

Analys av två stickprov ^{oberoende} (kap. 9)

I många statistiska undersökningar vill man jämföra två populationers parametrar, (ofta väntevärden). Anta att de två populationerna har väntevärde μ_X , μ_Y och var. σ_X^2 , σ_Y^2 . Man tar då två stickprov, ~~ett~~ ett från vardera population: X_1, X_2, \dots, X_{n_1} från den ena och Y_1, Y_2, \dots, Y_{n_2} från den andra.

Om det är populationernas väntevärden vi är intresserade av så vill vi såklart titta på stickprovsmedelvärdena \bar{X} och \bar{Y} som är våra punktskattare för μ_X och μ_Y . Kom ihåg att \bar{X} och \bar{Y} var för sig har samma egenskaper som i de tidigare kapitlen. Dvs om n stort så har vi $\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$ och $\bar{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{n})$. Men om vi är intresserade av skillnaden mellan de två väntevärdena, dvs $\mu_X - \mu_Y$, så vill vi veta fördelningarna för $\bar{X} - \bar{Y}$.

④ Det visar sig att om stickproven är oberoende av varandra så ges väntevärdet för $\bar{X} - \bar{Y}$ av $\mu_X - \mu_Y$ och variansen av $\bar{X} - \bar{Y}$ som betecknas med $\sigma_{(\bar{X} - \bar{Y})}^2$ ges av

$$\sigma_{(\bar{X} - \bar{Y})}^2 = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}$$

Enligt centrala gränsvärdesatsen är även $\bar{X} - \bar{Y}$ normalfördelad om n_1 och n_2 är stort. Vi har alltså att

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right)$$

Tack vare detta så kan vi göra konfidensintervall och ~~hypotestest~~.

hypotestest för skillnaden mellan väntevärdena: $\mu_X - \mu_Y$.
Om σ_X^2 och σ_Y^2 är okända så byter vi ut dem mot punktskattorna s_X^2 respektive s_Y^2 .

Några exempel

- Man vill undersöka två träningsmetoder, vilken ger mest viktminskning? (Exempel 1 i boken)
(i genomsnitt)
- Är det en genomsnittlig skillnad i gymnasiebetyger hos intagna studenter på Chalmers och KTH?

Disclaimer: I boken används dom \bar{X}_1 och \bar{X}_2 istället för \bar{X} och \bar{Y}

~~Skilnad i medelvärden~~

Om vi vill göra konfidsintervall för $\mu_X - \mu_Y$ har vi en följande:

Om n_1 och n_2 är stora och σ_X^2 och σ_Y^2 är kända så ges ett $(1-\alpha)\%$ -igt konfidsintervall av:

$$L = (\bar{X} - \bar{Y}) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

$$U = (\bar{X} - \bar{Y}) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

Om n_1 och n_2 är stora och σ_X^2 och σ_Y^2 är kända så ges ett $(1-\alpha)\%$ -igt konfidsintervall för skillnaden $\mu_X - \mu_Y$ av:

$$L = (\bar{X} - \bar{Y}) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$$

$$U = (\bar{X} - \bar{Y}) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$$

Män kan även under samma förutsättningar göra hypotestest för en skillnad $\mu_X - \mu_Y$. Är skillnaden skild från (större än, mindre än) något värde D_0 ?

Tvåsidigt

Ensidigt

(eller)

$H_0: \mu_X - \mu_Y = D_0$

$H_0: \mu_X - \mu_Y \geq D_0$

$H_0: \mu_X - \mu_Y \leq D_0$

$H_1: \mu_X - \mu_Y \neq D_0$

$H_1: \mu_X - \mu_Y < D_0$

$H_1: \mu_X - \mu_Y > D_0$

⑥ Om n_1 och n_2 är stora och σ_X^2 och σ_Y^2 är kända använder vi teststatistiken

$$Z_1 = \frac{(\bar{X} - \bar{Y}) - D_0}{\sigma_{(\bar{X} - \bar{Y})}}$$

där $\sigma_{(\bar{X} - \bar{Y})} = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$. Om σ_X^2 och σ_Y^2 är

okända så använder vi istället $\sigma_{(\bar{X} - \bar{Y})} = \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$
i formeln för Z_1 . I bägge fallen gäller att

$Z_1 \sim N(0,1)$ vilket betyder att vi använder normalfördelningsstabellen när vi tar fram de kritiska värdena.

Övning 6: Man vill jämföra ~~två~~ väntevärden hos två olika populationer. 400 observationer från vardera population har gett följande punktskattningar

$$\bar{x} = 5275$$

$$\bar{y} = 5240$$

$$s_x = 150$$

$$s_y = 200$$

(a) Gör ett 95% -igt konfidensintervall för att skatta skillnaden $\mu_X - \mu_Y$. Tolk resultatet.

Lösning: Eftersom vi inte vet σ_X^2 och σ_Y^2 så måste vi använda S_X^2 och S_Y^2 . Vi använder alltså formeln

$$l = (\bar{x} - \bar{y}) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$$

$$u = (\bar{x} - \bar{y}) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$$

Vi har att $n_1 = n_2 = 400$, $\bar{x} = 5275$, $\bar{y} = 5240$ (7)
 $s_x = 150$ och $s_y = 200$. Vi får då ($\alpha = 0.05 \Rightarrow z_{0.05} = 1.96$)

$$l = (5275 - 5240) - 1.96 \cdot \sqrt{\frac{150^2}{400} + \frac{200^2}{400}} = 10.5$$

$$u = (5275 - 5240) + 1.96 \cdot \sqrt{\frac{150^2}{400} + \frac{200^2}{400}} = 59.5$$

Vårt observerade konfidensintervall blir alltså $[10.5; 59.5]$

Vi är 95% säkra på att den riktiga skillnaden $\mu_x - \mu_y$ ligger i intervallet $[10.5; 59.5]$.

(b) Testa nollhypotesen $H_0: (\mu_x - \mu_y) = 0$ mot den alternativa hypotesen $H_1: (\mu_x - \mu_y) \neq 0$.

Ta fram p-värdet för testet och tolka resultatet.

Vi vill använda teststatistikan

$$Z = \frac{(\bar{x} - \bar{y}) - D_0}{\sigma_{(\bar{x} - \bar{y})}}$$

där vi sätter $\sigma_{(\bar{x} - \bar{y})}$ med $\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}} =$

$$= \sqrt{\frac{150^2}{400} + \frac{200^2}{400}} = 12.5$$

~~Vårt~~ Vårt observerade teststatistika blir alltså

$$Z = \frac{(5275 - 5240) - 0}{12.5} = 2.8$$

Vi tittar i tabellen och får fram värdet 0.4974

$$\frac{p\text{-värde}}{2} = 0.5 - 0.4974 = 0.0026 \Rightarrow p\text{-värde} = 2 \cdot 0.0026 = 0.0052$$

⑧

Våra observationer talar alltså väldigt starkt emot att $\mu_X - \mu_Y = 0$. Om det skulle gälla att $\mu_X - \mu_Y = 0$, dvs om H_0 är sann, så är sannolikheten att observera så här extrema (eller extremare) data endast 0.0052.

Vad händer om stickproven inte är stora nog?

Precis som tidigare kan man, under antagandet att populationerna är normalfördelade, göra konfidensintervall och utföra hypotes test med hjälp av t-fördelningen.