Föreläsning 11

Statistisk inlärning och dataanalys (Kungliga Tekniska Högskolan)

Föreläsning 11

11.1 Intervallskattning

Istället för att skatta värdet av ett visst parameter θ via en punktskattning $W(\boldsymbol{x})$ eller att testa om parametern har ett visst värdet θ_0 via ett hypotesttest kan vi istället bestämma en delmängd $C(\boldsymbol{x})$ av parameterrummet Ω som vi förvantar att innehålla θ med stor sannolikhet. Det vill säger, givet ett stickprov $\boldsymbol{x} = (x_1, \ldots, x_n)$ från ett $f_X(x|\theta)$-population skulle vi vilja uppskatta en delmängd $C(\boldsymbol{x})$ av parameterrummet Ω som innehåller θ med stor sannolikhet. Delmängden $C(\boldsymbol{x})$ kallas för en *mängdskattning*.

Det händer ofta att $C(\boldsymbol{x})$ är en intervall men det behövs inte att vara så. I situationerna där $C(\boldsymbol{x})$ är en intervall kallas $C(\boldsymbol{x})$ för en *intervallskattning*. När $C(\boldsymbol{x})$ är en intervallskattning kan det skrivas som

$$C(\boldsymbol{x}) = [L(\boldsymbol{x}), U(\boldsymbol{x})]$$

där $L(\boldsymbol{x})$ och $U(\boldsymbol{x})$ är funktioner som uppfyller $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ för alla $\boldsymbol{x}$. Vi skriver $C(\boldsymbol{x})$ som en sluten intervall men det kan också vara öppen. Det kan även händer att $L(\boldsymbol{x}) = -\infty$ eller $U(\boldsymbol{x}) = \infty$. Sådana intervallskattningar kallas för *ensidig* (eller *asymmetrisk*) *intervallskattningar*. När $L(\boldsymbol{x}) > -\infty$ och $U(\boldsymbol{x}) < \infty$ kallas $C(\boldsymbol{x})$ för en *tvåsidig intervallskattning*.

Vi skulle vilja ha att sannolikheten att $\theta \in C(\boldsymbol{x})$ är stor. Så behöver vi att beräkna *täckningssannolikheten*

$$P(\theta \in C(\boldsymbol{x})|\theta).$$

För att säkerställa tillräcklig täckning behöver vi att täckningssannolikhet minst en viss nivå. Givet $\alpha \in (0, 1)$ har vi en $1 - \alpha$ *konfidensmängd* (eller en $1 - \alpha$ *konfidensintervall* när $C(\boldsymbol{x})$ är en intervall) när

$$P(\theta \in C(\boldsymbol{x})|\theta) \geq 1 - \alpha \qquad \text{för alla } \theta \in \Omega.$$

Värdet $1 - \alpha$ kallas för *konfidensgraden* av $C(\boldsymbol{x})$.

Exempel 11.1. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{U}(0, \theta)$-fördelade. Låt $Y = \max_i(x_i)$. Vi betraktar två kandidater för intervallskattningar:

$$[aY, bY] \qquad \text{och} \qquad [Y + c, Y + d]$$

där $1 \leq a < b$ och $0 \leq c < d$. Notisera att $\theta \geq Y$ eftersom $P(Y > \theta) = 0$. För den första kandidat kan vi beräkna täckningssannolikheten som

$$P(\theta \in [aY, bY]|\theta) = P(aY \leq \theta \leq bY|\theta) = P\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right).$$

Vi har sett redan i Exempel 5.6 att $Y = \max_i(X_i)$ här täthetsfunktion

$$f_Y(y|\theta) = n\frac{y^{n-1}}{\theta^n}.$$

Så följer det att $T = Y/\theta$ här täthetsfunktionen

$$f_T(t|\theta) = nt^{n-1}.$$

Därför har intervallen $[aY, bY]$ täckningssannolikhet

$$P\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) = \int_{1/b}^{1/a} nt^{n-1}dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.$$

Detta är konstant på θ. Så det följer att intervallen $[aY, bY]$ har konfidensgrad $a^{-n} - b^{-n}$.

Å andra sidan har intervallen $[Y + c, Y + d]$ täckningssannolikheten

$$\begin{aligned} P(\theta \in [Y + c, Y + d]|\theta) &= P(Y + c \leq \theta \leq Y + d|\theta), \\ &= P\left(1 - \frac{d}{\theta} \leq \frac{Y}{\theta} \leq 1 - \frac{c}{\theta}|\theta\right), \\ &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n. \end{aligned}$$

Detta täckningssannolikhet beror på θ och vi kan se att

$$\lim_{\theta \to \infty} \left(\left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n\right) = 0.$$

Det följer att konfidensgraden för $[Y + c, Y + d]$ är 0. Uppenbarligen är vissa konfidensmängder bättre än andra.

I det föregående exemplet var våra två kandidater för konfidensintervaller ganska intuitiva att tänka ut. Vi skulle vilja naturligtvis ha metoder för att bestämma konfidensmängder i generellt sett.

11.1.1 Inversionen av en testvariabel

Ett sätt att konstruera en konfidensmängd (eller konfidensintervall) är att använda ett hypotestest.

Exempel 11.2. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{N}(\theta, \sigma^2)$-fördelade där σ^2 är känt. Vi såg i Exempel 10.4 (för $\sigma^2 = 1$) att ett α-nivå test för

$$H_0 : \theta = \theta_0 \qquad \text{mot} \qquad H_1 : \theta \neq \theta_0$$

ges av

$$\text{Förkasta } H_0 : \theta = \theta_0 \text{ om } |\bar{x} - \theta_0| > z_{\alpha/2}\sigma/\sqrt{n}.$$

Eftersom testet har nivå α gäller det att

$$P(\text{acceptera } H_0|\theta_0) = 1 - P(\text{förkasta } H_0|\theta_0) \geq 1 - \alpha.$$

Vi kan också skriva $P(\text{acceptera } H_0|\theta_0)$ som

$$\begin{aligned} P(\text{acceptera } H_0|\theta_0) &= P(|\bar{X} - \theta_0| < z_{\alpha/2}\sigma/\sqrt{n}|\theta_0), \\ &= P(-z_{\alpha/2}\sigma/\sqrt{n} < \bar{X} - \theta_0 < z_{\alpha/2}\sigma/\sqrt{n}|\theta_0), \\ &= P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \theta_0 < \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}|\theta_0). \end{aligned}$$

Eftersom testet har nivå α följer det att

$$P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \theta_0 < \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}|\theta_0) \geq 1 - \alpha.$$

Därför är

$$\left[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right]$$

en $1 - \alpha$ konfidensintervall.

Det föregående exemplet generaliseras. Givet ett test med kritiskt område $R = \{\boldsymbol{x} : \text{testet förkastar } H_0 : \theta = \theta_0\}$ kallas komplementet av R

$$A(\theta_0) = \{\boldsymbol{x} : \text{testet accepterar } H_0 : \theta = \theta_0\},$$

för *acceptansområdet* för testet. I Exempel 11.2 inverterade vi det α-nivå testet – eller verkligen inverterade vi dess acceptansområde) för att få en $1-\alpha$ konfidensmängd. Vi har i exemplet en korrespondens mellan acceptansområdet

$$A(\theta_0) = \{\boldsymbol{x} : \theta_0 - z_{\alpha/2}\sigma/\sqrt{n} < \bar{x} < \theta_0 + z_{\alpha/2}\sigma/\sqrt{n}\}$$

och konfidensmängden

$$C(\boldsymbol{x}) = \{\theta : \bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \theta_0 < \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\}$$

Denna korrespondens ges av

$$\boldsymbol{x} \in A(\theta_0) \qquad \text{om och endast om} \qquad \theta_0 \in C(\boldsymbol{x}).$$

Både mängder $A(\theta_0)$ och $C(\boldsymbol{x})$ försöker att svara det samma frågan: "Vilka parametrar θ överensstämmer med datan $\boldsymbol{x}$?" Denna korrespondens stämmer i generellt sett:

Sats 11.1. *Låt $A(\theta_0)$ vara acceptansområdet för ett test med nivå α för $H_0 : \theta = \theta_0$. För varje $\boldsymbol{x}$ definierar vi*

$$C(\boldsymbol{x}) = \{\theta \in \Omega : \boldsymbol{x} \in A(\theta)\}.$$

Så gäller det att $C(\boldsymbol{x})$ är en $1-\alpha$ konfidensmängd.

Bevis. Eftersom $A(\theta_0)$ är acceptansområdet för ett test med nivå α har vi att

$$P(\boldsymbol{X} \in A(\theta_0)|\theta_0) = 1 - P(\boldsymbol{X} \notin A(\theta_0)|\theta_0) \geq 1 - \alpha.$$

Det följer att

$$P(\theta_0 \in C(\boldsymbol{X})|\theta_0) = P(\boldsymbol{X} \in A(\theta_0)|\theta_0) \geq 1 - \alpha.$$

Därför är $C(\boldsymbol{X})$ en $1-\alpha$ konfidensmängd. □

Notisera att Sats 11.1 specificerar inte alternativhypotesen. Det kunde vara, exempelvis, $H_1 : \theta = \theta_1$, $H_1 : \theta \neq \theta_0$, $H_1 : \theta > \theta_0$ eller $H_1 : \theta < \theta_0$. Valet för alternativhypotesen kan påverka formen av $C(\boldsymbol{X})$ – ofta i användbara sätt.

Exempel 11.3. Låt $X_1, \ldots, X_n$ vara oberoende och Exp(λ)-fördelade där vi ta skalaparametriseringen av en exponentialfördelningen:

$$f_X(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}.$$

Vi kan testa

$$H_0 : \lambda = \lambda_0 \qquad \text{mot} \qquad H_1 : \lambda \neq \lambda_0$$

med en likelihood-kvottest ges av

$$\text{Förkasta } H_0 \text{ om } \frac{L(\lambda_0|\boldsymbol{x})}{L(\hat{\lambda}|\boldsymbol{x})} \leq c$$

för någon $c \in (0,1)$ där $\hat{\lambda}$ betecknas ML-skattningen av λ. Man kan verifiera att $\hat{\lambda} = \bar{x}$. Därför kan vi förenkla testet till

$$\text{Förkasta } H_0 \text{ om } \left(\frac{\sum_{i=1}^n x_i}{\lambda_0}\right)^n e^{-\left(\sum_{i=1}^n x_i\right)/\lambda_0} \leq k$$

för någon k som är vald för att ge ett test med nivå α. Acceptansområdet av testet är då

$$A(\lambda_0) = \left\{\boldsymbol{x} : \left(\frac{\sum_{i=1}^n x_i}{\lambda_0}\right)^n e^{-\left(\sum_{i=1}^n x_i\right)/\lambda_0} \geq k\right\}.$$

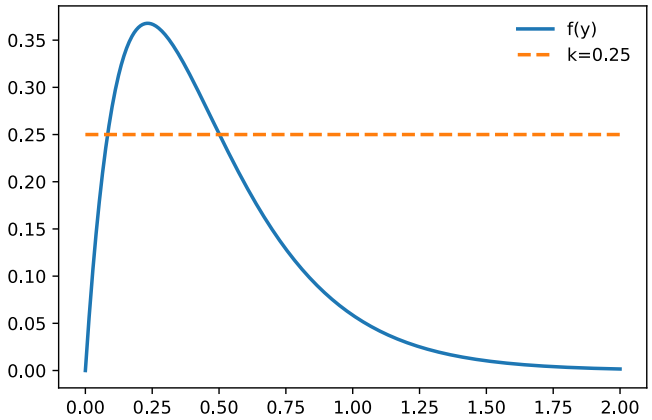
Det följer från Sats 11.1 att

$$C(\boldsymbol{x}) = \left\{\lambda > 0 : \left(\frac{\sum_{i=1}^n x_i}{\lambda}\right)^n e^{-\left(\sum_{i=1}^n x_i\right)/\lambda} \geq k\right\}.$$

Men hur ser mängden $C(\boldsymbol{x})$ ut? Notisera först att mängden beror bara på $\boldsymbol{x}$ genom $y = \sum_{i=1}^n x_i$ för $y \in [0, \infty)$ (eftersom y är en tillräcklig statistika för λ). Om vi plottar funktionen

$$f(\lambda) = \left(\frac{y}{\lambda}\right)^n e^{-y/\lambda}$$

för, exempelvis $n = 10$, $y = 4.29$ och $k = 0.25$ ser vi



Det följer att $C(\boldsymbol{x})$ är mängden som består av alla λ så att värdet $f(\lambda)$ är över den orange linjen i bilden. Vi ser att området har även en enda lokal maximum (dvs., $f(\lambda)$ är *unimodal*) och därför måste vara en intervall av formen

$$C(\boldsymbol{x}) = [Ly), U(y)]$$

där

$$\left(\frac{y}{L(y)}\right)^n e^{-y/L(y)} = k = \left(\frac{y}{U(y)}\right)^n e^{-y/U(y)}.$$

Om vi låter $L(y) = y/a$ och $U(y) = y/b$ får vi ut begränsningen

$$a^n e^{-a} = b^n e^{-b}.$$

Detta är bara en begränsning men täckningssannolikheten är en andra:

$$\begin{aligned} P(\lambda_0 \in C(\boldsymbol{x})|\lambda_0) &= P(L(y) \leq \lambda_0 \leq U(y)|\lambda_0), \\ &= P\left(\frac{y}{a} \leq \lambda_0 \leq \frac{y}{b}|\lambda_0\right), \\ &= P(b \leq \frac{y}{\lambda_0} \leq a|\lambda_0) \geq 1-\alpha. \end{aligned}$$

Eftersom $X_1, \ldots, X_n$ är exponential i skalaparametrisering kan man visa att $Y/\lambda \sim \text{Gamma}(n, 1)$. Så har $Z = Y/\lambda$ täthetsfunktionen

$$f_Z(z|\lambda) = \frac{1}{\Gamma(n)} z^{n-1} e^{-z}.$$

Det följer att a och b bestämts genom att löser systemet av ekvationer

$$\begin{aligned} a^n e^{-a} &= b^n e^{-b}, \\ \int_b^a \frac{1}{\Gamma(n)} z^{n-1} e^{-z} dz &= 1-\alpha. \end{aligned}$$

Detta system har ingen sluten form lösning men kan vara lösas numeriskt. Med hjälp av partialintegration kan man visa att

$$\int_b^a \frac{1}{\Gamma(n)} z^{n-1} e^{-z} dz = e^{-b} \sum_{k=0}^{n-1} \frac{b^k}{k!} - e^{-a} \sum_{k=0}^{n-1} \frac{a^k}{k!}.$$

För $n = 2$ får vi då

$$e^{-b}(b+1) - e^{-a}(a+1).$$

I detta fall skulle vi lösa systemet

$$\begin{aligned} a^2 e^{-a} &= b^2 e^{-b}, \\ e^{-b}(b+1) - e^{-a}(a+1) &= 1-\alpha. \end{aligned}$$

Om vi skulle vilja ha en 0.9 konfidensintervall tar vi $\alpha = 0.1$ och får

$$a \approx 5.480 \qquad \text{och} \qquad b \approx 0.441.$$

Det följer att en 0.9 konfidensintervall ges av

$$\left[\frac{x_1 + x_2}{5.480}, \frac{x_1 + x_2}{0.441}\right].$$

Här använde vi en likelihood-kvottest för att bestämma konfidensintervallen. Dessa brukar ha acceptansområde av formen

$$A(\theta_0) = \left\{\boldsymbol{x} : \frac{L(\theta_0|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})} \geq k(\theta_0)\right\}$$

för någon funktion $k(\theta_0)$ av θ_0. Om vi tar $k'(\theta_0, \boldsymbol{x}) = k(\theta_0) L(\hat{\theta}|\boldsymbol{x})$ får vi

$$A(\theta_0) = \{\boldsymbol{x} : L(\theta_0|\boldsymbol{x}) \geq k'(\theta_0, \boldsymbol{x})\}.$$

När funktionen $k'(\theta_0, \boldsymbol{x})$ är konstant på θ_0 har vi att $A(\theta_0)$ består av alla θ som har likelihood större än något värde bestämts av datan $\boldsymbol{x}$. I detta fall har vi ett område av högsta likelihood som en konfidensmängd – vilken kan vara en användbara interpretation av en konfidensmängd. Detta ska hända för vissa tester men inte alla.

Nu betraktar vi ett exempel av en ensidig konfidensintervall.

Exempel 11.4. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{N}(\theta, \sigma^2)$-fördelade med både θ och σ^2 okända. Vi skulle vilja ha en övre gräns för θ; dvs. vi skulle vilja ha en konfidensintervall av formen $(-\infty, U(\boldsymbol{x})]$.

Vi ska invertera ett test för $H_0 : \theta = \theta_0$ mot $H_1 : \theta < \theta_0$. Ett sådant test ska förkasta H_0 när data $\boldsymbol{x}$ föreslår låg värden för μ och därför acceptansområde ska består av $\boldsymbol{x}$ som föreslår hög värden för μ. Därför acceptansområdet ska vara avgränsad underifrån. Det följer att konfidensintervallen ska vara avgränsad ovan.

Ett likelihood-kvottest för $H_0 : \theta = \theta_0$ mot $H_1 : \theta < \theta_0$ är

$$\text{Förkasta } H_0 \text{ om } \tfrac{\bar{x}-\theta_0}{s/\sqrt{n}} < -t_\alpha(n-1).$$

Testet har därmed acceptansområdet

$$A(\theta_0) = \{\boldsymbol{x} : \bar{x} \geq \theta_0 - t_\alpha(n-1)s/\sqrt{n}\}.$$

Det följer att en $1-\alpha$ konfidensintervall är

$$C(\boldsymbol{x}) = (-\infty, \bar{x} + t_\alpha(n-1)s/\sqrt{n}].$$

11.1.2 Pivotvariabel

Ett annat sätt för att konstruera intervallskattningar är genom *pivotvariabler*. En *pivotvariabel* är en stokastisk variabel $Q(\boldsymbol{X}, \theta)$ som beror inte på θ. Givet en mängd A beror inte sannolikheten $P(Q(\boldsymbol{X}, \theta) \in A|\theta)$ på θ. Vi kan därför fixa mängden A oberoende av θ och invertera A för att få en konfidensmängd

$$C(\boldsymbol{x}) = \{\theta : Q(\boldsymbol{x}, \theta) \in A\}.$$

För läge-skala familjer har vi enkla pivotvariabler:

- läge familj: $\{f(x-\mu) : \mu\} \longrightarrow Q(\boldsymbol{X}, \mu) = \bar{X} - \mu$.
- skala familj: $\{(1/\sigma)f(x/\sigma) : \sigma > 0\} \longrightarrow W(\boldsymbol{X}, \sigma) = \bar{X}/\sigma$.
- läge-skala familj: $\{(1/\sigma)f((x-\mu)/\sigma) : \sigma > 0, \mu\} \longrightarrow Q(\boldsymbol{X}, (\mu, \sigma)) = (\bar{X} - \mu)/\sigma$.

Exempel 11.5. Vi återgår till vårt exempel (Exempel 11.3) där vi hade $X_1, \ldots, X_n$ oberoende och exponentialfördelade i skalaparameteriseringen med skala parameter λ. Vi såg att $Y = \sum_{i=1}^n X_i \sim \mathrm{Gamma}(n, \lambda)$ och $Y/\lambda \sim \mathrm{Gamma}(n, 1)$ (där vi använder den skala parametriseringen av gammafördelningar). Vi kan transformera detta till en χ^2-fördelning om vi notiserar att

$$\mathrm{Gamma}(n, 2) = \chi^2(2n). \qquad \text{(skala parametrisering av gamma)}$$

Det följer att

$$T = \frac{2Y}{\lambda} \sim \chi^2(2n),$$

vilken beror bara på n, antalet utfall, som är känt. Därför är T en pivotvariabel.

I Exempel 11.5 hittade vi en pivotvariabel genom en transformation av stokastiska variabler. Generellt sett, om täthetsfunktionen av en statistika W kan skrivas som

$$f_W(w|\theta) = g(Q(w, \theta)) \left| \frac{d}{dw} Q(w, \theta) \right|$$

för $g : \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}$ och $Q : \mathbb{R}^2 \longrightarrow \mathbb{R}$ som är monoton på θ är $Q(W, \theta)$ en pivotvariabel engligt en transformation av variabler.

För att få en $1-\alpha$ konfidensintervall från en pivotvariabel kan vi bestämma a och b så att

$$P(a \leq Q(\boldsymbol{X}, \theta) \leq b) \geq 1 - \alpha.$$

Konfidensintervall blir då

$$C(\boldsymbol{x}) = \{\theta : a \leq Q(\boldsymbol{X}, \theta) \leq b\}.$$

När $Q(\boldsymbol{X}, \theta)$ är monton på θ kan intervallen $[a, b]$ vara inverterad för att ge en konfidensintervall för θ.

Exempel 11.6. Vi fortsätter med Exempel 11.5. Eftersom $T \sim \chi^2(2n)$ kan vi ta $a = \chi^2_{1-\alpha/2}(2n)$ och $b = \chi^2_{\alpha/2}(2n)$ för att få

$$P\left(\chi^2_{1-\alpha/2}(2n) \leq \frac{2Y}{\lambda} \leq \chi^2_{\alpha/2}(2n)\right) \geq 1-\alpha.$$

Detta är detsamma som

$$P\left(\frac{2Y}{\chi^2_{1-\alpha/2}(2n)} \leq \lambda \leq \frac{2Y}{\chi^2_{\alpha/2}(2n)}\right) \geq 1-\alpha.$$

Därför har vi $1-\alpha$ konfidensintervallen

$$C(\boldsymbol{x}) = \left[\frac{2}{\chi^2_{1-\alpha/2}(2n)}\sum_{i=1}^n x_i, \frac{2}{\chi^2_{\alpha/2}(2n)}\sum_{i=1}^n x_i\right].$$

Om vi tar $n = 2$ och $1-\alpha = 0.9$ som i Exempel 11.3 får vi

$$C(\boldsymbol{x}) = \left[\frac{x_1+x_2}{4.75}, \frac{x_1+x_2}{0.355}\right]$$

Notisera att detta intervall är inte detsamma som intervallen från Exempel 11.3. I vissa fall ska båda metoder ger oss detsamma intervallskattning men inte alltid.

Ett fall när både metoder ger samman intervallen är för normalfördelningar.

Exempel 11.7. Låt $X_1, \ldots, X_n$ vara oberoende och $\mathrm{N}(\theta, \sigma^2)$-fördelade där σ^2 är känt. Det följer att

$$Z = \frac{\bar{X}-\theta}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

och därför kan vi ta Q som en pivotvariabel. Vi har då att

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

och därför

$$P\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \theta \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha.$$

Det följer att

$$C(\boldsymbol{x}) = \left[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right]$$

är en $1-\alpha$ konfidensintervall.

På samma sätt, om σ^2 är okänt har vi att

$$\frac{\bar{X}-\theta}{S/\sqrt{n}} \sim t(n-1)$$

är en pivotvariabel. Detta pivotvariabel ger oss $1-\alpha$ konfidensintervallen

$$C(\boldsymbol{x}) = \left[\bar{x} - t(n-1)_{\alpha/2}S/\sqrt{n}, \bar{x} + t(n-1)_{\alpha/2}S/\sqrt{n}\right].$$

Vi har även en annan pivotvariabel

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

som ger oss

$$P\left(\chi^2_{1-\alpha/2}(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha.$$

Detta är detsamma som

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha.$$

Därför en $1-\alpha$ konfidensintervall för σ^2 är

$$C(\boldsymbol{x}) = \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right].$$

11.1.3 Kortaste intervaller

Vi har nu två metoder för att bestämma en konfidensintervall och vi har sett att de kan ge olika resultat. Exempelvis, för stickprovet x_1, x_2 från en Exp(λ)-population beräknade vi de $1-\alpha$ konfidensintervallerna

$$\left[\frac{x_1+x_2}{5.480}, \frac{x_1+x_2}{0.441}\right] \qquad \text{och} \qquad \left[\frac{x_1+x_2}{4.75}, \frac{x_1+x_2}{0.355}\right]$$

för $\alpha = 0.1$. Frågan blir då, "Vilken intervall borde vi använda?" Vi kan jämföra deras längder som är

$$2.1(x_1+x_2) \qquad \text{och} \qquad 2.6(x_1+x_2).$$

Vi ser att den första intervallen är kortare (intervallen som beräknas med hjälp av likelihood-kvottestet). Vi kan då hävda att denna intervall är bättre eftersom den är mer exakt än den andra men den har detsamma täckningssannolikheten. (Å andra sidan kan vi säga att den andra intervallen är bättre eftersom det var både lättare att beräkna och kan beskrivas mer exakt med hjälp av χ^2-kvantiler... men detta beror såklart på tillämpningen.)

Notisera också att vi kunde få en tredje 0.90 konfidensintervall genom pivotvariablen $2Y/\lambda$. Detta ger oss intervallen

$$\left[\frac{2}{\chi^2_{0.90}(4)}(x_1+x_2), \infty\right) = \left[\frac{2}{3.90}(x_1+x_2), \infty\right).$$

Denna intervall har detsamma täckningssannolikheten som de andra två men oändlighet längd.

Exempel 11.8. Som ett annat exempel låter vi $x_1, \ldots, x_n$ vara ett stickprov från en N(μ, σ^2)-population med σ^2 känt. Vi har att

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

är en pivotvariabel. Så vi vill hitta a och b så att

$$P(a \le Z \le b) = 1-\alpha.$$

Innan tog vi $a = -z_{\alpha/2}$ och $b = z_{\alpha/2}$ men det finns ett oändlighet antal (a,b) som löser ekvationen. Nämligen, för godtycklig $\beta \in (0,1)$ har vi att

$$a = z_{1-\alpha-\beta} = -z_{\alpha+\beta} \qquad b = z_\beta$$

uppfyller $P(a \le Z \le b) = 1-\alpha$. En sådan lösning ger $1-\alpha$ konfidensintervallen

$$\left[\bar{x} - z_\beta \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha+\beta}\frac{\sigma}{\sqrt{n}}\right]$$

som har längden $(z_{\alpha+\beta}-z_\beta)\sigma/\sqrt{n}$. Givet ett visst stickprov $x_1, \ldots, x_n$ är $\sigma/\sqrt{n}$ fixt (eftersom σ är känt). Därför kan vi bara kontrollera längden av intervallen mellan $z_{\alpha+\beta}-z_\beta$. Om vi test flera val för β ser vi följande för $1-\alpha = 0.90$:

β	a	b	$b-a = z_{\alpha+\beta}-z_\beta$
0.01	-1.34	2.33	3.67
0.025	-1.44	1.96	3.40
0.05	-1.65	1.65	3.30
0.075	-1.96	1.44	3.40

Valet $\beta = 0.05 = \alpha/2$ verkar att vara optimalt över dessa värde. Följande satsen säger att detta stämmer.

Sats 11.2. *Låt $f_X(x)$ vara en unimodal täthetsfunktion (dvs. $f_X(x)$ har en enda lokal maximum). Om intervallen $[a,b]$ uppfllyer*

1. $\int_a^b f_X(x)dx = 1-\alpha$,
2. $f_X(a) = f_X(b) > 0$, *och*
3. $a \le x^* \le b$, *där x^* ger lokal maximum av $f_X(x)$*

så gäller det att $[a, b]$ är den kortaste intervall som uppfyller (a).

Vi har även följande som tillämpar till vårt exempel med normalfördelningen:

Följsats 11.1. *Om $f_X(x)$ uppfyller konditionerna i Sats 11.2 och det är också symmetrisk är den kortaste $1-\alpha$ konfidensintervall $[a, b]$ där*

$$\int_{-\infty}^{a} f_X(x)dx = \alpha/2 = \int_{b}^{\infty} f_X(x)dx.$$

Detta resultat förklara varför det bäste valet för β i Exempel 11.8 var $\beta = 0.05$.

11.2 Bayesiansk Intervallskattning

Från det bayesianska perspektivet är intervallskattningar även mer centrala eftersom de ger en mycket bättre sammanfattning av aposterioriförderlningen än en punktskattning, exempelvis, aposterioriväntevärdet. I bayesiansk inställningen ersätter vi täckningssannolikheter med aposteriorisannolikheter. Frekventistiska intervallskattningar är stokastiska (eftersom $L(\boldsymbol{X})$ och $U(\boldsymbol{X})$ är stokastiska variabler) och intervallen täckar den riktiga parametern θ med en viss sannolikhet (täckningssannolikheten). Å andra sidan är en bayesiansk intervallskattningen fixt givet data men parametern θ har inget riktigt värde och kan variera. Därför är parametern hittas i intervallen med en viss sannolikhet.

Vi definiera våra bayesiansk skattningar som mängder (på samma sett som våra frekventistiska skattningar) men dessa mängder är igen nästan alltid intervaller. En $1-\alpha$ *kredibilitetsmängd* för en viss parameter θ givet data $\boldsymbol{X} = \boldsymbol{x}$ är en mängd A så att

$$P(\theta \in A|\boldsymbol{X} = \boldsymbol{x}) = \int_A f_\Theta(\theta|\boldsymbol{X} = \boldsymbol{x})d\theta = 1 - \alpha.$$

Denna sannolikhet kallas för *kredibilitetssannolikheten* för A. När A är en intervall säger vi att kredibilitetsmängden A är en *kredibilitetsintervall*. För att bestämma en $1-\alpha$ kredibilitetsintervall behöver vi hitta c och d för vilka vi har

$$\int_{-\infty}^{c} f_\Theta(\theta|\boldsymbol{x})d\theta + \int_{d}^{\infty} f_\Theta(\theta|\boldsymbol{x})d\theta = \alpha.$$

Ett naturligt val för en sådan intervall är $[c, d]$ där

$$\int_{-\infty}^{c} f_\Theta(\theta|\boldsymbol{x})d\theta = \frac{\alpha}{2} \qquad \text{och} \qquad \int_{d}^{\infty} f_\Theta(\theta|\boldsymbol{x})d\theta = \frac{\alpha}{2}.$$

En sådan intervall kallas för en $1-\alpha$ *symmetriskt kredibilitetsintervall*.

Exempel 11.9. Låt $X_1, \ldots, X_n$ vara oberoende och $\text{Po}(\lambda)$-fördelade där $\Lambda \sim \text{Gamma}(a, b)$ och a är ett positivt heltal. Det följer att

$$\Lambda|\boldsymbol{x} \sim \text{Gamma}(a + \sum_{i=1}^{n} x_i, b + n).$$

För att bestämma en $1-\alpha$ kredibilitetsintervall behöver vi hitta c och d för vilka vi har

$$\int_{-\infty}^{c} f_\Lambda(\lambda|\boldsymbol{x})d\lambda + \int_{d}^{\infty} f_\Lambda(\lambda|\boldsymbol{x})d\lambda = \alpha.$$

Vi ska bestämma en symmetriskt kredibilitetsintervall. Därför skulle vi vilja hitta c och d så att

$$\int_{-\infty}^{c} f_\Lambda(\lambda|\boldsymbol{x})d\lambda = \frac{\alpha}{2} \qquad \text{och} \qquad \int_{d}^{\infty} f_\Lambda(\lambda|\boldsymbol{x})d\lambda = \frac{\alpha}{2}.$$

Vi använder igen faktumet att

$$\text{Gamma}(n, 1/2) = \chi^2(2n)$$

för att få

$$2(b+n)\Lambda|\boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(a + \sum_{i=1}^{n} x_i, 1/2) = \chi^2(2a + 2\sum_{i=1}^{n} x_i).$$

(Här använder vi att $a + \sum_{i=1}^n x_i$ är ett positivt heltal.) Då kan vi använda χ^2-kvantilerna för att se

$$P\left(\chi^2_{1-\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right) \leq 2(b+n)\Lambda \leq \chi^2_{\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right)\right) = 1 - \alpha.$$

Därför är

$$c = \frac{1}{2(b+n)}\chi^2_{1-\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right) \qquad \text{och} \qquad d = \frac{1}{2(b+n)}\chi^2_{\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right).$$

$1-\alpha$ symmetriskt kredibilitetsintervallen är då

$$\left[\frac{1}{2(b+n)}\chi^2_{1-\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right), \frac{1}{2(b+n)}\chi^2_{\alpha/2}\left(2a + 2\sum_{i=1}^n x_i\right)\right].$$

Som ett konkret exempel kan vi låta $\alpha = 0.1$, $a = b = 1$, $n = 10$ och $\sum_{i=1}^n x_i = 6$. Då blir 0.90 symmetriskt kredibilitetsintervallen

$$[0.299, 1.077].$$

Som ett alternativ till $1-\alpha$ symmetriskt kredibilitets intervaller kan vi betrakta $1-\alpha$ kredibilitetsintervaller med minsta längd. I sådana fall kan vi tillämpa Sats 11.2 till aposteriorifördelning för att få följande resultat:

Följsats 11.2. *Om $f_\Theta(\theta|\boldsymbol{X} = \boldsymbol{x})$ är unimodal gäller det för $\alpha > 0$ att den kortaste $1-\alpha$ kredibilitetsintervallen ges av $[c, d]$ där $c \leq \theta^* \leq d$ (där θ^* är aposterioritypvärdet) och*

1. *$f_\Theta(c|\boldsymbol{X} = \boldsymbol{x}) = f_\Theta(d|\boldsymbol{X} = \boldsymbol{x}) > 0$, och*
2. *$\int_c^d f_\Theta(\theta|\boldsymbol{X} = \boldsymbol{x})d\theta = 1 - \alpha$.*

En intervall som uppfyller villkoren av Följsats 11.2 kallas för en *HPD-kredibilitetsintervall* och det är intervallen som har den högsta aposteriorisannolikheten att innehålla parametern. Eftersom täthetsfunktionen är unimodal kan man skriva en sådan krediblitetsintervall som

$$\{\theta \in \Omega : f_\Theta(\theta|\boldsymbol{X} = \boldsymbol{x}) \geq k\}$$

för någon $k > 0$. Notisera att en $1-\alpha$ HPD-kredibilitetsintervall är detsamma som den $1-\alpha$ symmetriskt kredibilitetsintervall när $f_\Theta(\theta|\boldsymbol{X} = \boldsymbol{x})$ är både unimodal och symmetrisk.

Exempel 11.10. Vi återgår till Exempel 11.9 där vi hade ett stickprov $x_1, \ldots, x_n$ från en Po(λ)-population där vi hade apriori fördelningen $\Lambda \sim \text{Gamma}(1, 1)$. Vi fick aposteriorifördelninen $\Lambda|\boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(1 + \sum_{i=1}^n x_i, 1 + n)$. För att hitta c och d så att $f_\Lambda(c|\boldsymbol{X} = \boldsymbol{x}) = f_\Lambda(d|\boldsymbol{X} = \boldsymbol{x})$ och $\int_c^d f_\Lambda(\lambda|\boldsymbol{X} = \boldsymbol{x})d\theta = 1 - \alpha$ behöver vi att lösa systemet av ekvationer

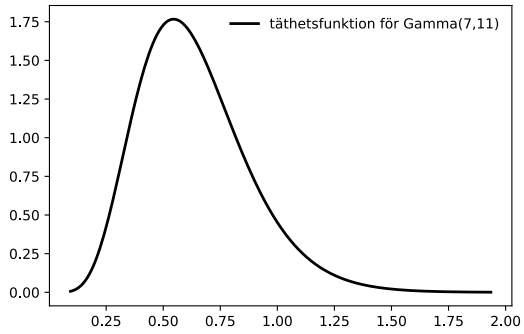
$$c^{\sum_{i=1}^n x_i} e^{-(n+1)c} = d^{\sum_{i=1}^n x_i} e^{-(n+1)d},$$

$$\int_c^d \frac{\Gamma(n+1)}{\Gamma\left(1 + \sum_{i=1}^n x_i\right)} \lambda^{\sum_{i=1}^n x_i} e^{-(n+1)\lambda} d\lambda = 1 - \alpha.$$

Vi kan använda partiellintegration som i Exempel 11.3 för att lösa systemet. Om vi göra det för $n = 10$ och $\sum_{i=1}^n x_i = 6$ får vi

$$[c, d] \approx [0.253, 1.055].$$

Denna interval har längd 0.75 medan den $1-\alpha$ symmetriska kredibilitetsintervallen som beräknas i Exempel 11.9 har längd 0.78. Därför är intervallen $[0.253, 1.055]$ kortare trots att den har detsamma kredibilitetssannolikhet. Det stämmer eftersom vi kan verifiera att Gamma(7, 11) har inte en symmetrisk täthetsfunktion:



11.3 Beslutsteori för intervallskattning

Precis som vi gjorde för punktskattningar och hypotestest kan vi betraka intervaller som handlingar i ett handlingsrum och intervallskattningar som ett beslut $W(\boldsymbol{x})$ som skickar ett stickprov $\boldsymbol{x}$ till en intervall. Då kan vi definiera förlustfunktioner på dessa beslut och studerar deras riskfunktioner och bayesbeslut.

Låt oss betrakta en-sidig intervaller i formen $(-\infty, a]$. Vi kan definiera förlustfunktionen

$$L(\theta, a) = \begin{cases} c(a-\theta) & a \geq 0 \quad \text{(intervallen täcker } \theta\text{)} \\ (1-c)(\theta - a) & a < \theta \quad \text{(intervallen täcker inte } \theta\text{)} \end{cases}$$

för någon c som uppfllyer $c < 1-c$. Vi straffar därför intervaller som täcker θ med c gånger så mycket de "övertäcker" θ och vi straffar intervaller som täcker inte θ med $1-c$ gånger så mycket de "undertäcker" θ. Notisera att vi straffar intervaller som täcker inte θ mer eftersom $c < 1-c$.

Riskfunktionen och aposterioririsken kan beräknas som vanligt. För detta förlustfunktion kan det visas att det resultande bayesbeslut (dvs., intervallskattningen som minimeras aposterioririsken) ges av lösningen till

$$\int_{-\infty}^{W(\boldsymbol{x})} f_\Theta(\theta | \boldsymbol{X} = \boldsymbol{x}) d\theta = 1 - c$$

för alla $\boldsymbol{x}$. Det vill säga att bayesbeslutet är aposteriorifördelningens $(1-c)$-kvantil.

Om vi är intresserad av två-sidig intervaller $[a_1, a_2]$ kan vi betrakta förlustfunktionen

$$L(\theta, (a_1, a_2)) = a_2 - a_1 + \begin{cases} c_1(a_1 - \theta) & \theta < a_1, \\ 0 & a_1 \leq \theta \leq a_2, \\ c_2(\theta - a_2) & \theta > a_2. \end{cases}$$

Denna förlustfunktion straffar för längre intervaller (med $a_2 - a_1$) och då straffar den intervaller som täcker inte θ. Vi kan straffar mer eller mindre om intervallen är ovan eller under θ med våra val till c_1 och c_2. I både fall är straffet skalad enligt distans mellan intervallen och parametern θ.

För denna förlustfunktion kan man visa att bayesbeslutet ges av $W(\boldsymbol{x}) = [a_1, a_2]$ där a_1 är aposteriorfördelningens $1/c_1$-kvantil och a_2 är $1 - 1/c_2$-kvantilen när $c_1, c_2 > 1$. Det följer att om vi tar $c_1 = c_2 = \alpha/2$ är bayesbeslutet den $1-\alpha$ symmetriskt kredibilitetsintervallen.

Vi kan också betrakta en förlustfunktion som använder ett konstant straff om vi täcker inte parametern:

$$L(\theta, (a_1, a_2)) = a_2 - a_1 + c(1 - \mathbf{1}_{[a_1, a_2]}(\theta).$$

Bayesbeslutet blir då

$$\{\theta : f_\Theta(\theta | \boldsymbol{X} = \boldsymbol{x}) \geq 1/c\}$$

när aposteriorifördelningen är unimodal och kontinuerlig. Det vill säger att bayesbeslutet är den $1-\alpha$ HPD-kredibilitetsintervallen med undergränsen $1/c$. Ju högre kostnaden för att missa parametern desto längre blir intervallen (och större blir kredibilitetssannolikheten).