

Konfidensintervall

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 undersökas på detta sätt är "Ger medicinen A större effekt än placebo?" och "Ger man ner i vikt av dieten B?"

I det första fallet kan en grupp patienter slumpmässigt delas in i två undergrupper där ena undergruppen får medicin A och andra undergruppen får placebo. Efter behandlingstidens slut mäts patienternas tillstånd med lämplig metod. Vi får

Placebo:	$n_1 = 43$	$\bar{x} = 21,57$	$s_1 = 3,87$	(värde över 20 mot+ svarar här sjukdom)
Medicin A:	$n_2 = 33$	$\bar{y} = 20,38$	$s_2 = 3,91$	

Vi antar att de båda gruppernas resultat är stickprov från $N(\mu_1, \sigma_1)$ -respektive $N(\mu_2, \sigma_2)$ -fördelningar. Vi vill bestämma ett konfidensintervall för $\mu_1 - \mu_2$. Låt $x_1, \dots, x_{n_1}$ beteckna stickprovet från $N(\mu_1, \sigma_1)$ och $y_1, \dots, y_{n_2}$ stickprovet från $N(\mu_2, \sigma_2)$ och anta att alla motsvarande s.v. är oberoende. Antag först att σ_1 och σ_2 är kända. Från tidigare sats om normalfördelningen vet vi att $\bar{X} - \bar{Y} \in N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$, så

$$\text{pivot-} \rightarrow \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \in N(0,1) \Rightarrow \text{IP} \left(-\lambda_{\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < \lambda_{\alpha/2} \right) = 1 - \alpha$$

och ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ är $I_{\mu_1 - \mu_2} = (\bar{X} - \bar{Y} - \lambda_{\alpha/2} D, \bar{X} - \bar{Y} + \lambda_{\alpha/2} D)$, där $D = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Antag nu att $\sigma_1 = \sigma_2 = \sigma$ är okända. Eftersom $D = \sigma \sqrt{1/n_1 + 1/n_2}$ då är okänt, behöver vi skatta D med ett lämpligt medelfel. Om vi väljer

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{så gäller} \quad \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s \sqrt{1/n_1 + 1/n_2}} \in t(n_1 + n_2 - 2)$$

viktat medelvärde av s_1^2 och s_2^2 pivotvariabel

Motsvarande tvåsidiga konfidensintervall för $\mu_1 - \mu_2$ är:

$$I_{\mu_1 - \mu_2} = (\bar{X} - \bar{Y} - t_{\alpha/2}(f)d, \bar{X} - \bar{Y} + t_{\alpha/2}(f)d), \quad \text{där } f = n_1 + n_2 - 2 \text{ och } d = s \sqrt{1/n_1 + 1/n_2}.$$

Ex: I den medicinska undersökningen kan vi anta att $\sigma_1 = \sigma_2 = \sigma$ är okända. Vi får $s^2 = 15,11$ och gränserna för ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ med konfidensgrad 95% blir

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}(43+33-2) s \sqrt{1/43 + 1/33} = 1,19 \pm 1,793 \Rightarrow I_{\mu_1 - \mu_2} = (-0,60, 2,94)$$

Med konfidensgrad 95% är skillnaderna mellan medeleffekterna för medicinen A och placebo mellan -0,6 och 2,9. Notera att vi nte 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 nte en fråga om sannolikhet.

Situationen i det andra fallet (dieten) är snarlik men kräver en annan analys eftersom det nu är samma personer som undersöks före och efter dieten. Stickproven $x_1, \dots, x_n$ och $y_1, \dots, y_n$ är därför beroende. Variationen inom varje stickprov är generellt större än skillnaderna 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 $N(\mu_1, \sigma_1)$ -fördelning och att y_i kommer från en $N(\mu_1 + \Delta, \sigma_2)$ -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 Δ mellan stickproven $x_1, \dots, x_n$ och $y_1, \dots, y_n$ så undersöker vi stickprovet av differenser $z_1, \dots, z_n$ där $z_i = x_i - y_i$. Eftersom y_i och x_i kommer från $N(\mu_1 + \Delta, \sigma_2)$ - respektive $N(\mu_1, \sigma_1)$ -fördelningar, så kommer z_i från en $N(\Delta, (\sigma_1^2 + \sigma_2^2))$ -fördelning. Vi har därför reducerat problemet till att bestämma konfidensintervall för väntevärdet av ett normalfördelat stickprov. Om $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ är känt för vi

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

och om $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ är okänt för vi $I_\Delta = \left(\bar{z} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{z} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right)$, där $s^2 = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2$

Ex: Tre personer

vägs före och efter en två månader lång diet.

	1	2	3	4	5	6	7	8	9	10
vikt före, y_i	80	65	79	87	104	115	76	95	85	88
vikt efter, x_i	74	61	75	89	96	112	77	92	80	84
differens, z_i	6	4	4	2	8	3	-1	3	5	4

Från stickprovet $z_1, \dots, z_{10}$ får vi $\bar{z} = 3,8$ och $s = 2,394$. Eftersom $t_{0,95}(9) = 2,26$, så är ett konfidensintervall för Δ med konfidensgrad 95%

$$I_\Delta = \left(3,8 - 2,26 \cdot \frac{2,394}{\sqrt{10}}, 3,8 + 2,26 \cdot \frac{2,394}{\sqrt{10}} \right) = (3,8 - 1,711, 3,8 + 1,711) = (2,09, 5,51)$$

så med konfidensgrad 95% ger vikten en viktminskning på mellan 2 och 5,5 kg.

Vi har hittills bara bestämt konfidensintervall för normalfördelade stickprov, men i praktiken kan stickprov komma från många olika fördelningar. Läck vare centrala gränsvärdesatsen 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 approximation av $X \in \text{Bin}(n, p)$, där p är okänt (jfr exempel på FL 9). En punktskattning för p är $p_{\text{obs}}^* = x/n$, men det är svårt att konstruera ett exakt konfidensintervall för p . För stora n ($np(1-p) \geq 10$) gäller approximativt att

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

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 $1-\alpha$.

$$I_p = \left(\frac{x}{n} - \lambda_{\alpha/2} d, \frac{x}{n} + \lambda_{\alpha/2} d \right)$$

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

Ex: Antalet nederbördsdagar i Stockholm på ett år

är $\text{Bin}(365, p)$ -fördelat,

där p är sannolikheten att det regnar en given dag. Givet att Stockholm hade 166 nederbördsdagar 2015, så är ett approximativt konfidensintervall för p med konfidensgrad 95%

$$I_p = \left(\frac{166}{365} - 1,96 \sqrt{\frac{166}{365} \frac{199}{365}}, \frac{166}{365} + 1,96 \sqrt{\frac{166}{365} \frac{199}{365}} \right) = (0,49, 0,50)$$

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

$$\lambda_{\alpha/2} \sqrt{p_{\text{obs}}^*(1-p_{\text{obs}}^*)/n} < 0,01 \quad \Rightarrow \quad 1,96 \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{n}} < 0,01 \quad \Rightarrow \quad n > \left(\frac{1,96}{2 \cdot 0,01} \right)^2 = 9604 \approx 27 \text{ år}$$