Matematisk statistik
KTH

Formelsamling i matematisk statistik

SF1910, SF1912, SF1914, SF1915, SF1916, SF1917, SF1918, SF1919, SF1920, SF1921, SF1922, SF1923, SF1924, SF1925, SF1935

Januari 2023

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

$V(X) = E\left(X^2\right) - (E(X))^2$
$C(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$
$\rho(X,Y) = \frac{C(X,Y)}{D(X)\,D(Y)}$

3 Diskreta fördelningar

Binomialfördelningen

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

"För-första-gången"-fördelningen

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

Hypergeometriska fördelningen

X är 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 $N \ge 2$, $n < N$, $0 < p < 1$. $E(X) = np, \quad V(X) = \frac{N-n}{N-1} np(1-p)$

Poissonfördelningen

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

4 Kontinuerliga fördelningar

Likformig fördelning

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

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

Exponentialfördelningen

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

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

Normalfördelningen

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

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

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

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

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

Gammafördelningen

X är Gamma(c, λ) om

$f_X(x) = \begin{cases} \frac{\lambda^c}{\Gamma(c)} \cdot x^{c-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0, \end{cases}$

där $\Gamma(c) = \int_0^{+\infty} x^{c-1} e^{-x} dx$. Om c positivt heltal har vi $\Gamma(c) = (c-1)!$.

$E(X) = \dfrac{c}{\lambda}, \quad V(X) = \dfrac{c}{\lambda^2}$

Betafördelningen

X är Beta(α, β) om

$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \leq x \leq 1, \\ 0, & \text{annars.} \end{cases}$

$E(X) = \dfrac{\alpha}{\alpha+\beta}, \quad V(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

5 Centrala gränsvärdessatsen

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

6 Approximation

Hyp(N, n, p) approximeras av Bin(n, p) om $\dfrac{n}{N} \leq 0.1$

Bin(n, p) approximeras av Po(np) om $p \leq 0.1$

Bin(n, p) approximeras av $N\left(np, \sqrt{np(1-p)}\right)$ om $np(1-p) \geq 10$

Po(μ) approximeras av $N(\mu, \sqrt{\mu})$ om $\mu \geq 15$

7 Tjebysjovs olikhet

Om $E(X) = \mu$ och $D(X) = \sigma > 0$ så gäller för varje $k > 0$ att
$P(|X - \mu| > k\sigma) \leq \dfrac{1}{k^2}$

8 Statistiskt material

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

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

9 Punktskattningar

9.1 Maximum-likelihoodmetoden

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

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

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

9.2 Minsta-kvadratmetoden

Låt x_i vara en observation av X_i, $i = 1, 2, ..., n$, och antag att
$E(X_i) = \mu_i(\theta_1, \theta_2, ..., \theta_k)$ och $V(X_i) = \sigma^2$,
där $\theta_1, \theta_2, ..., \theta_k$ är okända parametrar och $X_1, X_2, ..., X_n$ är oberoende.
Minsta-kvadratskattningarna (MK-skattningarna) av $\theta_1, \theta_2, ..., \theta_k$ är de värden $(\theta_1)^*_{\text{obs}}, (\theta_2)^*_{\text{obs}}, ..., (\theta_k)^*_{\text{obs}}$ som minimerar kvadratsumman

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

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

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

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

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

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

10 Några vanliga fördelningar i statistiken

χ^2-fördelningen

Om $X_1, X_2, ..., X_f$ är oberoende $N(0,1)$, så gäller det 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 det att $\dfrac{X}{\sqrt{Y/f}}$ är $t(f)$-fördelad.

11 Stickprovsvariablernas fördelningar vid normalfördelade stickprov

11.1 Ett normalfördelat stickprov

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

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

b) $\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2}$ 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 Två normalfördelade stickprov med samma varians

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

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

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

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

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

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

11.3 Två normalfördelade stickprov med olika varians

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

$\overline{X} - \overline{Y}$ 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

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

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

13 Linjär regression

13.1 Fördelningar

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

a) $\beta^* = \dfrac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}$ 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 $S^2 = \dfrac{1}{n-2} \sum\limits_{i=1}^{n} (Y_i - \alpha^* - \beta^* x_i)^2$

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

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

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

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

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

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 $A_1, A_2, ..., A_r$ med respektive sannolikheter $P(A_1), P(A_2), ..., P(A_r)$. Låt, för $j = 1, 2, ..., r$, den stokastiska variableln X_j beteckna antalet försök som ger resultatet A_j.

Test av given fördelning

Vi vill testa $H_0 : P(A_1) = p_1, P(A_2) = p_2, ..., P(A_r) = p_r$ för givna sannolikheter $p_1, p_2, ..., p_r$. 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 $np_j \geq 5$, $j = 1, 2, ..., r$.
Beräkningsaspekt: $Q = \sum_{j=1}^{r} \left(\frac{x_j^2}{np_j} \right) - n$
Om vi skattar k parametrar ur data, $\theta = (\theta_1, ..., \theta_k)$ för att skatta $p_1, p_2, ..., p_r$ med $p_1(\theta^*_{\text{obs}}), p_2(\theta^*_{\text{obs}}), ..., p_r(\theta^*_{\text{obs}})$, så är
$Q' = \sum_{j=1}^{r} \frac{(x_j - np_j(\theta^*_{\text{obs}}))^2}{np_j(\theta^*_{\text{obs}})}$ ett utfall av en approximativt $\chi^2(r-k-1)$-fördelad stokastisk variabel.

Homogenitetstest

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

Serie	Antal observationer av					Antal försök
	A_1	A_2	A_3	$\cdots$	A_r	
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1r}	n_1
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s	x_{s1}	x_{s2}	x_{s3}	$\cdots$	x_{sr}	n_s
Kolonnsumma	m_1	m_2	m_3	$\cdots$	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 $\chi^2\left((r-1)(s-1)\right)$-fördelad stokastisk variabel om $n_i m_j / N \geq 5$, för alla $i = 1, 2, ..., s$ och $j = 1, 2, ..., r$.

Oberoendetest

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

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

Samma teststorhet och fördelning kan användas som vid homogenitetstest.

15 Bayesiansk inferens

15.1 Apriori- och aposteriorifördelning

Givet en parameter Θ i parameterrummet Ω_θ med apriorifördelning $f_\Theta(\theta)$ och en datapunkt X med datafördelning $f_{X|\Theta}(x \mid \theta)$ har vi aposteriorifördelningen

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{f_X(x)}$$

där den aprioriprediktiva fördelningen (marginalfördelningen) för X är $f_X(x) = \int_{\Omega_\theta} f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta$ om Θ är kontinuerlig och $f_X(x) = \sum_{\Omega_\theta} f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)$ om Θ är diskret.

15.2 Konjugatfamiljer och uppdateringsregler

Om $\Theta \sim \text{Beta}(\alpha, \beta)$ och $X_1, X_2, \ldots, X_k$ är betingat oberoende givet Θ med $X_i \mid \Theta = \theta \sim \text{Bin}(n_i, \theta)$ för $i = 1, 2, \ldots, k$ så har vi

$$\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \sim \text{Beta}\left(\alpha + \sum_{i=1}^{k} x_i, \beta + \sum_{i=1}^{k} n_i - x_i\right).$$

Om $\Theta \sim N(\mu_0, \tau_0)$ och $X_1, X_2, \ldots, X_k$ är betingat oberoende givet Θ med $X_i \mid \Theta = \theta \sim N(\theta, \sigma)$ för $i = 1, 2, \ldots, k$ så har vi

$$\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \sim N\left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^{k} x_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}}, \frac{1}{\sqrt{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}}}\right).$$

Om $\Theta \sim \text{Gamma}(c, \lambda)$ och $X_1, X_2, \ldots, X_k$ är betingat oberoende givet Θ med $X_i \mid \Theta = \theta \sim \text{Po}(\theta)$ för $i = 1, 2, \ldots, k$ så har vi

$$\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \sim \text{Gamma}\left(c + \sum_{i=1}^{k} x_i, \lambda + k\right).$$

Om $\Theta \sim \text{Gamma}(c, \lambda)$ och $X_1, X_2, \ldots, X_k$ är betingat oberoende givet Θ med $X_i \mid \Theta = \theta \sim \text{Exp}(\theta)$ för $i = 1, 2, \ldots, k$ så har vi

$$\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \sim \text{Gamma}\left(c + k, \lambda + \sum_{i=1}^{k} x_i\right).$$