

TMS136

Föreläsning 13

Jämförelser mellan två populationer

- Hittills har vi gjort konfidensintervall och tester kring parametrar i EN population
- I praktiska sammanhang är man ofta intresserad av att jämföra två populationer
- Man kanske vill jämföra andelen som röstar "röd-grönt" i Göteborg med motsvarande andel i Stockholm eller medelåldern i Halmstad med medelåldern i Falkenberg

Tester för två normalfördelade populationer

- Antag att vi har stickprov $X_{11}, \dots, X_{1n_1}$ från en normalfördelning med väntevärde μ_1 och standardavvikelse σ_1 och stickprov $X_{21}, \dots, X_{2n_2}$ från en normalfördelning med väntevärde μ_2 och standardavvikelse σ_2
- Då kan vi utföra hypotestester och skapa konfidensintervall för $\mu_1 - \mu_2$ och σ_1/σ_2

Tester för två normalfördelade populationer

- Hur testerna utförs beror på vad man anser sig veta "från början"
- Det första man behöver fundera på är om populationerna är oberoende
- Om vi ska jämföra medelåldern i Halmstad med medelåldern i Falkenberg antar vi att populationerna är oberoende men om vi mäter kolesterolhalten hos patienter före och efter behandling är populationen "före" inte oberoende av populationen "efter"

Jämförelse av väntevärden

- Vi vill testa

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

$$H_1: \mu_1 - \mu_2 \neq \Delta_0$$

- Eller motsvarande med ensidig H_1 , alltså $\mu_1 - \mu_2 > \Delta_0$ eller $\mu_1 - \mu_2 < \Delta_0$
- Precis som i tester för en population sätter vi "det vi anar" eller "vill visa" i H_1
- Om vi tror att medelåldern i Halmstad är högre än medelåldern i Falkenberg och vill visa/statistiskt säkerställa detta blir mothypotesen $\mu_1 - \mu_2 > 0$ där index 1 och 2 betyder Halmstad respektive Falkenberg

Jämförelse av väntevärden

- Den "snällaste" situationen är att populationerna är oberoende och standardavvikelserna är kända
- Om så är fallet använder vi den standard normalfördelade testfunktionen

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Jämförelse av väntevärden

- Förekastelseregionen bestäms (som vanligt) av testets signifikansnivå och om nollhypotesen är en eller tvåsidig
- Om testet på signifikansnivå α görs ensidigt nedåt ($H_1: \mu_1 - \mu_2 < \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-z_\alpha$
- Om testet på signifikansnivå α görs ensidigt uppåt ($H_1: \mu_1 - \mu_2 > \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är större än z_α
- Om testet på signifikansnivå α görs tvåsidigt ($H_1: \mu_1 - \mu_2 \neq \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-z_{\alpha/2}$ eller större än $z_{\alpha/2}$

Jämförelse av väntevärden

- Vi får ett 95% konfidensintervall för skillnaden i väntevärden ur

$$\bar{x}_1 - \bar{x}_2 \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Att talet Δ_0 inte finns med i intervallet är ekvivalent med att i testet observera testfunktionens värde i förkastelseregionen

Exempel

- Vi vill testa

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

på 0.05-nivå då vi vet att $\sigma_1 = 1$ och $\sigma_2 = 1.5$ och observerar $\bar{x}_1 = 1.8$ och $\bar{x}_2 = 0.7$ från stickprov av storlek $n_1 = 12$ och $n_2 = 10$

Exempel

- Detta ger oss det observerade värdet på testfunktionen

$$\frac{1.8 - 0.7}{\sqrt{\frac{1^2}{12} + \frac{1.5^2}{10}}} \approx 1.98$$

- Det observerade värdet är större än $z_{0.05} = 1.64$ vilket innebär att vi kan förskasta nollhypotesen
- Vi har alltså statistiskt säkerställt att väntevärdet för population 1 är större än väntevärdet för population 2

Exempel

- Vi kunde dragit samma slutsats utifrån den (ensidiga) 95%-konfidensgränsen

$$1.8 - 0.7 - 1.64 \sqrt{\frac{1^2}{12} + \frac{1.5^2}{10}} \approx 0.19$$

- Alltså, den undre konfidensgränsen är positiv så "nollan finns inte med"

Om stdavv är okända

- Om man inte känner standardavvikelsema måste de skattas från data
- För att göra detta används stickprovsstandardavvikelsema
- Hur testfunktionen ser ut beror på om standardavvikelsema kan anses lika eller ej...

Jämförelse av väntevärde

- Om standardavvikelserna är okända men kan anses lika använder vi den $t_{n_1+n_2-2}$ -fördelade testfunktionen

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

där den *poolade variansen* S_p^2 ges av

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

Jämförelse av väntevärden

- Förkastelseregionen bestäms (som vanligt) av testets signifikansnivå och om nollhypotesen är en eller tvåsidig
- Om testet på signifikansnivå α görs ensidigt nedåt ($H_1: \mu_1 - \mu_2 < \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-t_{\alpha, n_1+n_2-2}$
- Om testet på signifikansnivå α görs ensidigt uppåt ($H_1: \mu_1 - \mu_2 > \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är större än t_{α, n_1+n_2-2}
- Om testet på signifikansnivå α görs tvåsidigt ($H_1: \mu_1 - \mu_2 \neq \Delta_0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-t_{\alpha/2, n_1+n_2-2}$ eller större än $t_{\alpha/2, n_1+n_2-2}$

Exempel

- Antag att vi i situationen där vi för två oberoende stickprov (från normalfördelningar), av storlek $n_1 = 16$ respektive $n_2 = 19$, observerat $s_1 = 2.76$ och $s_2 = 1.98$ också observerat $\bar{x}_1 = 3.46$ och $\bar{x}_2 = 1.92$ och vill på 0.05-nivå testa

$$H_0: \mu_1 - \mu_2 = 1$$

$$H_1: \mu_1 - \mu_2 \neq 1$$

Exempel

- Om vi antar (baserat på erfarenhet eller tester) att standardavvikelserna är lika så vi använder den $t_{n_1+n_2-2} = t_{33}$ -fördelade testfunktionen

$$\frac{\bar{X}_1 - \bar{X}_2 - 1}{s_p \sqrt{\frac{1}{16} + \frac{1}{19}}}$$

Exempel

- Vi observerar

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} =$$

$$\frac{15 \cdot 2.76^2 + 18 \cdot 1.98^2}{33} \approx 5.60$$

Så det observerade värdet på testfunktionen blir

$$\frac{3.46 - 1.92 - 1}{\sqrt{5.60} \sqrt{\frac{1}{16} + \frac{1}{19}}} \approx 0.67$$

Exempel

- Eftersom vi testar tvåsidigt på 0.05-nivå ska vi förskasta nollhypotesen om det observerade värdet på testfunktionen är mindre än $-t_{0.025,33}$ eller större än $t_{0.025,33}$
- I appendix finner vi inte några t -kvantiler för 33 frihetsgrader så vi kikar istället i raden för 30 frihetsgrader
- Vi använder hellre ett lägre frihetsgradstal än ett högre eftersom vi då minskar risken att göra ett typ-I fel. Vi har ju att t_{30} har tyngre svansar än t_{33} så vi måste "gå längre ut" för att kunna förkasta om vi använder ett lägre frihetsgradstal

Exempel

- Vi finner i appendix att $t_{0.025,30} = 2.042$ så vi kan inte förskasta nollhypotesen
- Vi har alltså inte fog för att säga att skillnaden i väntevärden är något annat än 1
- Minns att detta inte nödvändigtvis betyder att skillnaden i väntevärden är 1 utan bara att vi inte har "bevis nog" för att skillnaden i väntevärden inte är 1

Jämförelse av väntevärden

- Vi får ett 95% konfidensintervall för skillnaden i väntevärden ur

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Att talet Δ_0 inte finns med i intervallet när ekvivalent med att i testet observera testfunktionens värde i förkastelseregionen

Exempel

- Med data enligt exemplet ovan får vi det 95%-iga konfidensintervallet

$$3.46 - 1.92 \pm \sqrt{5.60} \sqrt{\frac{1}{16} + \frac{1}{19}}$$

alltså $[0.74, 2.34]$

- Eftersom talet 1 finns med i intervallet kan vi inte förkasta nollhypotesen

Jämförelse av väntevärde

- Om standardavvikelserna är okända men ej kan anses lika använder vi den t_v -fördelade testfunktionen

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

där frihetsgradstalet v ges som heltalsdelen av

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Jämförelse av väntevärden

- Förkastelseregionen bestäms (som vanligt) av testets signifikansnivå och om nollhypotesen är en eller tvåsidig
- Om testet på signifikansnivå α görs ensidigt nedåt ($H_1: \mu_1 - \mu_2 < \Delta_0$) förskastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-t_{\alpha, \nu}$
- Om testet på signifikansnivå α görs ensidigt uppåt ($H_1: \mu_1 - \mu_2 > \Delta_0$) förskastar vi nollhypotesen om vi observerar att testfunktionen är större än $t_{\alpha, \nu}$
- Om testet på signifikansnivå α görs tvåsidigt ($H_1: \mu_1 - \mu_2 \neq \Delta_0$) förskastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-t_{\alpha/2, \nu}$ eller större än $t_{\alpha/2, \nu}$

Jämförelse av väntevärden

- Vi får ett 95% konfidensintervall för skillnaden i väntevärden ur

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Att talet Δ_0 inte finns med i intervallet när ekvivalent med att i testet observera testfunktionens värde i förkastelseregionen

Reflektion

- Att betrakta de okända varianserna som olika blir "alltid rätt" eftersom om stickprovsvarianserna blir lika så konvergerar frihetsgradtalet och testfunktionen till de för "lika varianser"
- När man räknar i datorn blir detta aldrig ett problem, det är bara när man tvingas räkna förhand som det är behändigt att anta att varianserna är lika
- Man kan övertyga sig empiriskt om att man får ungefär samma svar från sina tester oavsett antagande så länge

$$\frac{1}{2} \leq \frac{s_1^2}{s_2^2} \leq 2$$

Jämförelse av väntevärden

- Om vi har parvisa stickprov $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ från till exempel en "före/efter-studie" är ju inte populationerna oberoende
- För att testa hypoteser kring skillnaden i väntevärden mellan före och efter betraktar vi differenserna

$$D_1 = X_{11} - X_{21}, \dots, D_n = X_{1n} - X_{2n}$$

Jämförelse av väntevärden

- Om standardavvikelsen σ_D för differenserna är känd använder vi den standard normalfördelade testfunktionen

$$\frac{\bar{D} - \Delta_0}{\sigma_D / \sqrt{n}} = \frac{\frac{1}{n} \sum_{i=1}^n D_i - \Delta_0}{\sigma_D / \sqrt{n}}$$

för att testa och skapa konfidensintervall

Jämförelse av väntevärden

- Förekastelseregioner och konfidensintervall blir precis som för enstickprovsfallet, alltså för ett tvåsidigt test på signifikansnivå α förkastar vi nollhypotesen om vi observerar att kvoten är mindre än $-z_{\alpha/2}$ eller större än $z_{\alpha/2}$ och motsvarande konfidensintervall ges av

$$\bar{d} \pm z_{\alpha/2} \sigma_D / \sqrt{n}$$

Jämförelse av väntevärden

- Om standardavvikelsen σ_D för differenserna är okänd använder vi den t_{n-1} -fördelade testfunktionen

$$\frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$$

för att testa och skapa konfidensintervall, S_D är stickprovsstandardavvikelsen för differenserna

Jämförelse av väntevärden

- Förkastelseregioner och konfidensintervall blir precis som för enstickprovsfallet, alltså för ett tvåsidigt test på signifikansnivå α förkastar vi nollhypotesen om vi observerar att kvoten är mindre än $-t_{\alpha/2, n-1}$ eller större än $t_{\alpha/2, n-1}$ och motsvarande konfidensintervall ges av

$$\bar{d} \pm t_{\alpha/2, n-1} s_D / \sqrt{n}$$

Hypotestest för skillnad i varians mellan två populationer

- Att testa för skillnader i varians mellan populationer kan vara av vikt i exempelvis kvalitetstyrningssammanhang där det är önskvärt att tillverka produkter med så liten spridning som möjligt
- Vi skall alltså testa (om vi nu vill testa tvåsidigt)

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

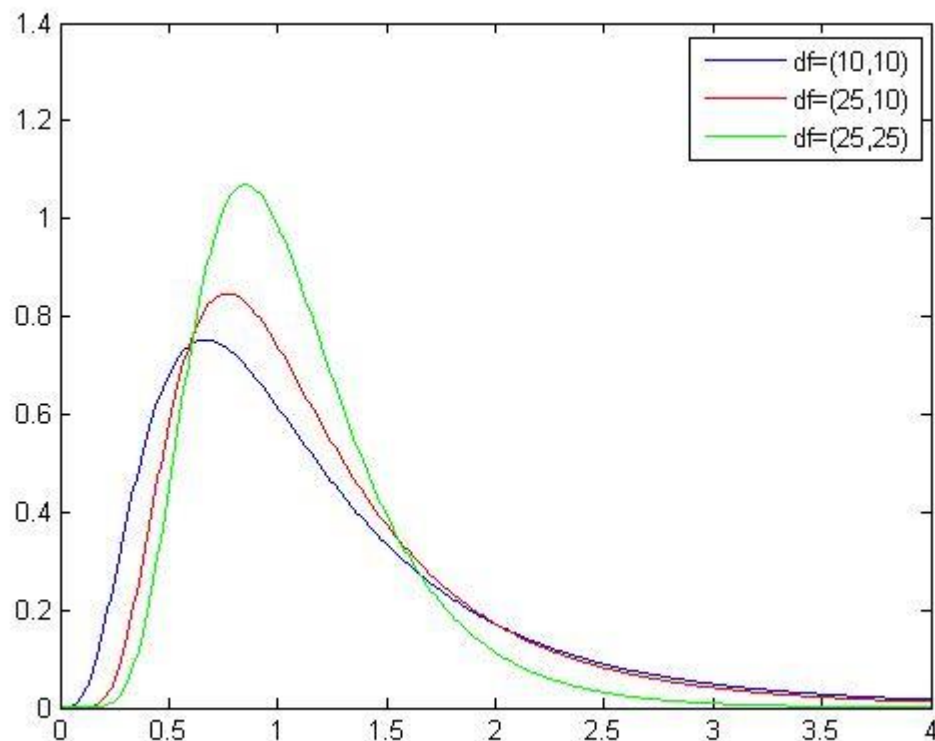
F -fördelning

- Testfunktionen som används är

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

där S_1^2 och S_2^2 är stickprovsvarianserna från population 1 respektive population 2 följer en så kallad F -fördelning med $n_1 - 1$ och $n_2 - 1$ frihetsgrader

F -fördelning (täthetsfunktion)



F -fördelning

- Under nollhypotesen om lika (populations)varianser har vi att

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

- För kvantilerna gäller att (verifiera detta)

$$f_{\alpha, v_1, v_2} = 1 / f_{1-\alpha, v_2, v_1}$$

Exempel

- Antag nu att vi för två oberoende stickprov (från normalfördelningar), av storlek $n_1 = 16$ respektive $n_2 = 19$, observerat $s_1 = 2.76$ och $s_2 = 1.98$ och vill testa om varianserna är olika, dvs

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Exempel

- Vi vill göra detta på $\alpha = 0.10$ -nivå och ska förkasta nollhypotesen om värdet på kvoten

$$\frac{s_1^2}{s_2^2}$$

blir större än $f_{0.05,15,18} = 2.27$

- Efter att ha observerat $s_1 > s_2$ behöver vi inte fundera på alternativet att $\frac{s_1^2}{s_2^2} < f_{0.95,15,18} < 1$

Exempel

- Vi har att

$$\frac{s_1^2}{s_2^2} = \left(\frac{2.96}{1.98} \right)^2 \approx 2.23 < 2.27$$

- Vi kan alltså inte förkasta nollhypotesen
- Observera (kolla tabellen i appendix) att om stickprovsstorleken hade ökat tillräckligt mycket och stickprovsvarianserna hållits konstanta hade vi förkastat nollhypotesen

Konfidensintervall

- Vi kan även skapa konfidensintervall för kvoten av varianser
- Ett tvåsidigt $(1 - \alpha) \cdot 100\%$ konfidensintervall för $\frac{\sigma_1^2}{\sigma_2^2}$ ges av

$$\left[\frac{s_1^2}{s_2^2} f_{1-\alpha/2, n_2-1, n_1-1}, \frac{s_1^2}{s_2^2} f_{\alpha/2, n_2-1, n_1-1} \right]$$

Jämförelse av proportioner

- Det kan så klart också vara av intresse att jämföra proportioner mellan populationer
- Om vi antar att vi i två oberoende populationer kan ta stickprov på ja/nej-frågor kanske vi vill statistiskt säkerställa en skillnad i "ja-frekvens" mellan populationerna
- Vi antar alltså att andelen p_1 i den ena populationen svarar ja och andelen p_2 i den andra populationen svarar ja

Jämförelse av proportioner

- Om vi vill testa för skillnad i proportioner blir våra hypoteser

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

- Om vi istället vill visa/statistiskt säkerställa att $p_1 < p_2$ blir $H_1: p_1 < p_2$
- Eller om vi istället vill visa/statistiskt säkerställa att $p_1 > p_2$ blir $H_1: p_1 > p_2$

Jämförelse av proportioner

- För utföra testet använder vi den approximativt standard normalfördelade testfunktionen

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

där $\hat{p}_1$ och $\hat{p}_2$ och n_1 och n_2 är stickprovsandelarna respektive stickprovsstorlekarna för population 1 och 2 och $\hat{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$

Jämförelse av proportioner

- Förekastelseregionen bestäms (som vanligt) av testets signifikansnivå och om nollhypotesen är en eller tvåsidig
- Om testet på signifikansnivå α görs ensidigt nedåt ($H_1: p_1 - p_2 < 0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-z_\alpha$
- Om testet på signifikansnivå α görs ensidigt uppåt ($H_1: p_1 - p_2 > 0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är större än z_α
- Om testet på signifikansnivå α görs tvåsidigt ($H_1: p_1 - p_2 \neq 0$) förkastar vi nollhypotesen om vi observerar att testfunktionen är mindre än $-z_{\alpha/2}$ eller större än $z_{\alpha/2}$

Konfidensintervall

- Motsvarande konfidensintervall ges av

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

- Eftersom "standardavvikelsebiten" ser annorlunda ut för konfidensintervallet jämfört med testfunktionen kan vi inte dra ekvivalenta slutsatser kring proportionerna på det sätt vi kunde göra för väntevärdena

Exempel

- Vi observerar att 42 av 100 slumpmässigt tillfrågade personer i Göteborg säger att de skulle rösta rödgrönt och att 72 av 200 slumpmässigt tillfrågade Stockholmare skulle rösta rödgrönt
- Kan vi baserat på detta på 0.05-nivå statistiskt säkerställa det är större andel som röstar rödgrönt i Göteborg än i Stockholm?

Exempel

- Vi har alltså hypoteserna (Göteborg är 1 och Sthlm 2)

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2$$

- Vi observerar

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.42 - 0.36}{\sqrt{0.38 \cdot 0.62 \cdot 0.015}} \approx 1.01$$

Exempel

- Vi vet att om vi testar ensidigt uppåt på 0.05-nivå så ska vi förkasta nollhypotesen om vi observerar att testfunktionen är större än 1.64
- Vi observerar 1.01 och kan alltså inte förkasta nollhypotesen
- Vi kan alltså inte säkerställa en skillnad i andelar som röstar rödgrönt
- Observera återigen att detta inte behöver betyda att nollhypotesen är sann. Om stickprovsproportionerna hållits konstanta och stickprovsstorlekarna ökats hade vi kunnat förkasta nollhypotesen...

p-värden

- p-värdet är ett tal som kan användas för att bestämma sig för att förkasta eller inte förkasta en nollhypotes
- Om p-värdet är mindre än signifikansnivån förkastas nollhypotesen
- I fallet med ensidigt test nedåt/uppåt sannolikheten att testfunktionen blir mindre/större än det observerade värdet
- I fallet med tvåsidigt test två gånger sannolikheten att testfunktionen större än beloppet av det observerade värdet

Exempel

- I fallet med andelarna observerade vi värdet 1.01 vid ett ensidigt test uppåt och vet att testfunktionen är approximativt standard normalfördelad
- p-värdet blir därför

$$P(Z > 1.01) = 1 - \Phi(1.01) \approx 0.16$$

p-värden

- Om vi arbetar med t -fördelade testfunktioner kan vi inte hitta p-värden i tabellerna eftersom vi bara har specifika t -kvantiler
- Om man vill kan man i dessa fall använda Excel eller Matlab eller något annat program som ger de sökta p-värdena