s$ och $j = 1, 2, \dots, r$.

Oberoendetest

Antag att värdemängden för den stokastiska variabeln X kan delas in i kategorierna $A_1, A_2, \dots, A_r$ och att värdemängden för den stokastiska variabeln Y kan delas in i kategorierna $B_1, B_2, \dots, B_s$. Vi vill testa om de stokastiska variablerna X och Y är oberoende.

Antal observationer	A_1	A_2	A_3	$\dots$	A_r	Radsumma
B_1	x_{11}	x_{12}	x_{13}	$\dots$	x_{1r}	n_1
B_2	x_{21}	x_{22}	x_{23}	$\dots$	x_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_s	x_{s1}	x_{s2}	x_{s3}	$\dots$	x_{sr}	n_s
Kolonnsumma	m_1	m_2	m_3	$\dots$	m_r	N

Samma teststorhet och fördelning kan användas som vid homogenitetstest.