Högskolan i Halmstad

Intervallskattning

Antag att vi har ett stickprov $x_1, \ldots, x_n$ på X som vi vet är $N(\mu, \sigma)$ men vi vet ej värdet av $\mu = E(X)$. Då kan vi beräkna $\bar{x}$, vvr skattning av μ. För att få reda på hur tillförlitlig denna skattning är kan vi beräkna $s/\sqrt{n}$ som är den vvr skattningen av standardavvikelsen för $\bar{x}$.

Denna information kan även utnyttjas för *konfidensintervall* och *hypotestest*.

Eftersom vi vet värdet av μ beräknar vi $\bar{x}$ men vi kanske också vill veta mellan vilka gränser μ ligger med någon viss sannolikhet (säg $1-\alpha$). Dessa gränser utgör ett konfidensintervall (konf.int.). Eftersom gränserna, C_1 och C_2, beräknas m.h.a. $\bar{X}$ och S, är C_1 och C_2 stokastiska variabler och intervallet ett stokastiskt intervall.

Definition 1 *En* **intervallskattning** *är ett intervall med stokastiska gränser.* **Konfidensgraden** *är sannolikheten att intervallet innehåller parametern i fråga. Ett* **konfidensintervall** *är en observation av en intervallskattning.*

På samma sätt som $p = P(a \leq X \leq b)$ betyder att variabeln X "hamnar inom" eller "träffar" intervallet (a, b) med sannolikhet p, kan man säga att om (C_1, C_2) är en intervallskattning för μ med konfidensgrad $1-\alpha$ så "innehåller" eller "träffar" det μ med sannolikhet $1-\alpha$.

Konfidensintervall för medianen m vid kontinuerlig fördelning

Exempel Låt X vara tiden det tar för en elev att skriva en provräkning. Antag att vi har det stokastiska stickprovet $X_1, X_2, \ldots, X_{10}$, skrivtider för olika elever. Vilken konfidensgrad har intervallet $(\min_i X_i, \max_i X_i)$ för medianen m? Hur stor konfidensgrad har intervallet med den näst minsta och den tredje största observationen som gränser?

Lösning: Kom ihåg definitionen av medianen:
m är det tal sådant att $P(X \leq m) = P(X > m) = 0.5$.

Antag att vi sorterar stickprovet: $\underbrace{X_{(1)}}_{\min_i X_i}, X_{(2)}, \ldots, X_{(9)}, \underbrace{X_{(10)}}_{\max_i X_i}$. Då är

$$\begin{aligned}
P(X_{(1)} \leq m \leq X_{(10)}) &= P(X_{(1)} \leq m, X_{(10)} \geq m) \\
&= 1 - P\Big(\Big\{\{X_{(1)} \leq m\} \cap \{X_{(10)} \geq m\}\Big\}^C\Big) \\
&= 1 - P\Big(\underbrace{\{X_{(1)} > m\} \cup \{X_{(10)} < m\}}_{\text{disjunkta händelser!}}\Big) \\
&= 1 - \Big(\underbrace{P(X_{(1)} > m)}_{I} + \underbrace{P(X_{(10)} < m)}_{II}\Big)
\end{aligned}$$

$I = P(\min_i X_i > m) = P(X_1 > m, \ldots, X_{10} > m) = \underbrace{P(X_1 > m)}_{0.5} \cdots \underbrace{P(X_{10} > m)}_{0.5} = 0.5^{10}$.
På samma sätt är även $II = 0.5^{10}$ varmed $P(X_{(1)} \leq m \leq X_{(10)}) = 1 - 2 \cdot 0.5^{10} \approx 0.998$, dvs $(X_{(1)}, X_{(10)})$ är en intervallskattning för m med konfidensgrad 99.8%.

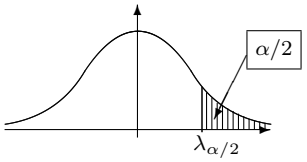
Med den näst minsta och tredje största observationen som gränser menas intervallet $(X_{(2)}, X_{(8)})$. Detta har konfidensgrad
$P(X_{(2)} \leq m \leq X_{(8)}) = P(m \leq X_{(8)}) - P(m \leq X_{(2)}) = \underbrace{P(X_{(2)} \leq m)}_{I} - \underbrace{P(X_{(8)} \leq m)}_{II}$
där
$I = 1 - P(X_{(2)} > m) = 1 - P(\text{högst 1 obs.} \leq m) =$
$= 1 - (P(0 \text{ obs.} \leq m) + P(1 \text{ obs.} \leq m)) = 1 - 0.5^{10} - \binom{10}{1}0.5^{10} = 1 - 11 \cdot 0.5^{10}$.
och
$II = P(\text{högst 2 obs.} > m) = P(0 \text{ obs.} > m) + P(1 \text{ obs.} > m) + P(2 \text{ obs.} > m) =$
$= 0.5^{10} + \binom{10}{1}0.5^9(1-0.5)^1 + \binom{10}{2}0.5^8(1-0.5)^2 = 56 \cdot 0.5^{10}$.
Därmed är konfidensgraden
$I - II = 1 - 11 \cdot 0.5^{10} - 56 \cdot 0.5^{10} = 1 - 67 \cdot 0.5^{10} = 0.9346$
dvs $(X_{(2)}, X_{(8)})$ är en intervallskattning med konfidensgrad 93.5%. □

Konfidensintervall för μ då σ är känt

Antag att σ känt. Då är (C_1, C_2) ett $1-\alpha$ konf.int. (egentligen intervallskattning) för μ om $1-\alpha = P(C_1 \leq \mu \leq C_2)$. Antag nu att C_1 och C_2 ska ligga "symmetriskt" kring μ, dvs att att $P(\mu < C_1) = P(\mu > C_2) = \frac{\alpha}{2}$. Eftersom $\bar{X}$ är det bästa vi kan utgå ifrån då vi inte känner μ, låt $C_1 = \bar{X} - a$ och $C_2 = \bar{X} + a$. Då är

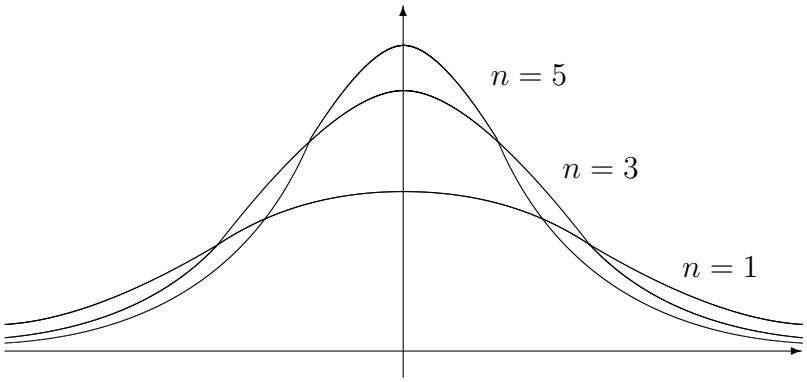
$$\begin{aligned}
\tfrac{\alpha}{2} &= P(\mu < C_1) \\
&= 1 - P(C_1 \leq \mu) \\
&= 1 - P(\bar{X} - a \leq \mu) \\
&= 1 - P(\bar{X} \leq \mu + a) \\
&= 1 - P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu + a - \mu}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{a}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

dvs $\Phi(\frac{a}{\sigma/\sqrt{n}})$ varmed $\frac{a}{\sigma/\sqrt{n}} = \lambda_{\alpha/2}$, där $\lambda_{\alpha/2}$ är normalfördelningens $\frac{\alpha}{2}$-**percentil**, dvs $\lambda_{\alpha/2}$ är det x-värde bortom vilket $\frac{\alpha}{2}$ av arean under täthetsfunktionen ligger, dvs x-värdet sådant att $1 - \Phi(x) = \frac{\alpha}{2}$



Därmed är $a = \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}$ och $(C_1, C_2) = (\bar{X} - \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}\ ,\ \bar{X} + \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2})$ en intervallskattning av μ med konfidensgrad $1-\alpha$.

t-fördelningen



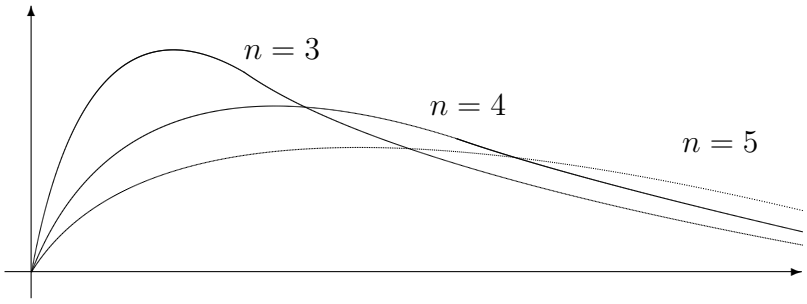
Liknar väldigt mycket standard normalfördelningen. Enda parametern till t-fördelningen är antalet frihetsgrader n. Värden för dess fördelningsfunktion fås från tabell med rätt frihetsgradstal.

$t(n-1)$ anger t-fördelningen med $n-1$ frihetsgrader.

$t_{\alpha/2}(n-1)$ anger $\frac{\alpha}{2}$-percentilen av t-fördelningen med frihetsgradstal $n-1$.

T.ex. får man med $\alpha = 0.05$ och $n = 10$ percentilen $t_{\alpha/2}(n-1) = t_{0.025}(9) = 2.26$.

χ^2-fördelningen (uttalas "tji-två-fördelningen")



Även χ^2-fördelningen har endast en parameter, frihetsgradstalet n och percentilvärden för den fås genom att slå i tabell.

$\chi^2(n-1)$ anger χ^2-fördelningen med $n-1$ frihetsgrader.

$\chi^2_\alpha(n-1)$ anger α-percentilen av χ^2-fördelningen med frihetsgradstal $n-1$.

T.ex. fås med $\alpha = 0.05$ och $n = 10$ percentilvärdet $\chi^2_\alpha(n-1) = \chi^2_{0.05}(9) = 3.33$.

Konfidensintervall för μ då σ är okänt

Antagandet att σ är känt är naturligtvis i princip aldrig realistiskt. Denna situation var med bara som ett trappsteg mot den mer tillämpbara situationen där såväl μ som σ antas vara okända och man vill göra en intervallskattning av väntevärdet μ.

Resonemanget blir i princip detsamma men den stokastiska variabeln $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ blir ej normalfördelad utan t-fördelad (eftersom σ i nämnaren bytts mot skattningen S). Därmed blir proceduren densamma som förut fast σ byts mot S och normalpercentilen $\lambda_{\alpha/2}$ byts mot t-percentilen $t_{\alpha/2}(n-1)$ (som alltså även beror på stickprovsstorleken). Summan av kardemumman är att en intervallskattning av μ med konfidensgrad $1-\alpha$ nu är

$$\left(\bar{X} - \frac{S}{\sqrt{n}}\, t_{\alpha/2}(n-1) \quad , \quad \bar{X} + \frac{S}{\sqrt{n}}\, t_{\alpha/2}(n-1)\right)$$

Konfidensintervall för σ^2

I detta fall antas både μ och (givetvis) σ vara okända. En intervallskattning av σ^2 med konfidensgrad $1-\alpha$ är då

$$\left(0 \quad , \quad \frac{(n-1)S^2}{\chi^2_\alpha(n-1)}\right)$$

Exempel Vid ett sjukhus förlöses havande kvinnor. Vid 5 slumpvis valda förlossningar var tiderna

15.3 16.5 13.8 14.7 13.9

timmar. Vad blir ett 95% konfidensintervall

(a) för den förväntade förlossningstiden om σ är känt $= 1$?

(b) för den förväntade förlossningstiden om σ är okänt?

(c) för variansen av förlossningstiden?

Lösning:

(a) $\bar{x} = \frac{1}{5}(15.3 + 16.5 + 13.8 + 14.7 + 13.9) = 14.84$
$\frac{\sigma}{\sqrt{n}} \cdot \lambda_{\alpha/2} = \frac{1}{\sqrt{5}} \cdot 1.96 = 0.876.$
Därmed blir intervallet $(14.84 - 0.876\ ,\ 14.84 + 0.876) \approx (13.96\ ,\ 15.72)$.
Observera att under gränsen avrundas nedåt och övre gränsen avrundas uppåt!

(b) $s = \sqrt{\frac{1}{5-1}(15.3^2 + 16.5^2 + 13.8^2 + 14.7^2 + 13.9^2 - 5 \cdot 14.84^2)} = \sqrt{1.238} = 1.113$
$\frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) = \frac{1.113}{\sqrt{5}} \cdot 2.776 = 1.382.$
Därmed blir intervallet $(14.84 - 1.113\ ,\ 14.84 + 1.113) \approx (13.45\ ,\ 16.23)$.
Observera återigen avrundningsprincipen!!

(c) Intervallet blir $(0\ ,\ \frac{(n-1)s^2}{\chi^2_\alpha(n-1)}) = (0\ ,\ \frac{4 \cdot 1.238}{9.49}) = (0\ ,\ 0.522)$. □

Exempel Koldioxidhalten är approximativt normalfördelad med standardavvikelse 0.12. Man vill bilda ett 98% konfidensintervall för väntevärdet. Hur många observationer måste man ha för att intervallet ska bli 0.1 brett?

Lösning: $X \in N(\mu, \sigma)$ och σ känt varmed konfidensintervallet är
$(\bar{x} - \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}\ ,\ \bar{x} + \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2})$ varmed bredden blir
$(\bar{x} + \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}) - (\bar{x} - \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}) = 2\frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}$ vilket ska vara $= 0.1$.
Om vi ur detta löser ut n fås $n = (2\sigma\lambda_{\alpha/2}\frac{1}{0.1})^2 = (2 \cdot 0.12 \cdot 2.3263 \cdot 10)^2 = 31.17\ldots$
dvs man måste ha *minst* 32 observationer. (Observera avrundningen uppåt för att konfidensintervallet garanterat ska ha konfidensgrad 98%.) □