MS-A0509 Grundkurs i sannolikhetskalkyl och statistik
Exempel etc., del II

G. Gripenberg

Aalto-universitetet

11 februari 2014

Ett konfidensintervall

Antag att 175 *personer blir tillfrågade om de gärna äter en viss grönsak och* 80 *svarar ja. För att bestämma ett konfidensintervall med konfidensgraden* 90% *för att en slumpmässigt vald person gärna äter denna grönsak kan vi gå tillväga på följande sätt:*
De svar vi fått är ett observerat stickprov av en Bernoulli(p) *fördelad slumpvariabel och bestimatornde moment och maximum likelihood-estimatron för* p *är medelvärdet* $\overline{X}$ *som ofta i dessa fall betecknas med* $\hat{p}$.
Den central gänsvärdessatsen säger att

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim_a \mathrm{N}(0,1),$$

eftersom variansen varj X_j *är* $p(1-p)$ *så att variansen av medelvärdet blir* $\frac{p(1-p)}{n}$. *Nu måste vi göra ännu en approximation eftersom vi inte känner till* p *och approximera det talet med* $\hat{p} = \overline{X}$. *Nu är* $z_{0.05} = -F^{-1}_{\mathrm{N}(0,1)}(0.05) = 1.645$ *vilket betyder att*

Ett konfidensintervall, forts.

$$\Pr\left(\frac{\overline{X}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le -1.645\right) \approx 0.05 \quad \textit{och} \quad \Pr\left(\frac{\overline{X}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \ge 1.645\right) \approx 0.05.$$

Detta betyder i sin tur att

$$\Pr\left(p \ge \overline{X} + 1.645 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.05,$$

och

$$\Pr\left(p \ge \overline{X} - 1.645 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.05,$$

Det observerade approximativa konfidensintervallet blir därför

$$\left[0.46 - 1.645\sqrt{\frac{0.46 \cdot 0.54}{175}}, 0.46 + 1.645\sqrt{\frac{0.46 \cdot 0.54}{175}}\right] \approx [0.40, 0.52].$$

Hur får man ett konfidensintervall för σ^2 då $X \sim \mathrm{N}(\mu, \sigma^2)$

Om $X_1, X_2, \ldots, X_n$ är ett stickprov med stickprovsvarians S^2 av en $\mathrm{N}(\mu, \sigma^2)$ fördelad slumpvariabel så är

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Om vi nu vill bestämma ett symmetriskt slumpintervall med konfidensgraden α så skall sannolikheten för värden mindre än den nedre gränsen vara $\frac{1-\alpha}{2}$ och detsamma för värden större än den övre gränsen. Om nu $C \sim \chi^2(n-1)$ så är

$$\Pr\left(C \le F^{-1}_{\chi^2(n-1)}\left(\frac{1-\alpha}{2}\right)\right) = \frac{1-\alpha}{2}, \quad \Pr\left(C \ge F^{-1}_{\chi^2(n-1)}\left(\frac{1+\alpha}{2}\right)\right) = \frac{1-\alpha}{2}.$$

Nu gäller

$$\frac{(n-1)S^2}{\sigma^2} \le F^{-1}_{\chi^2(n-1)}\left(\frac{1-\alpha}{2}\right) \quad \Leftrightarrow \quad \sigma^2 \ge \frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}\left(\frac{1-\alpha}{2}\right)},$$

med ett motsvarande resultat för den andra gränsen.

Oberoende

Antag att en viss sorts frukter indelas i tre olika färgklasser och två olika smakklasser, och att då man undersökt 245 frukter fått följande resultat:

	Färg 1	*Färg 2*	*Färg 3*
Smak 1	*23*	*47*	*38*
Smak 2	*46*	*32*	*59*

Om vi nu vill ha svar på frågan om färg och smak är oberoende av varandra så väljer vi som nollhypotes att de är oberoende och använder testvariabeln

$$C = \sum_{i=1}^{2}\sum_{k=1}^{3} \frac{(O_{i,k} - E_{i,k})^2}{E_{i,k}}.$$

Här är $O_{i,k}$ de observerade antalen, i detta fall tex. $O_{2,3} = 59$ och $E_{i,k}$ är de förväntade antalen när vi antar att nollhypotesen gäller, dvs. $E_{i,k} = \frac{1}{245}\left(\sum_{m=1}^{3} O_{i,m}\right)\left(\sum_{m=1}^{2} O_{m,k}\right)$.

Oberoende, forts.

För att räkna ut testvariabeln och sedan p-värdet kan vi först skriva

```
o=[23 47 38; 46 32 59]
```

sedan får vi antalet n med

```
n= sum(sum(o))
```

de förväntade antalen med `e=sum(o')'*sum(o)/n`

(`sum(o')` *räknar radsummorna och* `sum(o')'` *sätter dessa i en kolumnvektor) och testvariabeln med kommandot*

```
c= sum(sum((o-e).^2./e))
```

Till sist kan vi räkan ut p-värdet med

```
p = 1-chi2cdf(c,(2-1)*(3-1))
```

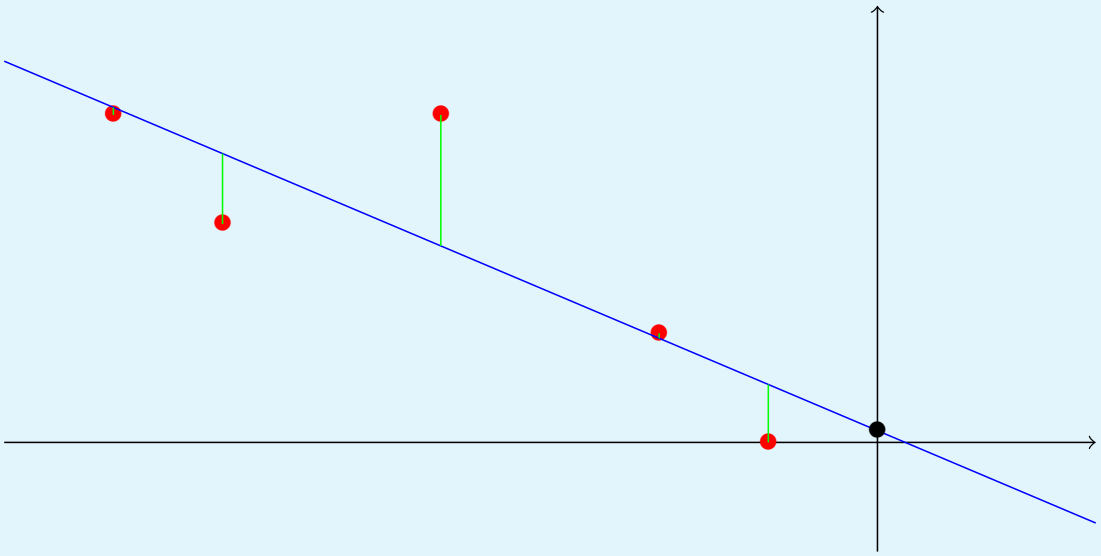
och eftersom detta blir `0.0027480` *kan vi förkasta nollhypotesen på signifikansnivån* 0.5%.

Regression och extrapolering

Anta att vi har följande uppgifter om punkterna (x_j, y_j), $j = 1, \ldots, 5$*:*
`x=[-5 -4 -2 0 1], y=[3 2 3 1 0]`
och vi vill försöka bestämma värdet av y *då* $x = 2$ *och också testa nollhypotesen att detta värde är högst* -1.8*.*
Det första vi gör är att subtrahera 2 från alla x_j *dvs.* `x=x-2;` *så att vi skall bestämma värdet av* y *då* $x = 0$*.*
Sedan räknar vi ut koefficienterna b_0 *och* b_1 *i regressionslinjen* $y \approx b_0 + b_1 x$ *tex. så att vi räknar medelvärdena* $\overline{x} = \frac{1}{5}\sum_{j=1}^{5} x_j = -4$ *och* $\overline{y} = \frac{1}{5}\sum_{j=1}^{5} y_j = 1.8$ *med kommandona* `mean(x)` *och* `mean(y)`*. Sedan räknar vi ut stickprovsvarianserna* $s_x^2 = \frac{1}{4}\sum_{j=1}^{5}(x(j) - \overline{x})^2 = 6.5$ *och* $s_y^2 = \frac{1}{4}\sum_{j=1}^{5}(y_j - \overline{y})^2 = 1.7$ *med kommandona* `var(x)` *och* `var(y)`*. Stickprovskovariansen blir* $s_{xy} = \frac{1}{4}\sum_{j=1}^{5}(x_j - \overline{x})(y_j - \overline{y}) = -2.75$ *vilket i* `ocatve` *fås med kommandot* `cov(x,y)` *medan man i* `matlab` *måste räkna* `c=cov(x,y); c(1,2)`*. Estimaten för regressionskoefficienterna blir nu* $b_1 = \frac{s_{xy}}{s_x^2} = -0.42308$ *och* $b_0 = 0.10769$*.*

Regression och extrapolering, forts.

Om man beaktar förskjutningen av $x = 2$ *till origo ser regressionslinjen ut på följande sätt:*



Regression och extrapolering, forts.

Om vi nu vill testa nollhypotesen att regressionslinjenm skär y-axeln i en punkt som är högst -1.8 *så skall vi också räkna ut restvariansen* $s^2 = frac1n-2\sum_{j=1}^{5}(y_j - b_0 - b_1x_j)^2 = 0.71538$. *Detta kan också göras med hjälp av formeln* $s^2 = \frac{(n-1)s_y^2(1-r_{xy}^2)}{n-2}$ *där* $r_{xy} = s_{xy}\sqrt{\frac{s_x^2}{s_y^2}}$.

I `matlab/octave` *kan detta göras med kommandot* `sum((y-b0-b1*x).^2)/3` *förutsatt att man har räknat ut* `b0` *och* `b1`. *Värdet av testvariabeln blir nu*

$$t = \frac{b_0 - (-1.8)}{\sqrt{s^2(\frac{1}{5} + \frac{\bar{x}^2}{(5-1)s_x^2})}} = 2.4978.$$

Eftersom testvariabelns fördelning är $\mathrm{t}(5-2)$ *och noolhypotesen* $\beta_0 \leq -1.8$ *innebär att negativa värden för testvariabeln bekräftar nollhypotesen blir p-värdet*

$$p = \Pr(T \geq 2.4978) = 1 - F_{\mathrm{t}(3)}(2.4978) = F_{\mathrm{t}(3)}(-2.4978) = 0.43939,$$

vilket betyder att nollhypotesen kan förkastas på signifikansnivån 5%.

Regression och extrapolering, forts.

Ett annat sätt att komma till samma resultat är att konstatera att $\mathrm{t}_{0.05}(3) = F^{-1}_{\mathrm{t}(3)}(0.95) = 2.3534$ *vilket man också hittar i tabellen (men observera att vi här har ett ensidigt alternativ).*

TAULUKKO 2. ***t*-DISTRIBUTION *t*(*df*)**
TABLE 2. ***t*-JAKAUMA *t*(*df*)**

Kriittisiä arvoja / Critical values

	Merkitsevyystaso 1-suuntaisissa testeissä / Sigficance level in 1-side						
df	0.4	0.3	0.2	0.1	0.05	0.025	0.01
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143

Eftersom testvaribeln är större än detta kritiska värde förkastas nollhypotesen på signifikansnivån 5%.