

Matematisk Statistik och Diskret Matematik, MVE051/MSG810, VT21

Timo Vilkas

Chalmers - Göteborgs Universitet

Föreläsning 11
6 maj 2021



Jämförelse av två medelvärden

På ett liknande sätt som vi jämförde proportioner i olika populationer kan man också jämföra medelvärden i två populationer.

- Antag att vi har två populationer med motsvarande medelvärde μ_1 resp. μ_2 och varians σ_1^2 resp. σ_2^2 .
- Återigen drar vi ett stickprov från vardera populationen så att stickproven är oberoende av varandra.
- Sedan beräknar vi en punktskattning för medelvärdena: $\hat{\mu}_1 = \bar{X}_1$ och $\hat{\mu}_2 = \bar{X}_2$.
- Och en för differensen $\mu_1 - \mu_2$: $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$.

KI för differensen av två medelvärden

Populationer normalfördelade (eller n stor) med kända varianser

- Om populationerna är normalfördelade (eller stickproven är av tillräcklig storlek) så är $\bar{X}_1 - \bar{X}_2$ (approximativt) normalfördelad med väntevärde $\mu_1 - \mu_2$ och varians $\sigma_1^2/n_1 + \sigma_2^2/n_2$ (jfr. föreläsning 7). *OBS: varianserna adderas då $\bar{X}_1$ och $\bar{X}_2$ är oberoende, $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$*
- Ett $100(1 - \alpha)\%$ konfidensintervall för differensen $\mu_1 - \mu_2$ ges då av

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$
 då $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = Z \sim N(0,1)$
- Ifall båda populationer har samma varians ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) blir det

$$\bar{X}_1 - \bar{X}_2 \pm \sigma \cdot z_{\alpha/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

KI för differensen av två medelvärden

Populationer normalfördelade med samma okänd varians

Antag att variansen är samma, dvs. $\sigma_1 = \sigma_2 = \sigma$, dock okänd. Då definierar vi den **kombinerade variansen** (eng. *pooled variance*) av stickproven som

$$S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

$$\text{korrig.: } S^2 = \frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n-1}$$

vvr. estimator för σ^2

där S_1^2 och S_2^2 är motsvarande stickprovsvarianser.

- S_p^2 är en väntevärdesriktig estimator för σ^2 :

$$\mathbb{E}(S_1^2) = \underbrace{\sigma_1^2}_{=\sigma^2}, \mathbb{E}(S_2^2) = \underbrace{\sigma_2^2}_{=\sigma^2} \Rightarrow \mathbb{E}(S_p^2) = \frac{n_1-1}{n_1+n_2-2} \cdot \mathbb{E}(S_1^2) + \frac{n_2-1}{n_1+n_2-2} \mathbb{E}(S_2^2) = \sigma^2 \left(\frac{n_1-1+n_2-1}{n_1+n_2-2} \right) = \sigma^2$$

- $\frac{n_1-1}{\sigma^2} S_1^2$ och $\frac{n_2-1}{\sigma^2} S_2^2$ har en χ^2 -fördelning med $n_1 - 1$ resp. $n_2 - 1$ frihetsgrader (jfr. föreläsning 8).

KI för differensen av två medelvärden

Populationer normalfördelade med samma okänd varians

- Om $Y_1 \sim \Gamma(\alpha_1, \beta)$ och $Y_2 \sim \Gamma(\alpha_2, \beta)$ är oberoende, så gäller $Y_1 + Y_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$ (se föreläsning 13)
- Då stickproven är oberoende är också $\frac{n_1-1}{\sigma^2} S_1^2 \sim \Gamma(\frac{n_1-1}{2}, 2)$ och $\frac{n_2-1}{\sigma^2} S_2^2 \sim \Gamma(\frac{n_2-1}{2}, 2)$ det och som följd

$$\chi^2_{\gamma} = \Gamma\left(\frac{\gamma}{2}, 2\right)$$

$$\frac{n_1-1}{\sigma^2} S_1^2 + \frac{n_2-1}{\sigma^2} S_2^2 = \frac{n_1+n_2-2}{\sigma^2} S_p^2 \sim \Gamma\left(\frac{n_1+n_2-2}{2}, 2\right),$$

alltså χ^2 -fördelad med $n_1 + n_2 - 2$ frihetsgrader.

- Då populationerna är normalfördelade har slumpvariabeln

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{(1/n_1 + 1/n_2)}}$$

en standard normalfördelning: $\Delta \bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$

KI för differensen av två medelvärden

Populationer normalfördelade med samma okänd varians

- Analogt fallet av en population har då slumpvariabeln

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

jfr. sats i föreläsning 8:
För $Z \sim N(0,1)$ och $Y \sim \chi^2_r$ oberoende
 $\Rightarrow T = \frac{Z}{\sqrt{Y/r}}$ är t-fördelad
med r frihetsgrader.

en t-fördelning med $n_1 + n_2 - 2$ frihetsgrader:

$$Z := \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1), \quad Y := \frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 \sim \chi^2_{n_1 + n_2 - 2} \quad (\text{se ovan; de är oberoende})$$

$$T = \frac{Z}{\sqrt{\frac{Y}{n_1 + n_2 - 2}}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \cdot \sqrt{\frac{(n_1 + n_2 - 2) \cancel{\sigma^2}}{(n_1 + n_2 - 2) S_p^2}} = \sqrt{\frac{n_1 + n_2 - 2}{Y}} \quad \text{är t-fördelad med } n_1 + n_2 - 2 \text{ frihetsgrader.}$$

KI och hypotesprövning för $\mu_1 - \mu_2$

Populationer normalfördelade (eller n stor) med samma okänd varians

- Ett $100(1 - \alpha)\%$ KI för $\mu_1 - \mu_2$ när populationerna är normalfördelade med okänd standardavvikelse $\sigma_1 = \sigma_2 = \sigma$ blir då

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{s_p^2(1/n_1 + 1/n_2)}.$$

- Precis som i fallet av en population kan vi argumentera på samma sätt för icke-normalfördelade populationer så länge n_1, n_2 är stora (och centrala gränsvärdessatsen får användas).
- För att jämföra två medelvärden kan vi alternativt genomföra en hypotesprövning med alternativet H_1 som antingen

$$\mu_1 > \mu_2, \quad \mu_1 < \mu_2 \quad \text{eller} \quad \mu_1 \neq \mu_2.$$

Strategin är då exakt samma som för en parameter (genom att betrakta $\mu_1 - \mu_2$ istället).

Exempel (Exercise 10.14 i boken)

Ett företag vill undersöka om det är värt att skaffa nya laser-handskannare som används för inventering i lagret och testar då nuvarande skannare mot en nyare modell. Följande data återger stickprov på hur många 7-inch barkoder som kunde avläsas per sekund (man utgår ifrån normalfördelad data med samma varians):

Nya	Gamla
$n_1 = 61$	$n_2 = 61$
$\bar{x}_1 = 40$	$\bar{x}_2 = 29$
$s_1^2 = 24.9$	$s_2^2 = 22.7$

- 1 Beräkna den kombinerade variansen s_p^2 .
- 2 Ange ett 90% konfidensintervall för $\mu_1 - \mu_2$.
- 3 Ger stickproven stöd för antagandet att de nya skannare är mer effektiva i genomsnitt?

Lösning till Exercise 10.14

- 1 $s_p^2 = \frac{60 \cdot 24.9 + 60 \cdot 22.7}{120} = 23.8$.
- 2 Slumpvariabeln $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$ är t -fördelad med 120

frihetsgrader.

Tabell VI återger bara värden fram till 100 frihetsgrader, vilket dock tillåter en lämplig approximation: $t_{\alpha/2} = t_{0.05} \approx 1.658$

Ett 90% KI ges alltså av

$$\left(40 - 29 \pm 1.658 \sqrt{23.8 \left(\frac{1}{61} + \frac{1}{61}\right)}\right) = (9.53, 12.47)$$

- 3 Absolut, då intervallet inte innehåller 0 och ligger på den positiva sidan.

Kan man komma fram till samma slutsats med hypotesprövning? $H_0 : \mu_1 - \mu_2 = 0$ och $H_1 : \mu_1 - \mu_2 > 0$. Teststatistiken ges av

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

som är t -fördelad med $n_1 + n_2 - 2 = 120$ frihetsgrader.

Det blir ett ensidigt test med avseende på högra svansen och kritiskt värde $t_\alpha = t_{0.1} \approx 1.289$.

Utfallet är $t = \frac{40-29}{\sqrt{23.8 \cdot 2/61}} = 12.45 \gg t_{0.1}$.

Därför, förkastar vi nollhypotesen och är 90% säkra på att alternativet H_1 stämmer istället.

Anmärkning

- För att ens kunna lägga ihop stickprovsvarianserna behöver vi anta att varianserna σ_1^2 resp. σ_2^2 i båda populationer är lika. Detta antagande kan verifieras med en tvåsidig hypotesprövning med $H_0 : \sigma_1^2 = \sigma_2^2$ och en rimlig signifikansnivå (något vi inte gått genom).
- I exemplet ovan sades inget om att antalet skannade barkoder var normalfördelat (dessvärre är det ju bara heltal, vilket talar emot en normalfördelning). Dock är stickproven tillräckligt stora så att centrala gränsvärdessatsen kan tillämpas!

Definition

För en talföljd $(a_k)_{k \in \mathbb{N}_0}$ definierar vi motsvarande **genererande funktion** (eng. *generating function*) som

$$g(x) = \sum_{k=0}^{\infty} a_k x^k.$$

t.ex. motsvarar $g(x) = 1 + \frac{x^2}{5}$
följden $a_0 = 1, a_1 = 0, a_2 = \frac{1}{5}, a_3 = a_4 = \dots = 0$

Det spelar i första hand ingen roll om summan konvergerar (och för vilka värden på x den gör det) utan ska anses (lite mer abstrakt) som en form av avancerad bokföring.

Exempel

- För $a_k = 1$ får vi en geometrisk summa (som vi redan stötte på tidigare): $g(x) = 1 + x + x^2 + x^3 + \dots = \frac{1}{1-x}$ ($|x| < 1$)

- Också $a_k = 2^k$ ger en sådan och vi får:

$$g(x) = \sum_{k=0}^{\infty} 2^k x^k = \frac{1}{1-2x} \quad (|x| < \frac{1}{2})$$

Vanliga genererande funktioner

- Mer allmänt ger $a_k = c^k$ för fast $c \in \mathbb{R}$ den genererande funktionen $g(x) = \sum_{k=0}^{\infty} c^k x^k = \frac{1}{1-cx}$ ($|cx| < 1$).

- Ur binomialsatsen $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$ samt $\binom{n}{k} = 0$ för $k > n$, kan vi dra slutsatsen att $a_k = \binom{n}{k}$, för fixt $n \in \mathbb{N}$, har den genererande funktionen:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!}$$

$$g(x) = \sum_{k=0}^{\infty} \binom{n}{k} x^k = (1+x)^n \quad (\text{jfr. föreläsning 2}).$$

- Om vi däremot låter $a_n = \binom{n+k}{k}$, för $k \in \mathbb{N}$ fixt, fås

$$g(x) = \sum_{n=0}^{\infty} \binom{n+k}{k} x^n = \frac{1}{(1-x)^{k+1}} \quad (\text{uppgift 6.17}).$$

Leksakstillämpning

Genererande funktioner kan med fördel användas för att hitta *antalet* heltalslösningar till ekvationer av typen $y_1 + y_2 + \dots + y_k = n$ för $k, n \in \mathbb{N}$.

Exempel

Antag att n stycken kanelbullar ska fördelas mellan Timo, Oskar och Karl-Gustav. Oskar vill inte ha fler än två, Timo och Karl-Gustav minst en. Timo kan äta upp till fyra stycken, Karl-Gustav inte fler än två. På hur många sätt kan man dela upp n stycken så att alla blir nöjda?

Översatt till matte letar vi efter heltalslösningar till

$y_1 + y_2 + y_3 = n$ där $0 \leq y_1 \leq 2$ är antalet Oskar får, $1 \leq y_2 \leq 4$ antalet Timo får och resterande $1 \leq y_3 \leq 2$ går till Karl-Gustav.

Leksakstillämpning

Exempel

Svaret på frågan är koefficienten framför x^n i följande genererande funktion:

$$\begin{aligned} g(x) &= (x^0 + x^1 + x^2) \cdot (x^1 + x^2 + x^3 + x^4) \cdot (x^1 + x^2) \\ &= x^8 + 3x^7 + 5x^6 + 6x^5 + 5x^4 + 3x^3 + x^2 \end{aligned}$$

För $n = 3$ är det $(y_1, y_2, y_3) = (0, 1, 2)$, $(0, 2, 1)$ samt $(1, 1, 1)$. *3 stycken*

För $n = 4$ är det $(y_1, y_2, y_3) = (0, 2, 2)$, $(0, 3, 1)$, $(1, 1, 2)$, $(1, 2, 1)$ och $(2, 1, 1)$. *5 stycken*

Tillämpning

Exempel

Hur många heltalslösningar har $y_1 + y_2 + y_3 = 15$ givet restriktionerna $y_1 \leq 5$ och $y_2 \geq 5$?

På samma sätt som ovan handlar det om koefficienten framför x^{15} i den genererande funktionen

$$\begin{aligned} g(x) &= (1 + x^1 + x^2 + x^3 + x^4 + x^5) \cdot (x^5 + x^6 + x^7 + \dots) \\ &\quad \cdot (1 + x^1 + x^2 + x^3 + \dots) \\ &= \frac{1-x^6}{1-x} \cdot \frac{x^5}{1-x} \cdot \frac{1}{1-x} = \frac{x^5 - x^{11}}{(1-x)^3}. \end{aligned}$$

Tillämpning

Exempel

Med $\frac{1}{(1-x)^3} = \sum_{n=0}^{\infty} \binom{n+2}{2} x^n$ (se ovan) fås

$$g(x) = \sum_{n=0}^{\infty} \binom{n+2}{2} x^{n+5} - \sum_{n=0}^{\infty} \binom{n+2}{2} x^{n+11}.$$

Svaret på frågan (koefficienten framför x^{15}) blir därför alltså $\binom{12}{2} - \binom{6}{2} = 51$.