Föreläsning 11

Thomas Önskog

29/11 2017

Konfidensintervall för jämförelser mellan två stickprov

På förra föreläsningen bestämde vi konfidensintervall för väntevärdet och standardavvikelsen av ett normalfördelat stickprov. Vi ska nu göra jämförelser mellan två normalfördelade stickprov. Exempel på frågeställningar som kan besvaras på detta sätt är

1. Ger medicin A större effekt på en viss sjukdom än placebo?
2. Går man ner i vikt av dieten B?

För att besvara frågeställning 1 kan en grupp patienter slumpmässigt delas in i två undergrupper, där patienterna i den ena undergruppen får medicin A och patienterna i den andra undergruppen får placebo. Efter behandlingstidens slut mäts patienternas tillstånd med någon lämplig metod. Låt $x_1, ..., x_{n_1}$ beteckna stickprovet av mätvärden från patienterna som har fått placebo och låt $y_1, ..., y_{n_2}$ beteckna stickprovet av mätvärden från patienterna som fått medicin A. Vi får följande tabell (där ett värde över 20 i detta fall motsvarar att patienten är sjuk).

Placebo	$n_1 = 43$	$\overline{x} = 21.57$	$s_1 = 3.87$
Medicin A	$n_2 = 33$	$\overline{y} = 20.38$	$s_2 = 3.91$

Vi antar att $x_1, ..., x_{n_1}$ och $y_1, ..., y_{n_2}$ är stickprov från $N(\mu_1, \sigma_1)$- respektive $N(\mu_2, \sigma_2)$-fördelningar och att alla stickprovsvariabler är oberoende. Vi vill bestämma ett **konfidensintervall för skillnaden i väntevärde** $\mu_1 - \mu_2$. Antag först att σ_1 och σ_2 är kända. Från tidigare sats om linjärkombinationer av normalfördelade s.v. så vet vi att $\overline{X} \in \mathrm{N}(\mu_1, \sigma_1/\sqrt{n_1})$ och att $\overline{Y} \in \mathrm{N}(\mu_2, \sigma_2/\sqrt{n_2})$, så

$$\underline{\overline{X} - \overline{Y} \in \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right),}$$

vilket innebär att

$$\underbrace{\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}}_{\text{Pivotvariabel}} \in \mathrm{N}(0,1) \qquad \Rightarrow \qquad \mathbb{P}\Big(-\lambda_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < \lambda_{\alpha/2} \Big) = 1 - \alpha,$$

och ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ är därmed

$$\underline{I_{\mu_1-\mu_2} = (\overline{x} - \overline{y} - \lambda_{\alpha/2}D, \overline{x} - \overline{y} + \lambda_{\alpha/2}D)}, \text{ där } \underline{D = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Antag härnäst att σ_1 och σ_2 är okända men identiska och båda lika med σ. Eftersom D, som i detta fall kan förenklas till $D = \sigma\sqrt{1/n_1 + 1/n_2}$, är okänt, så behöver vi skatta D med ett lämpligt medelfel. Om vi som medelfel väljer

$$\underline{s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}, \qquad \text{så gäller} \qquad \underbrace{\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S\sqrt{1/n_1 + 1/n_2}}}_{\text{Pivotvariabel}} \in t(n_1+n_2-2),$$

där S är stickprovsvariabeln till punktskattningen s. Det tvåsidiga konfidensintervallet för $\mu_1 - \mu_2$ blir i detta fall

$$\underline{I_{\mu_1-\mu_2} = (\overline{x} - \overline{y} - t_{\alpha/2}(n_1+n_2-2)d, \overline{x} - \overline{y} + t_{\alpha/2}(n_1+n_2-2)d)}, \text{ där } \underline{d = s\sqrt{1/n_1 + 1/n_2}}.$$

Exempel. *I den medicinska undersökningen kan vi anta att $\sigma_1 = \sigma_2 = \sigma$ är okänt. Vi får då $s^2 = 15.11$ och gränserna för ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ med konfidensgrad 95% blir*

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(n_1+n_2-2)s\sqrt{1/n_1 + 1/n_2} = 1.19 \pm 1.793 \qquad \Rightarrow \qquad I_{\mu_1-\mu_2} = (-0.60, 2.94).$$

*Med konfidensgrad 95% är skillnaden mellan medeleffekterna för medicin A och placebo mellan -0.6 och 2.9. Notera att vi **inte** kan påstå att skillnaden mellan medeleffekterna med 95% sannolikhet är mellan -0.6 och 2.9, eftersom det sanna värdet på $\mu_1 - \mu_2$ är fixt och inte en fråga om sannolikhet.*

Situationen i frågeställning 2 ovan är snarlik men kräver en annan analys eftersom det nu är samma personer som undersöks före och efter dieten. Stickproven $x_1, ..., x_n$ och $y_1, ..., y_n$ är därför beroende. Variationen inom varje stickprov är generellt större än skillnaden mellan x_i och y_i. Eftersom x_i och y_i härrör från samma person talar vi om **stickprov i par**. Vi antar att x_i kommer från en $\mathrm{N}(\mu_i, \sigma_1)$-fördelning och att y_i kommer från en $\mathrm{N}(\mu_i + \Delta, \sigma_2)$-fördelning, där σ_1 och σ_2 kan vara olika. Parametern Δ anger den systematiska skillnaden mellan stickproven, i detta fall dietens eventuella effekt.

För att bestämma den systematiska skillnaden mellan stickproven $x_1, ..., x_n$ och $y_1, ..., y_n$, så undersöker vi stickprovet av differenser $z_1, ..., z_n$, där $z_i = y_i - x_i$. Eftersom y_i och x_i kommer från $\mathrm{N}(\mu_i + \Delta, \sigma_2)$- respektive $\mathrm{N}(\mu_i, \sigma_1)$-fördelningar, så kommer z_i från en $\mathrm{N}(\Delta, \sqrt{\sigma_1^2 + \sigma_2^2})$-fördelning.

Vi har därför reducerat problemet till att bestämma ett konfidensintervall för väntevärdet Δ av **ett** normalfördelat stickprov. Om $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ är känt, så får vi

$$\underline{I_\Delta = \left(\overline{z} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{z} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)},$$

och om $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ är okänt, så får vi

$$\underline{I_\Delta = \left(\overline{z} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{z} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)}, \text{ där } \underline{s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2}}.$$

Exempel. *Tio personer vägs före och efter två månaders användning av dieten B. De uppmätta vikterna är sammanfattade i följande tabell.*

Person	1	2	3	4	5	6	7	8	9	10
Vikt före diet (y_i)	80	65	79	87	104	115	76	95	85	88
Vikt efter diet (x_i)	74	61	75	85	96	112	77	92	80	84
Viktdifferens (z_i)	6	4	4	2	8	3	-1	3	5	4

Från stickprovet $z_1, ..., z_{10}$ *får vi* $\overline{z} = 3.8$ *och* $s = 2.394$. *Eftersom standardavvikelsen är okänd och* $t_{0.025}(9) = 2.26$, *så är ett konfidensintervall för* Δ *med konfidensgrad* 95%

$$I_\Delta = \Big(3.8 - 2.26\frac{2.394}{\sqrt{10}}, 3.8 + 2.26\frac{2.394}{\sqrt{10}}\Big) = (3.8 - 1.711, 3.8 + 1.711) = (2.09, 5.51).$$

Med konfidensgrad 95% *så ger dieten B en förväntad viktnedgång på mellan* 2 *och* 5.5 *kg. Notera att* 4 *av de* 10 *uppmätta viktdifferenserna ligger utanför konfidensintervallet, vilket är mer än förväntat.*

Konfidensintervall med approximativ konfidensgrad

Vi har hittills bara bestämt konfidensintervall för normalfördelade stickprov, men i praktiken kan stickprov komma från många olika fördelningar. Tack vare centrala gränsvärdessatsen är **stickprovsvariablerna för många vanliga punktskattningar** ofta **approximativt normalfördelade** för stora stickprov, även om stickprovet i sig inte är normalfördelat. Vi kan därför använda metoderna för konfidensintervall för normalfördelade stickprov även i dessa fall. Konfidensgraden blir dock inte längre exakt $1-\alpha$ utan bara approximativt $1-\alpha$.

Låt exempelvis x vara en observation av $X \in \mathrm{Bin}(n, p)$, där p är okänt. En punktskattning för p är $p^*_{\text{obs}} = x/n$, men det är svårt att konstruera ett exakt konfidensintervall för p. För stora n (närmare bestämt då $np(1-p) \geq 10$) gäller approximativt att

$$X \in \mathrm{N}(np, \sqrt{np(1-p)}), \qquad \text{så} \qquad p^* = \frac{X}{n} \in \mathrm{N}(p, \sqrt{p(1-p)/n}).$$

Standardavvikelsen av p^* beror på den okända parametern p, men kan approximeras med medelfelet $d = \sqrt{p^*_{\text{obs}}(1-p^*_{\text{obs}})/n}$. Vi får då följande konfidensintervall med approximativ konfidensgrad $1-\alpha$.

$$\underline{I_p = \Big(\frac{\overline{x}}{n} - \lambda_{\alpha/2}d, \frac{\overline{x}}{n} + \lambda_{\alpha/2}d\Big)}, \text{ där } \underline{d = \sqrt{p^*_{\text{obs}}(1-p^*_{\text{obs}})/n}}.$$

Notera att vi använder normalfördelningens kvantiler fast standardavvikelsen är okänd och detta beror på att konfidensgraden ändå endast är approximativ.

Exempel. *Antalet nederbördsdagar i Stockholm under ett år är Bin*$(365, p)$*-fördelat, där* p *är sannolikheten att det regnar eller snöar en given dag. Givet att Stockholm hade* 166 *nederbördsdagar under* 2015, *så är ett approximativt konfidensintervall för* p *med konfidensgrad* 95%

$$I_p = \Big(\frac{166}{365} - 1.96\sqrt{\frac{166}{365}\Big(1-\frac{166}{365}\Big)/365}, \frac{166}{365} + 1.96\sqrt{\frac{166}{365}\Big(1-\frac{166}{365}\Big)/365}\Big) = (0.40, 0.50),$$

så med konfidensgrad 95% *är sannolikheten för nederbörd i Stockholm en given dag mellan* 40 *och* 50 *procent.*

Vi kan använda konfidensintervallet för att avgöra i hur många dagar som vi måste registrera nederbörden för att uppnå en viss noggrannhet i resultatet. Om vi vill ha ett konfidensintervall med konfidensgrad 95% som är högst två procentenheter långt, så måste

$$2\lambda_{\alpha/2}\sqrt{p^*_{\text{obs}}(1-p^*_{\text{obs}})/n} < 0.02 \quad \underbrace{\Rightarrow}_{p(1-p)\leq 1/4} \quad 1.96\frac{1}{2\sqrt{n}} < 0.01$$

$$\Rightarrow \quad n > \left(\frac{1.96}{2\cdot 0.01}\right)^2 = 9604 \approx 27 \text{ år.}$$