Karlstads universitet
Institutionen för informationsteknologi
Avdelningen för Statistik

CIT B01 vt04,
Jan Rudander
15 april 2004

Formelsamling statistik

Sannolikhetslära

$\frac{g}{m}$-principen, kräver att alla utfall är lika sannolika.

$$P(A) + P(A^*) = 1$$

<u>Lagen om total sannolikhet</u>:

$$\left.\begin{aligned} i \neq j \Rightarrow A_i \cap A_j = \varnothing \\ \bigcup_i A_i = \Omega \end{aligned}\right\} \Rightarrow P(B) = \sum_i P(B \cap A_i)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

<u>Betingad sannolikhet</u>:

$P(A \cap B) = P(A) \cdot P(B \mid A)$, omformulerat: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$.

<u>Oberoende</u>:

A och B oberoende $\Leftrightarrow$ $P(A \cap B) = P(A) \cdot P(B)$ $\Leftrightarrow$
$P(B) = P(B \mid A) = P(B \mid A^*)$

Moment

$$E[g(X)] = \begin{cases} \sum_x g(x) P(x) & \textit{diskreta} \quad \textit{fallet} \\ \int_x g(x) f(x) & \textit{kontinuerliga} \quad \textit{fallet} \end{cases}$$

<u>Väntevärde och varians</u>:

$$\mu = E(X)$$
$$\sigma^2 = V(X) = E\left[(X - \mu)^2\right] = \ldots = E(X^2) - \mu^2$$

<u>Ändra läge och skala</u>:

$$E(aX + b) = aE(X) + b$$
$$V(aX + b) = a^2 V(X)$$

Summor:

$$E\left(\sum_i X_i\right) = \sum_i E(X_i)$$

$$V\left(\sum_i X_i\right) = \sum_i V(X_i)$$ om X och Y är oberoende.

X och Y oberoende $\Rightarrow$ $E(XY) = E(X) \cdot E(Y)$ $\Leftrightarrow$ $C(X,Y) = 0$
(obs: den första pilen är enkelriktad)

Gauss approximationsformler:

$E[g(X)] \approx g(\mu)$, där $\mu = E(X)$
$V[g(X)] \approx [g'(\mu)]^2 \cdot V(X)$

Tjebyshevs olikhet:

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

Konvergens:
$Z_1, Z_2, \ldots$ säges *konvergera i sannolikhet* mot konstanten c omm för varje $\varepsilon > 0$ gäller att $P(|Z_n - c| < \varepsilon) \to 1$ då $n \to \infty$. Vi skriver $Z_n \xrightarrow{p} \theta$.

STL (stora talens lag): Anta att $X_1, X_2, \ldots$ är *oberoende* och har *samma fördelning* samt att väntevärdet $\mu = E(X)$ är ändligt. Då gäller att

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

då $n \to \infty$.

Några vanliga diskreta fördelningar

Nedan är $p+q=1$ dvs. $q=1-p$.

beteckning	Sannolikhetsfunktion $p(x)=P(X=x)$	Fördelningsfunktion $F(x)=P(X\le x)$	Väntevärde $\mu=E(X)$	Varians $\sigma^2=V(X)$
$Hyp(N,n,p)$	$\dfrac{\binom{Np}{x}\cdot\binom{Nq}{n-x}}{\binom{N}{n}}$ för de möjliga x-värdena	$F(x)=\sum_{k=0}^{x}p(k)$	$\mu=np$	$\sigma^2=npq\cdot\dfrac{N-n}{N-1}$
$Ber(p)\equiv Bin(1,p)$	$p^x q^{1-x}$ för $x=0,1$	$F(x)=\sum_{k=0}^{x}p(k)$	$\mu=p$	$\sigma^2=pq$
$Bin(n,p)$	$\binom{n}{x}p^x q^{n-x}$ för $x=0,1,\ldots,n$	$F(x)=\sum_{k=0}^{x}p(k)$	$\mu=np$	$\sigma^2=npq$
$Geo(p)$	pq^{x-1} för $x=0,1,2,\ldots$	$F(x)=1-q^x$	$\mu=\dfrac{1}{p}$	$\sigma^2=\dfrac{q}{p^2}$
$NegBin(l,p)$	$\binom{x-1}{l-1}p^l q^{x-l}$ för $x=l,l+1,l+2,\ldots$	$F(x)=\sum_{k=0}^{x}p(k)$	$\mu=\dfrac{l}{p}$	$\sigma^2=\dfrac{lq}{p^2}$
$Po(\lambda)$	$\dfrac{\lambda^x e^{-\lambda}}{x!}$ för $x=0,1,2,\ldots$	$F(x)=\sum_{k=0}^{x}p(k)$	$\mu=\lambda$	$\sigma^2=\lambda$

Några vanliga kontinuerliga fördelningar

beteckning	Täthetsfunktion $f(x) = F'(x)$	Fördelningsfunktion $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$	Väntevärde $\mu = E(X)$	Varians $\sigma^2 = V(X)$
$R(a,b)$	$\frac{1}{b-a}$ för $a \le x \le b$	$\frac{x-a}{b-a}$	$\mu = \frac{a+b}{2}$	$\sigma^2 = \frac{(b-a)^2}{12}$
$Exp(\lambda) \equiv$ $\equiv Gam(1,\lambda)$	$\lambda e^{-\lambda x}$ för $x \ge 0$	$1 - e^{-\lambda x}$	$\mu = \frac{1}{\lambda}$	$\sigma^2 = \frac{1}{\lambda^2}$
$Gam(n,\lambda)$	$\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$ för $x \ge 0$	$F(x) = \int_0^x f(t)dt$	$\mu = \frac{n}{\lambda}$	$\sigma^2 = \frac{n}{\lambda^2}$
$Gam(\alpha,\lambda)$	$\frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$ för $x \ge 0$	$F(x) = \int_0^x f(t)dt$	$\mu = \frac{\alpha}{\lambda}$	$\sigma^2 = \frac{\alpha}{\lambda^2}$
$\chi^2(\nu) \equiv$ $\equiv Gam\left(\frac{\nu}{2}, \frac{1}{2}\right)$	$\frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)}$ för $x \ge 0$	$F(x) = \int_0^x f(t)dt$	$\mu = \nu$	$\sigma^2 = 2\nu$
$Wei(\alpha,\beta)$	$\alpha\beta(\alpha x)^{\beta-1} e^{-(\alpha x)^\beta}$ för $x \ge 0$	$F(x) = 1 - e^{-(\alpha x)^\beta}$	$\frac{1}{\alpha}\Gamma\left(1+\frac{1}{\beta}\right)$	σ^2_{Wei} (se nedan)
$N(0,1)$	$\varphi(x) = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}$ för $x \in R$	$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt = \int_{-\infty}^{x} \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} dt$	$\mu = 0$	$\sigma^2 = 1$
$N(\mu,\sigma)$	$\varphi\left(\frac{x-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} =$ $= \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$ för $x \in R$	$\Phi\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \varphi(t)dt$	$\mu = \mu$	$\sigma^2 = \sigma^2$
$T(\nu)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right) \cdot \left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}}$ för $x \in R$	$F(x) = \int_{-\infty}^{x} f(t)dt$	$\mu = 0$, förutsatt att $\nu \ge 2$	$\sigma^2 = \frac{\nu}{\nu-2}$, förutsatt att $\nu \ge 3$

Vanliga approximationer

Exakt fördelning	Förutsättningar/tumregel	Approximerande fördelning	parametrar
$Hyp(N, n, p)$	$\frac{n}{N} \leq 0.05$	$Bin(n, p)$	
CGS (centrala gränsvärdessatsen)			
$\sum_{i=1}^{n} X_i$	oberoende, likafördelade X-variabler, n "stor"	$N(n\mu, \sqrt{n}\sigma)$	
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	oberoende, likafördelade X-variabler, n "stor"	$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	
CGS, några specialfall			
$Hyp(N, n, p)$	$V(X) \geq 10$	$N(\mu, \sigma)$	$\mu = np$ och $\sigma^2 = npq \cdot \frac{N-n}{N-1}$
$Bin(n, p)$	$V(X) \geq 10$	$N(\mu, \sigma)$	$\mu = np$ och $\sigma^2 = npq$
$Po(\lambda)$	$E(X) = \lambda \geq 15$	$N(\mu, \sigma)$	$\mu = \lambda$ och $\sigma^2 = \lambda$
Approximationer rörande **Bernulliförsök** och **Poissonprocessen**			
$Bin(n, p)$	p liten, (n stor)	$Po(\lambda)$	$\lambda = np$
$Geo(p)$	p liten	$Exp(\lambda)$	$\lambda = p$
$NegBin(l, p)$	p liten	$Gam(n, \lambda)$	$l = n$ och $\lambda = p$

Slumptal/pseudoslumptal

<u>kontinuerlig invers transformmetod</u>:

$X_i = F_X^{-1}(U_i)$ där $U_i \in R(0,1)$

Skattningars egenskaper

En skattning θ^* av θ kan delas upp i tre delar enligt

$\theta^* = \theta + b + \varepsilon$, där

θ	är målparametern,
$b = E(\theta^*) - \theta$	är det systematiska felet (eller biasen),
$\varepsilon = \theta^* - E(\theta^*)$	är det rena slumpfelet (med väntevärde noll).

θ^* är en *vvr-skattning* (väntevärdesriktig) av θ omm $E(\theta^*) = \theta$.

θ^* är en *konsistent* skattning av θ omm $\theta^* \xrightarrow{p} \theta$.

Mean Square Error definieras:

$$(1) \qquad MSE(\theta^*) = E([\theta^* - \theta]^2) = \ldots = V(\theta^*) + b^2$$

Av två skattningar θ_1^* och θ_2^* är den *effektivast* som har minst MSE. Om båda är vvr är därmed den effektivast som har minst varians, se (1) ovan.

Beräkningsformel för stickprovsvarians

$$S^2 = \frac{1}{n-1}\Sigma(X_i - \overline{X})^2 = \ldots = \frac{1}{n-1}(\Sigma X_i^2 - n \cdot \overline{X}^2) = \frac{1}{n-1}\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right)$$

Härledning av konfidensintervall

Anta att en skattning $\hat{\theta}$ är NF (normalfördelad). Då gäller

$$(2) \qquad \frac{\hat{\theta} - \mu_{\hat{\theta}}}{\sigma_{\hat{\theta}}} \in N(0,1)$$

Ett konfidensintervall för $\mu_{\hat{\theta}}$ med konfidensgrad $1 - \alpha$ är då

$$(3) \qquad (\hat{\theta} \pm \lambda_{\alpha/2} \cdot \sigma_{\hat{\theta}})$$

Anta $X_1, X_2, ..., X_n$ stickprov från $N(\mu,\sigma)$. Då är $\overline{X} = \frac{\Sigma X_i}{n}$ och $S = \sqrt{\frac{1}{n-1}\Sigma\left(X_i - \overline{X}\right)^2}$ oberoende. Vidare gäller att

(4) $$\overline{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

(5) $$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \in N(0,1)$$

.

(6) $$\frac{\overline{X}-\mu}{S/\sqrt{n}} \in T(n-1),$$

(7) $$\frac{(n-1)\cdot S^2}{\sigma^2} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2} \in \chi^2(n-1)$$

(8) $$\frac{\Sigma\left(X_i - \mu\right)^2}{\sigma^2} \in \chi^2(n)$$

CIT B01 vt04,
Jan Rudander
15 april 2004

Chi-två-test

Chi-två-test, allmänt:

(12) $$Q = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i}$$

är approximativt $\chi^2(\nu)$-fördelad om nollhypotesen är sann och alla $np_i \geq 5$.
Antalet frihetsgrader är $\nu = r - 1 - t$, där t är antalet skattade parametrar.

Oberoendetest/homogenitetstest:

Y_{ij} = ”antalet observationer i cellen på rad i och kolumn j”. $i = 1,\ldots,r$ och $j = 1,\ldots,c$
$R_i = \sum_j Y_{ij}$ är summan av observationerna i rad i.
$C_j = \sum_i Y_{ij}$ är summan av observationerna i kolumn j.

$\hat{E}(Y_{ij}) = \dfrac{R_i \cdot C_j}{n}$ är skattat förväntat värde under H_0.

(13) $$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(Y_{ij} - \hat{E}(Y_{ij})\right)^2}{\hat{E}(Y_{ij})}$$

är approximativt $\chi^2(\nu)$-fördelad om nollhypotesen är sann och alla $\hat{E}(Y_i) \geq 5$.
Antalet frihetsgrader är $\nu = (r-1)(c-1)$.

Köteori

$A(t)$ = antalet ankomster till systemet/kön i intervallet $(0,t]$

$D(t)$ = antalet avgångar från systemet i intervallet $(0,t]$
$D_q(t)$ = antalet avgångar från kön i intervallet $(0,t]$

$N(t) = A(t) - D(t)$, antalet i systemet vid tidpunkt t,
$N_q(t) = A(t) - D_q(t)$, antalet i kön vid tidpunkt t,

En *jämviktsfördelning* $p_0, p_1, p_2, \dots$ existerar om

$$p_k = \lim_{t\to\infty} P(N(t) = k),$$

$$\sum_k p_k = 1.$$

Om jämviktsfördelningen existerar så existerar gränsvärdena

$\frac{A(t)}{t} \xrightarrow{p} \lambda_A$, genomsnittlig ankomstintensitet,

$\frac{\int_0^t N(\tau)d\tau}{t} = \xrightarrow{p} E(N) = \nu$, genomsnittligt antal i systemet,

$W(t) = \frac{\int_0^t N(\tau)d\tau}{A(t)} \xrightarrow{p} \omega$, genomsnittlig tid i systemet,

$\frac{\int_0^t N_q(\tau)d\tau}{t} = \xrightarrow{p} E(N_q) = \nu_q$, genomsnittligt antal i kön,

$W_q(t) = \frac{\int_0^t N_q(\tau)d\tau}{A(t)} \xrightarrow{p} \omega_q$, genomsnittlig tid i kön.

Här betecknar N och N_q två slumpvariabler för vilka gäller att

$$P(N = k) = p_k = \lim_{t\to\infty} P(N(t) = k),$$

$$P(N_q = k) = \lim_{t\to\infty} P(N_q(t) = k).$$

<u>Littles sats</u>:

$\nu = \lambda_A w$

$\nu_q = \lambda_A w_q$

<u>Födelse-/dödsprocess</u>:

födelseintensiteter $\lambda_0, \lambda_1, \lambda_2, \dots$

dödsintensiteter $\mu_1 \mu_2, \mu_3, \dots$

för $k = 1,2,\dots$ kan jämviktssannolikheterna skrivas $p_k = \frac{\lambda_0 \lambda_1 \dots \lambda_{k-1}}{\mu_1 \mu_2 \dots \mu_k} p_0$.