Markovkedjor

- tillståndsrum $S = \{s_1, \dots, s_r\}$ och övergångsmatris $P = (p_{ij})_{i,j=1,\dots,r}$ där
  $p_{ij} = \mathbb{P}(X_{n+1} = s_j \mid X_n = s_i)$ för godtyckligt $n \in \mathbb{N}_0$, $\sum_{j=1}^{r} p_{ij} = 1$ (P radstokastisk)
- flerstegs övergångssats. $\mathbb{P}(X_n = s_j \mid X_0 = s_i) = (P^n)_{ij}$ (element (i,j) i matrispotensen P^n)
- multiplikation från höger! $u^{(n)} := (\mathbb{P}(X_n = s_1), \mathbb{P}(X_n = s_2), \dots, \mathbb{P}(X_n = s_r)) \in [0,1]^r$
  $\rightarrow u^{(n)} = u^{(0)} \cdot P^n$
- MK är irreducibel om $(P^n)_{ij} > 0$ för godtyckliga i,j och något n
  MK är reguljär om $(P^n)_{ij} > 0$ för något fast $n \in \mathbb{N}_0$ och alla i,j
- För en reguljär MK gäller $P^n \to \Pi = \begin{pmatrix} -\pi- \\ \vdots \\ -\pi- \end{pmatrix}$ där $\pi = (\pi_1, \dots, \pi_r)$ kallas gränsfördelning.
- tillstånd s_j absorberande om $p_{jj} = 1$, transient om det har pos. sch. att aldrig återvända
- För en absorberande MK väljer vi kanonisk form $P = \begin{matrix} \\ \text{trans.} \\ \text{abs.} \end{matrix} \overset{\begin{matrix} \text{trans.} & \text{abs.} \end{matrix}}{\left(\begin{array}{c|c} Q & R \\ \hline 0 & I_a \end{array}\right)}$
  $\Rightarrow$ • $P^n \to \left(\begin{array}{c|c} 0 & NR \\ \hline 0 & I_a \end{array}\right)$ där $N = (I_t - Q)^{-1}$ är den fundamentala matrisen
  - n_{ij} = förväntade antalet besök i s_j med start i s_i (där s_i, s_j är transienta)
  - $d = N \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ anger: d_j = tiden till absorption efter start i tillstånd s_j (transient)

Genererande funktioner

$g(x) = \sum_{n=0}^{\infty} a_n x^n$ till en given talföljd $(a_n)_{n \in \mathbb{N}_0}$

viktiga g.f. $g(x) = \sum_{n=0}^{\infty} (cx)^n = \frac{1}{1-cx}$ $[|cx|<1]$ geometriska

$g(x) = \sum_{k=0}^{\infty} \binom{n}{k} x^k = (1+x)^n$ binomial

$g(x) = \sum_{n=0}^{\infty} \binom{n+k}{k} x^n = \frac{1}{(1-x)^{k+1}}$ $[|x|<1]$

- kan användas till att räkna antalet heltalslösningar till $y_1 + y_2 + \dots + y_k = n$ med givna restriktioner [koeff. framför x^n i $g_1(x) \cdot \dots \cdot g_k(x)$ där $g_i(x) = \sum_{\text{m tillåtet värde för } y_i} x^m$]
- kan användas till att lösa rekursioner (ibland tillsammans med partialbråksuppdelning)
  [gångra rekursionen med x^n, summera över tillåtna n, lösa ut $g(x)$, skriva om till potensserie]

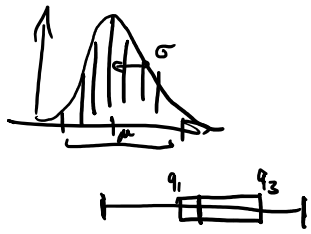
Exponentiella gen. fkt.

$g(x) = \sum_{n=0}^{\infty} a_n \frac{x^n}{n!}$ till en given talföljd $(a_n)_{n \in \mathbb{N}_0}$

- kan användas till att räkna antalet strängar med "bokstäver" ur ett givet alfabet med restriktioner [strategi: samma som ovan fast nu med exp. gen. fkt.]
- vanliga: $e^{cx} = \sum_{n=0}^{\infty} \frac{(cx)^n}{n!}$, $\cosh(x) = \frac{e^x + e^{-x}}{2} = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}$, $\sinh(x) = \frac{e^x - e^{-x}}{2} = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}$

Statistik

Beskrivande S.



Histogram (variationsbredd, rel. frekvenser)
boxplot (median, 1:a och 3:e quartil, iqr)

Inferentiell S.

- Generellt antagande: Stickprov består av n oberoende s.v., alla har samma fördelning som populationen.
- Estimator $\hat{\theta}$ för en parameter θ av populationens fördelning är en s.v. och kan vara väntevärdesriktig ($\mathbb{E}\hat{\Theta} = \theta$) och konsistent ($\text{var}(\hat{\theta}) \to 0$ med $n \to \infty$).
- Stickprovs medelvärdet $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ är väntevärdesriktig (som estimator för μ) och har varians $\frac{\sigma^2}{n}$ där $\sigma^2 = \text{var}(X_i)$, alltså konsistent.
- Stickprovsvariansen $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n\sum_i X_i^2 - (\sum_i X_i)^2}{n(n-1)}$ är väntevärdesriktig estimator för σ^2. ($\mathbb{E}S^2 = \sigma^2$)
- För normalfördelad population (eller stort stickprov, $n \geq 25$) är $\overline{X}$ (approximativt) normalfördelad med väntevärde μ och varians σ^2 (Centrala gränsvärdess.)
  $\Rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- Ett symmetriskt $(1-\alpha)\cdot 100\%$ konfidens intervall för μ ges då av
  $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ där $z_{\alpha/2}$ är sådan att $P(Z > z_{\alpha/2}) = \alpha/2$ ($Z \sim N(0,1)$)
  ensidigt: $\left(-\infty, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$ resp. $\left(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$. (högra, vänstra, svansen)
- För normalfördelad population (eller stor n) är
  $\frac{(n-1)S^2}{\sigma^2}$ (approximativt) χ^2-fördelad med $n-1$ frihetsgrader.
- Ett symmetriskt $(1-\alpha)\cdot 100\%$ konfidensintervall för σ^2 ges då av
  $\left(\frac{(n-1)S^2}{q_{1-\alpha/2}}, \frac{(n-1)S^2}{q_{\alpha/2}}\right)$ där q_α är sådan att $P(X < q_\alpha) = \alpha$. ($X \sim \chi^2_{n-1}$)
  ensidigt: $\left[0, \frac{(n-1)S^2}{q_\alpha}\right)$
- För en normalfördelad population är $\frac{\overline{X} - \mu}{S/\sqrt{n}}$ t-fördelad med $n-1$ frihetsgrader
  Symmetriskt $(1-\alpha)\cdot 100\%$ k.I. för μ:
  $\overline{x} \pm t_{\alpha/2}\cdot\frac{s}{\sqrt{n}}$ där $t_{\alpha/2}$ är sådan att $P(T > t_{\alpha/2}) = \alpha/2$.

Hypotesprövning

nollhypotes: $H_0: \theta = \theta_0$, $\theta \geq \theta_0$ eller $\theta \leq \theta_0$

alternativ: $H_1: \theta \neq \theta_0$, $\theta < \theta_0$, $\theta > \theta_0$

tvåsidigt — ensidigt

- Väljer signifikansnivå och bedömer om nollhypotesen är rimlig eller bör förkastas (typ I fel: H_0 förkastas men stämmer, typ II fel: H_0 behålls dock fel)
- Behöver en teststatistika och dess fördelning (givet antagandet att $\theta = \theta_0$)

vanligast (när det gäller $\theta = \mu$): $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ och $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ t-fördelad med $n-1$ frihetsgrader

- om utfallet av Z inte landar i $(-z_{\alpha/2}, z_{\alpha/2})$ (tvåsidigt)

resp. $(-\infty, z_\alpha)$ eller $(-z_\alpha, \infty)$ (ensidigt) förkastar vi H_0.

Motsvarande regioner för T: $(-t_{\alpha/2}, t_{\alpha/2})$

resp. $(-\infty, t_\alpha)$ eller $(-t_\alpha, \infty)$

- p-värde: minsta signifikansnivå α på vilken H_0 förkastas.

t.ex. för teststatistika Z med utfall z_0 är det

$$p = \begin{cases} 2\,\mathbb{P}(Z \leq z_0) \\ 2\,\mathbb{P}(Z \geq z_0) \end{cases} \text{ om } \begin{matrix} z_0 \leq 0 \\ z_0 > 0 \end{matrix} \text{ (tvåsidigt)} \quad \text{resp.} \quad \begin{matrix} p = \mathbb{P}(Z \leq z_0) \\ p = \mathbb{P}(Z \geq z_0) \end{matrix} \begin{matrix} \text{(vänster)} \\ \text{ensidigt} \\ \text{(höger)} \end{matrix}$$

(samma för andra symmetriskt fördelade teststatistikor, som t.ex. T)

Proportionsskattning

X_j: indikator om element j i stickprovet ingår

- $\hat{p} = \frac{\sum_j X_j}{n}$ är väntevärdesriktig och konsistent estimator av proportionen p.
- Ett symmetriskt $(1-\alpha)100\%$ konfidensintervall för p är $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Vill man ha max. avvikelse d från punktskattningen $\hat{p}$ i denna behövs stickprovsstorlek $n \geq \frac{z_{\alpha/2}^2\, p_0(1-p_0)}{d^2}$ med apriori värde p_0, $n \geq \frac{z_{\alpha/2}^2}{4d^2}$ utan.
- För hypotesprövning används man teststatistikan $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ så länge $\min\{np; n(1-p)\} > 5$
- Jämförelse av två populationer:

$\hat{p}_1 - \hat{p}_2$ approx. $N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$

→ symmetriskt $(1-\alpha)100\%$-KI för $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Jämförelse av medelvärden i två populationer

- $\sigma_1 \neq \sigma_2$: $\bar{X}_1 - \bar{X}_2$ (approximativt) $\sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

  $\Rightarrow$ $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ symmetriskt $(1-\alpha)100\%$-KI för $\bar{X}_1 - \bar{X}_2$.

- $\sigma_1 = \sigma_2 =: \sigma$ (okända)

  kombinerad stickprovsvarians $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

  väntevärdesriktig estimator för σ^2,

  $\frac{n_1+n_2-2}{\sigma^2} S_p^2$ χ^2-fördelad med n_1+n_2-2 frihetsgrader.

  $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$ t-fördelad ——— " ———.

kan användas som teststatistika eller till beräkning av ett (symmetriskt) $(1-\alpha)\cdot 100\%$-KI för $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Linjär regression

X : förklarande variabel, Y : responsvariabel

vill hitta ett linjärt samband $\mu_{Y|x} = \beta_0 + \beta_1 \cdot x$

datamängd $\{(x_i, y_i)\ ;\ i = 1 \dots n\}$

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

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_i Y_i^2 - \frac{1}{n}\left(\sum_i Y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_i x_i Y_i - \frac{1}{n}\left(\sum_i x_i\right)\left(\sum_i Y_i\right)$$

minsta kvadrats estimatorer

$B_1 = \frac{S_{xy}}{S_{xx}}$ är en väntevärdesriktig est. för β_1, $B_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

$B_0 = \bar{Y} - B_1\bar{x}$ ——— " ——— β_0, $B_0 \sim N\left(\beta_0, \frac{\sigma^2}{n S_{xx}}\sum_i x_i^2\right)$

$SSE = \sum_{i=1}^{n}(Y_i - (B_0 + B_1 x_i))^2 = S_{yy} - B_1 S_{xy}$ (sum of squared errors)

$S^2 = \frac{SSE}{n-2}$ är en väntevärdesriktig est. för σ^2.

- $Z = \frac{B_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$ eller $T = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$ t-fördelad med $n-2$ frihetsgrader

  kan användas för att bestämma konfidens intervall

  $B_1 \pm z_{\alpha/2} \frac{\sigma}{\sqrt{S_{xx}}}$ resp. $B_1 \pm t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}$ eller hypotesprövning.

- Samma gäller för $Z = \frac{B_0 - \beta_0}{\sigma \sqrt{\sum_i x_i^2 / n \cdot S_{xx}}} \sim N(0,1)$, $T = \frac{B_0 - \beta_0}{S \cdot \sqrt{\frac{\sum_i x_i^2}{n \cdot S_{xx}}}}$ t-fördelad med $n-2$ frihetsgrader

  $\Rightarrow B_0 \pm z_{\alpha/2} \sigma \sqrt{\frac{\sum_i x_i^2}{n \cdot S_{xx}}}$ resp. $B_0 \pm t_{\alpha/2} S \cdot \sqrt{\frac{\sum_i x_i^2}{n S_{xx}}}$

Centrala gränsvärdessatsen

$X_1, \dots, X_n$ oberoende, samma fördelning med väntevärde μ och varians σ^2

Då är $\bar{X}$ approximativt normalfördelad med väntevärde μ och varians σ^2/n

alltså $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $(n \geq 25)$.

Chebyshev

För X med $\mathbb{E}X = \mu$ och $\text{var}(X) = \sigma^2$ och godtyckligt $\varepsilon > 0$ gäller

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

Stora talens lag

$X_1, \dots, X_n$ oberoende, samma väntevärde μ och varians σ^2, godtyckligt $\varepsilon > 0$

$\Rightarrow \mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \to 0$ med $n \to \infty$.