Allmän metodik att härleda konfidensintervall

Vi har data $x_1, x_2, \cdots, x_n$ som vi ser som utfall av stokastiska variabler $X_1, X_2, \cdots, X_n$ vilkas fördelning påverkas av en parameter $\Theta$.

Ett konfidensintervall med konfidensgraden $1 - \alpha$ (typiskt 95% dvs att $\alpha = 0.05$) för parametern $\Theta$ är ju ett numeriskt intervall

$$\big(a(x_1, x_2, \cdots, x_n), b(x_1, x_2, \cdots, x_n)\big)$$

där funktionerna $a$ respektive $b$ skall välja så att

$$P\left(a(X_1, X_2, \cdots, X_n) \leq \Theta \leq b(X_1, X_2, \cdots, X_n)\right) = 1 - \alpha.$$

Detta betyder att det numeriska konfidensintervallet är att betrakta som ett utfall av det stokastiska intervallet

$$\big(a(X_1, X_2, \cdots, X_n), b(X_1, X_2, \cdots, X_n)\big)$$

och konfidensgraden anger sannolikheten att det stokastiska intervallet innehåller $\Theta$. Konfidensgraden anger alltså sannolikheten att metoden att räkna ut konfidensintervallet fungerar, men är inte en utsaga om det enskilda numeriska konfidensintervallet.

Det finns en standardiserad metodik i 5 steg som (nästan) maskinmässigt tar fram funktionerna $a$ och $b$, dvs som producerar konfidensintervall. Formelsamlingen utgör en rik källa av tips vad gäller det praktiska genomförande av dessa 5 steg. Dessutom finns hjälpresultat som genomför stegen 3-5 i ett svep. Nedan beskrivs dessa 5 steg först generellt och sen hur de praktiskt genomförs i 3 olika situationer.

1) Skatta $\Theta$ med någon listigt vald punktskattning $\Theta^*(x_1, x_2, \cdots, x_n)$ som t ex tas fram med hjälp Maximum Likelihood-metoden eller Minsta Kvadratmetoden. I vissa situationer finns andra okända parametrar än $\Theta$ och dessa skall också skattas.

2) Ta fram fördelningen för motsvarande stickprovsvariabel $\Theta^*(X_1, X_2, \cdots, X_n)$. Här kan formelsamlingen vara till stor hjälp.

3) Gör en listig transformation av stickprovsvariabeln så att resultatet ej beror av $\Theta$ eller andra parametrar som finns i problemet. Denna s k pivot-variabel (pivot betyder svängtapp - ett bra svenskt namn vore ”grundbult”) innehåller alltid den okända parametern $\Theta$. Formelsamlingen vimlar med tips om lämpliga transformationer i specifika situationer. I de konkreta situationer som beskrivs nedan ges exempel på dessa pivot-variabler. Vi skaffar oss alltså en ny variabel ur $\Theta^*$ som vi kallar t ex $T$ som har en fix (känd) fördelning oavsett värdet på $\Theta$.

4) Vi väljer tal $c$ och $d$ så att $P(c \leq T \leq d) = 1 - \alpha$. Detta innebär att vi skall hitta percentiler till fördelningen för $T$. Tabell 2-4 i formelsamlingen anger sådana s.k. percentiler för tre fördelningar som ofta dyker upp som fördelning för pivot-variabeln $T$ nämligen $N(0, 1)$-fördelningen, $t$-fördelningen och $\chi^2$-fördelningen. Om vi tar $c$ och $d$ så att $P(T < c) = P(T > d) = \alpha/2$ får vi tvåsidiga symmetriska konfidensintervall, som alltså ger både en undre och en övre gräns för parametern $\Theta$. Vill vi bara ha en undre eller bara en övre gräns för $\Theta$ (s k enkelsidigt konfidensintervall) skall vi ta $P(T < c) = \alpha$ och $d = +\infty$ eller $P(T > d) = \alpha$ och $c = -\infty$.

5) Vi vränger nu ovanstående uttryck för $T$ ut och in så att vi får $\Theta$ ensamt i mitten och att $\Theta$ ej finns någon annanstans. Vi har därmed får $\Theta$ instängt mellan två stokastiska gränser och konfidensintervallet är utfallet av dessa gränser, dvs typografiskt byter man $X$ mot $x$!

Exempel 1: Som exempel kan vi låta våra mätdata $x_1, x_2, \cdots, x_n$ vara utfall av $X_1, X_2, \cdots, X_n$ som är oberoende $N(\Theta, \sigma_0)$ där $\sigma_0$ är känd och vi vill göra ett konfidensintervall för $\Theta$ med konfidensgrad $1 - \alpha$. Stegen ovan motsvarar följande:

1) $\Theta^* = \bar{x} = (x_1 + x_2 + \cdots + x_n)/n$

2) $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ är $N\left(\Theta, \sigma_0/\sqrt{n}\right)$.

3) $T = \dfrac{\bar{X} - \Theta}{\sigma_0/\sqrt{n}}$ är $N(0,1)$

4) Vi har då p g a symmetrin i $N(0,1)$-fördelningen som gör $\lambda_{1-\alpha/2} = -\lambda_{\alpha/2}$ att

$$P\left(-\lambda_{\alpha/2} \leq T \leq \lambda_{\alpha/2}\right) = P\left(-\lambda_{\alpha/2} \leq \frac{\bar{X} - \Theta}{\sigma_0/\sqrt{n}} \leq \lambda_{\alpha/2}\right) = 1 - \alpha$$

5) Vi får

$$P\left(\bar{X} - \lambda_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \Theta \leq \bar{X} + \lambda_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

som ger konfidensintervallet $\bar{x} \pm \lambda_{\alpha/2}\dfrac{\sigma_0}{\sqrt{n}}$. De enkelsidiga intervallen blir då $\left(-\infty, \bar{x} + \lambda_\alpha\dfrac{\sigma_0}{\sqrt{n}}\right)$ respektive $\left(\bar{x} - \lambda_\alpha\dfrac{\sigma_0}{\sqrt{n}}, \infty\right)$.

Exempel 2: Om vi stället skulle ha $X_1, X_2, \cdots X_n$ som är oberoende $N(\Theta, \sigma)$ där både $\Theta$ och $\sigma$ är okända blir metodiken följande:

1) Vi skattar $\Theta$ med $\bar{x}$ och $\sigma$ med $s$ där $s^2 = \dfrac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$

2) Vi vet att $\bar{X}$ är $N(\Theta, \sigma/\sqrt{n})$ och formelsamlingen ger att $\dfrac{(n-1)s^2}{\sigma^2}$ är $\chi^2(n-1)$. Dessutom kan man visa (och läsa i formelsamlingen) att (förvånande nog) $\bar{X}$ och $s^2$ är oberoende. Notera att $s$ här är stickprovsvariabeln dvs $s^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ och ej det numeriska värdet $s$ där $s^2 = \dfrac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ som alltså räknas ut med våra observerade mätdata. När vi nedan tittar på utfallet av det stokastiska intervallet skall vi dock använda oss av det numeriska värdet!

3) Här ger formelsamlingen att $T = \dfrac{\bar{X} - \Theta}{s/\sqrt{n}}$ är $t(n-1)$-fördelad. $t(n-1)$-fördelningen införs och studeras just därför att den är fördelning för denna pivot-variabel.

4) Tabell 3 i formelsamlingen ger då p g a symmetrin i $t(n-1)$-fördelningen som gör att $t_{1-\alpha/2}(n-1) = -t_{\alpha/2}(n-1)$ att

$$P\left(-t_{\alpha/2}(n-1) \leq \frac{\bar{X} - \Theta}{s/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

5) Vi får

$$P\left(\bar{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \Theta \leq \bar{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

som ger intervallet

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

Exempel 3: Skulle vi i situationen i Exempel 2 vilja göra konfidensintervall för $\sigma$ får vi samma i 1) och 2) som ovan.

3) Pivotvariabel angavs redan i steg 2 dvs $\dfrac{(n-1)s^2}{\sigma^2}$ utgör pivot-variabel för $s$.

4) Tabell 4 ger nu att

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

5) Vi får här

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

som ger intervallet

$$\left(s\sqrt{\frac{n-1}{\chi^2_{\alpha/2}(n-1)}}, s\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}}\right)$$

för $\sigma$.

Detta schema kan nu genomföras för t ex två oberoende stickprov eller enkel linjär regression med små modifikationer på lämpliga ställen. För de s k approximativa metoderna väljer man som pivot-variabel en som är approximativt $N(0,1)$ oftast genom hänvisning till Centrala gränsvärdessatsen. Eftersom pivot-variabeln då bara är approximativt $N(0,1)$ blir konfidensgraden bara approximativt den avsedda $1-\alpha$. Dessa konfidensintervall kallas därför approximativa eftersom säkerheten i metoden bara är approximativt $1-\alpha$.

Man kan vidare notera att Formelsamlingens avsnitt om $\lambda$-metoden, $t$-metoden, approximativa metoden och $\chi^2$-metoden genomför steg 3, 4 och 5 i ett enda svep.