Föreläsning 12/10

Hypotestest för en proportion. (kap. 8.6)

Precis som för konfidensintervall så kan man modifiera hypotestest för ett väntevärde något så att analysen kan göras för en proportion istället. Hypoteserna kan t. ex. se ut så här (för tvåsidigt).

$$H_0: \; p = p_0$$
$$H_1: \; p \neq p_0$$

Teststatistikan blir här ($\hat{p}$ = punktskattaren)

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

som är $N(0,1)$-fördelad om n är stort.

Tumregel: $np_0 \geq 15$ och $n(1-p_0) \geq 15$

Övning 87: En undersökning visade att 401 av 835 tillfrågade amerikanska ungdomar växte upp i ett hushåll med endast en förälder. Baserat på denna information, kan du dra slutsatsen att mer än 45% av amerikanska ungdomar har växt upp i ett sådant hushåll? Testa på signifikansnivå $\alpha = 0.05$.

Lösning: Låt p vara den riktiga proportionen av amerikanska ungdomar som växt upp i enföräldershem.

Hypoteserna blir nu

$H_0: p \leq 0.45$

$H_1: p > 0.45$.

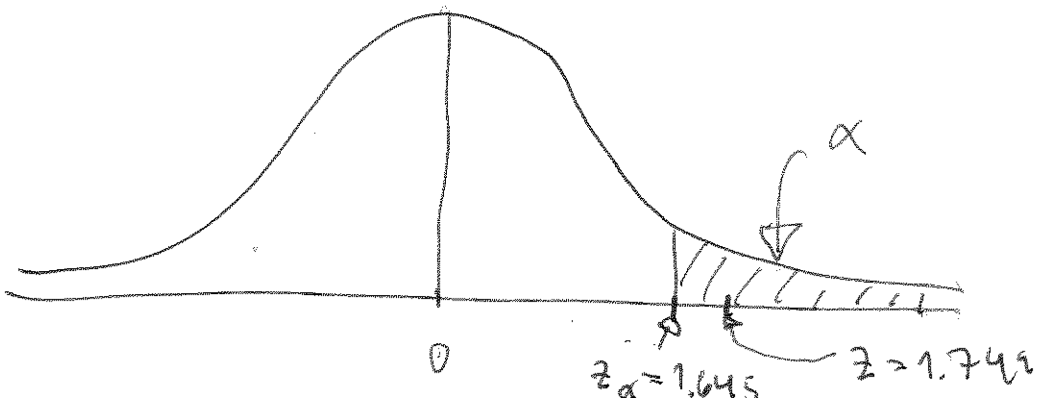
Kom ihåg att det är $p > 0.45$ vi vill statistiskt säkerställa, alltså blir det alternativhypotesen. Teststatistikan blir

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$ vilken är $N(0,1)$-fördelad.

Eftersom $\hat{p} = \frac{401}{835} = 0.4802$ blir den observerade teststatistikan

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.4802 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{835}}} = 1.749$$

Vi vill alltså jämföra detta värde med det kritiska värdet z_α. Tabellerna ger att $z_{0.05} = 1.645$. Eftersom $1.749 > 1.645$ så förkastar vi H_0 på signifikansnivå $\alpha = 0.05$



Slutsats: Ja, vi kan dra slutsatsen att mer än 45% av amerikanska ungdomar växt upp i ~~ett~~ ~~[illegible]~~ hushåll med en förälder.

Analys av två oberoende ~~[illegible]~~ stickprov. (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, ~~[illegible]~~ ett ~~från~~ från vardera population: $X_1, X_2, \ldots, X_{n_1}$ från den ena och

$Y_1, Y_2, \ldots, Y_{n_2}$ från den andra.

Om det är populationernas väntevärden vi är intresserade av så vill vi säkert 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 kapitlena. 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 fördelningen för $\bar{X} - \bar{Y}$.

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

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

Enligt centrala gränsvärdessatsen 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 för skillnaden mellan väntevärdena: $\mu_X - \mu_Y$

Om σ_X^2 och σ_Y^2 är okända så byter vi ut dem mot punktskattarna S_X^2 respektive S_Y^2

<u>Några exempel</u>

— Man vill undersöka två bantningsmetoder, vilken ger mest viktminskning? (i genomsnitt) (Exempel 1 i boken)

— Är det en genomsnittlig skillnad i gymnasiebetygen hos intagna studenter på chalmers och KTH?

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

Om vi vill göra konfidensintervall för $\mu_X - \mu_Y$ har vi nu 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 konfidensintervall 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 konfidensintervall för skillnaden $\mu_X - \mu_Y$ av:

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

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

Man 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?

<u>Tvåsidigt</u>

$H_0: \mu_X - \mu_Y = D_0$

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

<u>Ensidigt</u>

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

$H_1: \mu_X - \mu_Y < D_0$

(eller)

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

$H_1: \mu_X - \mu_Y > D_0$

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 normalfördelningstabellen när vi tar fram de kritiska värdena.

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

$\bar{x} = 5275$ $\qquad\qquad$ $\bar{y} = 5240$

$s_X = 150$ $\qquad\qquad$ $s_Y = 200$

(a) Gör ett 95%-igt konfidensintervall för att skatta skillnaden $\mu_X - \mu_Y$. Tolka 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 skattar $\sigma_{(\bar{X} - \bar{Y})}$ med $\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}} =$

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

Vår observerade teststatistika blir alltså

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

Vi tittar i tabellen och får fram värdet 0.4974

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

(B) 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 hypotestest med hjälp av t-fördelningen