Konfidensintervall

Notat til STK1110

Ørnulf Borgan, Ingrid K. Glad og Anders Rygh Swensen
Matematisk institutt, Universitetet i Oslo

August 2007

Formål

En vanlig metode for å angi usikkerheten til et estimat er å beregne et *konfidensintervall*. Hvordan dette gjøres er beskrevet på sidene 217–219 i læreboka til Rice. Siden dette er midt i et kapittel om utvalgsteori, som ikke er pensum i STK1110, tar vi opp konfidensintervall i dette notatet.

Generelt

Anta at en ukjent parameter θ blir estimert med estimatoren $\widehat{\theta}$. Et konfidensintervall er et (stokastisk) intervall $[\widehat{\theta}_L, \widehat{\theta}_U]$, beregnet på grunnlag av de samme observasjonene som $\widehat{\theta}$, som inneholder parameteren med gitt sannsynlighet:

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1 - \alpha.$$

Sannsynligheten $1 - \alpha$ kalles *konfidenskoeffisienten* til intervallet. Vi sier også at intervallet $[\widehat{\theta}_L, \widehat{\theta}_U]$ er et $100(1-\alpha)\%$ konfidensintervall. Vanlige valg av α er $0.10, 0.05$ og 0.01 svarende til konfidensintervall med konfidenskoeffisient $90\%, 95\%$ og 99%.

Vi vil se hvordan vi kan lage konfidensintervall for noen vanlige fordelinger.

Poisson fordelingen

Anta at $X_1, \ldots, X_n$ er uavhengige og Poissonfordelte med parameter λ. Da er

$$\widehat{\lambda} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1}$$

en forventningsrett estimator for λ med standardfeil

$$\sigma_{\widehat{\lambda}} = \sqrt{\frac{\lambda}{n}}$$

og estimert standardfeil

$$s_{\widehat{\lambda}} = \sqrt{\frac{\widehat{\lambda}}{n}} \tag{2}$$

(jfr. eksempel A i avsnitt 8.4 i læreboka). Hvis $n\lambda$ er "stor" (i praksis minst 15), vil

$$\frac{\sum_{i=1}^{n} X_i - n\lambda}{\sqrt{n\lambda}} = \frac{\widehat{\lambda} - \lambda}{\sigma_{\widehat{\lambda}}}$$

være tilnærmet standard normalfordelt (jfr. eksempel A i avsnitt 5.3 i læreboka). En kan vise at også

$$\frac{\widehat{\lambda} - \lambda}{s_{\widehat{\lambda}}} \tag{3}$$

er tilnærmet standard normalfordelt.

Vi vil bruke (3) til å bestemme et tilnærmet $100(1-\alpha)\%$ konfidensintervall for λ. Dette kan gjøres på følgende måte. La z_p være p-fraktilen i standard normalfordelingen. Siden $z_{\alpha/2} = -z_{1-\alpha/2}$ har vi at

$$P\left(-z_{1-\alpha/2} \leq \frac{\widehat{\lambda} - \lambda}{s_{\widehat{\lambda}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha. \tag{4}$$

Ved å løse ulikhetene ser vi at begivenheten

$$-z_{1-\alpha/2} \leq \frac{\widehat{\lambda} - \lambda}{s_{\widehat{\lambda}}} \leq z_{1-\alpha/2}$$

er den samme som

$$\widehat{\lambda} - z_{1-\alpha/2}\, s_{\widehat{\lambda}} \leq \lambda \leq \widehat{\lambda} + z_{1-\alpha/2}\, s_{\widehat{\lambda}}.$$

Dermed gir (4) at

$$P\left(\widehat{\lambda} - z_{1-\alpha/2}\, s_{\widehat{\lambda}} \leq \lambda \leq \widehat{\lambda} + z_{1-\alpha/2}\, s_{\widehat{\lambda}}\right) \approx 1 - \alpha.$$

Ved å velge

$$\widehat{\lambda}_L = \widehat{\lambda} - z_{1-\alpha/2}\, s_{\widehat{\lambda}}$$
$$\widehat{\lambda}_U = \widehat{\lambda} + z_{1-\alpha/2}\, s_{\widehat{\lambda}}$$

ser vi at $[\widehat{\lambda}_L, \widehat{\lambda}_U]$ er et (stokastisk) intervall som med sannsynlighet tilnærmet lik $1-\alpha$ inneholder λ. Vi har altså funnet at

$$\widehat{\lambda} \pm z_{1-\alpha/2}\, s_{\widehat{\lambda}} \tag{5}$$

er et tilnærmet $100(1-\alpha)\%$ konfidensintervall for λ. Merk $1-\alpha/2$-fraktilen i standard normalfordelingen enkelte ganger skrives som $z(\alpha/2)$ i boka til Rice. Altså har vi $z(\alpha/2) = z_{1-\alpha/2}$.

Anencefali er en medfødt, sjelden misdannelse med manglende utvikling av hjernen. Barn som har denne misdannelsen er oftest dødfødt eller dør kort etter fødselen. I tabellen nedenfor har vi gitt forekomsten av anencefali i Edinburgh hver måned i perioden 1956-1966.

Antall tilfeller x	0	1	2	3	4	5	6	7	8
Antall måneder med x tilfeller	18	42	34	18	11	6	0	2	1

Av tabellen ser vi at det i 18 av de 132 månedene var ingen tilfeller av anencefali, i 42 av månedene var det ett tilfelle, osv.

Siden anencefali er en sjelden misdannelse, er det rimelig å anta at antall tilfeller av anencefali i løpet av en måned er Poisson fordelt med parameter λ (= det forventede antall tilfeller per måned). Ved å bruke (1) finner vi estimatet

$$\widehat{\lambda} = \frac{0 \cdot 18 + 1 \cdot 42 + 2 \cdot 34 + 3 \cdot 18 + 4 \cdot 11 + 5 \cdot 6 + 6 \cdot 0 + 7 \cdot 2 + 8 \cdot 1}{132} = 1.97$$

for det forventede antall tilfeller av anencefali per måned. Estimatet (2) for standardfeilen er $\sqrt{1.97/132} = 0.12$. Ved å sette disse estimatene inn i (5) får vi at et 95% konfidensintervall blir $1.97 \pm 1.96 \cdot 0.12$, det vil si fra 1.73 til 2.21 tilfeller per måned.

Normalfordelingen

La nå $X_1, \ldots, X_n$ være uavhengige og normalfordelte $N(\mu, \sigma^2)$. Da er

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

en forventningsrett estimator for μ med standardfeil

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}.$$

Hvis variansen σ^2 hadde vært kjent, kunne vi brukt at

$$\frac{\widehat{\mu} - \mu}{\sigma_{\hat{\mu}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

er standard normalfordelt og funnet et $100(1-\alpha)\%$ konfidensintervall for μ fra

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

jfr. (4) i eksempelet ovenfor. Intervallet blir da

$$\overline{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Variansen er imidlertid som oftest ukjent og må estimeres fra observasjonene med

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

se avsnitt 6.3 i boken til Rice. Fra korollar B s. 198 i Rice har vi at

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

er t-fordelt med $n-1$ frihetsgrader. La nå $t_{n-1,p}$ være p-fraktilen i en t-fordeling med $n-1$ frihetsgrader. Vi har, tilsvarende som i normalfordelingen, at $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ og kan derfor skrive

$$P\left(-t_{n-1,1-\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha. \tag{6}$$

Ved å løse ulikheten på samme måte som ovenfor, gir (6)

$$P\left(\overline{X} - t_{n-1,1-\alpha/2}\,\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}\,\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

slik at

$$\overline{X} \pm t_{n-1,1-\alpha/2}\,\frac{S}{\sqrt{n}} \tag{7}$$

er et $100(1-\alpha)\%$ konfidensintervall for μ. Merk $1-\alpha/2$-fraktilen i t-fordelingen med $n-1$ frihetsgrader enkelte ganger skrives som $t_{n-1}(\alpha/2)$ i boka til Rice. Altså har vi $t_{n-1}(\alpha/2) = t_{n-1,1-\alpha/2}$.

I forbindelse med produksjon av en spesiell type latex-maling, måles tørketiden til malingen (i timer) i $n = 15$ forsøk:

3.4	2.5	3.8	2.9	3.6
2.8	3.3	3.6	3.7	3.8
4.4	3.0	3.2	3.0	3.8

På bakgrunn av disse målingene ønsker vi å finne et 90% konfidensintervall for forventet tørketid for denne malingstypen.

Det er rimelig å anta at målingene er normalfordelte. For et 90% konfidensintervall er $\alpha = 0.10$ og vi trenger derfor fraktilen $t_{14,0.95} = 1.761$ fra tabell 4 (appendiks B) i Rice. Ved å bruke MATLAB eller en lommeregner finner vi at $\overline{X} = 3.387$ timer og $S^2 = 0.246$.

Et 90% konfidensintervall for forventet tørketid for latex-malingen blir da ved (7)

$$3.387 \pm 1.761\frac{\sqrt{0.246}}{\sqrt{15}},$$

som blir fra 3.16 til 3.61 timer.

I de to eksemplene ovenfor har konfidensintervallene (5) og (7) formen

$$\text{estimator} \pm \text{fraktil} \cdot (\text{estimator av standardfeilen})$$

Slik vil det være i mange andre situasjoner der vi har symmetriske fordelinger. Men som det neste eksemplet viser, er det ikke alltid slik.

Eksponensialfordelingen

Anta at $T_1, \ldots, T_n$ er uavhengige og eksponensialfordelte med forventning τ. De har da sannsynlighetstettheten

$$f(t) = \begin{cases} \frac{1}{\tau}e^{-t/\tau} & \text{for } t > 0\,, \\ 0 & \text{ellers.} \end{cases}$$

Siden $\mathrm{E}T_i = \tau$, er

$$\widehat{\tau} = \overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i \tag{8}$$

en forventningsrett estimator for τ.

Vi ønsker å bestemme et $100(1-\alpha)\%$ konfidensintervall for τ. For å kunne gjøre det må vi først bestemme fordelingen til (en funksjon av) $\widehat{\tau}$. Det er lett å vise at $2T_i/\tau$ er gammafordelt med formparameter 1 og skalaparameter 1/2. (Gjør det!) Det følger at

$$\sum_{i=1}^{n} \frac{2T_i}{\tau} = \frac{2n\overline{T}}{\tau} = 2n\frac{\widehat{\tau}}{\tau} \tag{9}$$

er gammafordelt med formparameter n og skalaparameter 1/2 (jfr. eksempel F i avsnitt 4.5 i læreboka). Denne fordelingen kalles kjikvadratfordelingen med $2n$ frihetsgrader.
La x_p^2 være p-fraktilen i kjikvadratfordelingen med $2n$ frihetsgrader. Da gir (9) at

$$P\left(x_{\alpha/2}^2 \leq 2n\frac{\widehat{\tau}}{\tau} \leq x_{1-\alpha/2}^2\right) = 1-\alpha.$$

Ved å omforme ulikhetene har vi ekvivalent at

$$P\left(\frac{2n\widehat{\tau}}{x_{1-\alpha/2}^2} \leq \tau \leq \frac{2n\widehat{\tau}}{x_{\alpha/2}^2}\right) = 1-\alpha.$$

Dette betyr at

$$\left[\frac{2n\widehat{\tau}}{x_{1-\alpha/2}^2}, \frac{2n\widehat{\tau}}{x_{\alpha/2}^2}\right] \tag{10}$$

er et $100(1-\alpha)\%$ konfidensintervall for τ.

Kreftpasienter som gjennomgår behandling (stråling eller kjemoterapi) kan mange ganger oppnå symptomfrihet uten at sykdommen helbredes. De kan da senere få tilbakefall slik at symptomene igjen blir manifeste. I tabellen nedenfor er det gitt tid til tilbakefall (i uker) for 21 pasienter med akutt leukemi. (Tallene er fra 1960-tallet og er ikke representative for dagens leukemibehandling.)

1	1	2	2	3	4	4	5	5	8	8
8	8	11	11	12	12	15	17	22	23	

For disse dataene (og i mange andre tilfeller) kan tid til tilbakefall beskrives godt ved eksponensialfordelingen. Vi kan derfor betrakte tidene i tabellen som realisasjoner av 21 uavhengige eksponensialtfordelte T_i-er. Av (8) er et estimat for forventet tid til tilbakefall

$$\widehat{\tau} = \frac{1}{21}(1+1+2+2+3+\cdots+22+23) = 8.67.$$

La oss også bestemme et 90% konfidensintervall. Vi ser av (10) at vi da trenger 5% og 95% fraktilene i kjikvadratfordelingen med $2 \cdot 21 = 42$ frihetsgrader. Av MATLAB finner vi at disse fraktilene er henholdsvis 28.1 og 58.1. (Tabell 3 i appendiks B i læreboka gir fraktiler i kjikvadratfordelingen for utvalgte antall frihetsgrader.) Dermed blir et 90% konfidensintervall for forventet tid til tilbakefall

$$\left[\frac{2 \cdot 21 \cdot 8.67}{58.1}, \frac{2 \cdot 21 \cdot 8.67}{28.1}\right],$$

det vil si fra 6.27 til 12.9 uker.

Oppsummering

La oss til slutt oppsummere noen egenskaper ved konfidensintervall som det er verd å merke seg:

(i) Et konfidensintervall $[\widehat{\theta}_L, \widehat{\theta}_U]$ er et intervall med grenser som er stokastiske variable. Det er konstruert for å inneholde en ukjente parameter θ.

(ii) Sannsynligheten

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1 - \alpha.$$

(eventuelt $\approx 1 - \alpha$) kan tolkes som en relativ frekvens av et stort antall forsøk. Den gir andelen av forsøkene der intervallet inneholder θ.

(iii) Etter at intervallgrensene er regnet ut på grunnlag av faktiske observasjoner, er det beregnede intervallet en realisasjon av det stokastiske intervallet. Det har da ingen mening å snakke om sannsynligheten for at intervallet inneholder θ. For enten ligger θ i intervallet eller så gjør den ikke det (selv om vi ikke kan vite hva som er tilfellet). Dette er som for andre sannsynlighetsutsagn (så lenge vi holder oss til et frekvensbasert sannsynlighetsbegrep). Det er for eksempel ikke meningsfylt å snakke om sannsynligheten for at vi fikk sekser etter at vi har kastet en terning. For enten fikk vi en sekser eller så fikk vi ikke det.