

F14, Jämförelse mellan två populationer forts..

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Man kanske vet sedan tidigare att det föreligger en skillnad d_0 mellan populationerna, och vill testa om den skillnaden fortfarande finns. Nollhypotesen uttrycks mer generellt då som

$$H_0 : \mu_1 - \mu_2 = d_0$$

och mothypotesen som något av följande alternativ

$$H_1 : \mu_1 - \mu_2 \neq d_0$$

$$H_1 : \mu_1 - \mu_2 < d_0$$

$$H_1 : \mu_1 - \mu_2 > d_0.$$

Om $d_0 = 0$ hamnar vi i det första fallet.

Hypotesprövning vid jämförelser

Vi har två populationer, och vill testa om det föreligger någon skillnad mellan populationsmedelvärdena μ_1 och μ_2 . Nollhypotesen blir då

$$H_0 : \mu_1 = \mu_2$$

och beroende på hur frågeställningen ser ut ges mothypotesen av något av följande alternativ

$$H_1 : \mu_1 \neq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

$$H_1 : \mu_1 > \mu_2.$$

Ekvivalent är att formulera nollhypotesen på följande sätt

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

och mothypotesen på något av följande sätt

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

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

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

Normalfördelade variabler med känd varians

Vid en industri finns en maskin som producerar vissa enheter. Man tänker köpa en ny maskin om den i genomsnitt producerar fler enheter. Låt X_1 och X_2 vara produktionen under en dag för den gamla respektive nya maskinen. Vi antar att X_1 och X_2 är normalfördelade variabler med okända väntevärden μ_1 och μ_2 , och kända varianser $\sigma_1^2 = 25^2$ och $\sigma_2^2 = 100^2$. Båda maskinerna prövas under 100 dygn och man erhåller $\bar{x}_1 = 1305$ och $\bar{x}_2 = 1338$. Hypoteserna formuleras som

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

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

Vi har att

$$\bar{X}_i \sim N\left(\mu_i; \frac{\sigma_i^2}{n_i}\right), i = 1, 2,$$

och att differensen

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Dessutom är

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{sqrt}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \sim N(0; 1),$$

vilket ger följande testvariabel

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\text{sqrt}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \sim N(0; 1) \text{ då } H_0 \text{ är sann.}$$

Det observerade testvärdet i stickprovet blir

$$\begin{aligned} z_{obs} &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\text{sqrt}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \\ &= \frac{(1305 - 1338) - 0}{\text{sqrt}\left(\frac{25}{100} + \frac{100}{100}\right)} = -3.20 \end{aligned}$$

P-värdet = 0.0007. Alltså H_0 förkastas. Det kan anses statistiskt påvisat att den genomsnittliga produktionen är större för maskin 2 än för maskin 1 för alla $\alpha > 0.0007$.

Normalfördelade variabler med okänd varians

Antag att mängden tjära i cigaretter varierar som normalfördelade variabler med medelvärde μ_1 och varians σ_1^2 (med filter) och medelvärde μ_2 och varians σ_2^2 (utan filter), respektive, där populationsvarianserna är okända men antas vara lika, dvs $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Testa på 5% signifikansnivå om mängden tjära i cigaretter med filter är mindre än mängden tjära i cigaretter utan filter. Hypoteserna formuleras som

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

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

Två slumpmässiga stickprov av storlek $n_1 = 9$ och $n_2 = 11$ dras oberoende av varandra från X_1 och X_2 , respektive.

Då gäller att

$$\bar{X}_i \sim N\left(\mu_i; \frac{\sigma^2}{n_i}\right), i = 1, 2,$$

och att differensen

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2; \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

Variansen σ^2 är okänd och skattas med den poolade variansen

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

Då gäller att

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{sqr}t\left(s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)} \sim t(n_1 + n_2 - 2),$$

vilket ger följande testvariabel

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\text{sqr}t\left(s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)} \sim t(n_1 + n_2 - 2) \text{ då } H_0 \text{ är sann}$$

H_0 förkastas om $t_{obs} < -1.73$.

Skattade parametrar i stickproven är $\bar{x}_1 = 0.70$, $\bar{x}_2 = 1.16$, $s_1^2 = 0.11$ och $s_2^2 = 0.21$, och den poolade variansen blir

$$s_p^2 = \frac{(9-1)0.11 + (11-1)0.21}{9+11-2} = 0.17.$$

Det observerade testvärdet i stickprovet blir

$$\begin{aligned} t_{obs} &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\text{sqr}t\left(s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)} \\ &= \frac{(0.70 - 1.16) - 0}{\text{sqr}t\left(0.17\left(\frac{1}{9} + \frac{1}{11}\right)\right)} = -2.52 \end{aligned}$$

Då $t_{obs} < -1.73$ förkastas H_0 . Det kan anses statistiskt påvisat att den genomsnittliga tjärhalten är mindre i cigaretter med filter på 5% signifikansnivå.

Stora stickprov

Inkomsten (i tusental) för kvinnor och inkomsten (i tusental) för män varierar som stokastiska variabler (X_1 och X_2) med okända väntevärden μ_1 och μ_2 , och okända varianser σ_1^2 och σ_2^2 . Vi misstänker att medelinkomsten för kvinnor är mindre än medelinkomsten för män, och vi vill därför testa

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

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

Vi drar två slumpmässiga stickprov av storlek $n_1 = 112$ och $n_2 = 98$ oberoende av varandra från populationerna.

Då fördelningen för X_1 och X_2 är okänd, utnyttjar vi vid stora stickprov ($n_i \geq 30$) att

$$\bar{X}_i \sim \text{approx } N\left(\mu_i; \frac{\sigma_i^2}{n_i}\right) \text{ enligt CGS.}$$

och att differensen

$$(\bar{X}_1 - \bar{X}_2) \sim \text{approx } N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ enl. CGS.}$$

Dessutom är

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{sqr}t\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \sim \text{approx } N(0; 1),$$

vilket ger följande testvariabel

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\text{sqr}t\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \sim \text{approx } N(0; 1) \text{ då } H_0 \text{ är sann}$$

Om populationsvarianserna σ_i^2 också är okända skattas de med stickprovsvarianserna s_i^2 .

I stickprovet skattas parametrar till $\bar{x}_1 = 19.2$, $\bar{x}_2 = 22.6$, $s_1^2 = 169$ och $s_2^2 = 144$. Det observerade testvärdet i stickprovet blir

$$\begin{aligned} z_{obs} &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(19.2 - 22.6) - 0}{\sqrt{\frac{169}{112} + \frac{144}{98}}} \\ &= -1.97 \end{aligned}$$

Då p -värdet $= 0.024$ förkastas H_0 . Det kan anses statistiskt påvisat att medelinkomsten är större för män än för kvinnor för alla $\alpha > 0.024$.

Skillnader mellan andelar

Uppgift 812 Körner:

I en stor marknadsundersökning konstaterar man att 138 män av 200 föredrar en av två förpackningar. Av 300 kvinnor är det 186 som föredrar samma förpackning. Kan man påvisa någon skillnad mellan könen vad gäller preferens för denna förpackning? Hypoteserna blir

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 \neq 0$$

Då

$$1. n_1 \hat{p}_1 \hat{q}_1 > 5 \text{ och}$$

$$2. n_2 \hat{p}_2 \hat{q}_2 > 5$$

har vi att

$$\hat{P}_i \sim \text{approx} N \left(p_i; \frac{p_i(1-p_i)}{n_i} \right), \quad i = 1, 2.$$

Först noterar vi att testvariabelns fördelning gäller då nollhypotesen är sann. Om H_0 är sann är $p_1 = p_2 = p$, som skattas med

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{138 + 186}{200 + 300} = 0.648$$

Under nollhypotesen blir då

$$(\hat{P}_1 - \hat{P}_2) \sim \text{approx} N \left(0; p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

Dessutom är då

$$\frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{\hat{P}(1-\hat{P}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \text{approx} N(0; 1) \text{ då } H_0 \text{ är sann}$$

Det observerade testvärdet blir då

$$\begin{aligned} z_{obs} &= \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{0.69 - 0.62}{\sqrt{0.648(1-0.648) \left(\frac{1}{200} + \frac{1}{300} \right)}} \\ &= 1.61. \end{aligned}$$

Då P -värdet $= \Pr(|Z| > 1.61) = 2(1 - 0.9463) = 0.11$ kan H_0 inte förkastas. Det kan ej anses statistiskt påvisat att preferens av förpackning skiljer sig åt mellan kvinnor och män för alla $\alpha < 0.11$.

Parvisa observationer

Man vet av erfarenhet att en blodtrycksmedicin som säljs på marknaden sänker blodtrycket med igenomsnitt 3 enheter. Forskare har fått fram en ny blodtrycksmedicin och man vill undersöka om den är effektivare än den gamla, dvs att blodtrycket minskar med mer än tre enheter igenomsnitt. Ett slumpmässigt stickprov bestående av 5 patienter dras. På varje patient mäter man blodtrycket före och efter de har tagit medicinen. Vi drar alltså bara *ett stickprov* men får istället *två mätvärden* för varje individ. Vi har inte längre två *oberoende* stickprov.

Låt X vara mätvärdena före medicinen och Y mätvärdena efter medicinen. Eftersom det är ett litet stickprov är vi tvungna att göra följande antaganden:

$$\begin{aligned} X &\sim N(\mu_x, \sigma_x^2) \\ Y &\sim N(\mu_y, \sigma_y^2) \end{aligned}$$

Tyvärr har vi ett beroende mellan X och Y eftersom vi har mätt blodtrycket före och efter medicinen på samma individer. Detta gör att variansen $V(\bar{X} - \bar{Y})$ blir komplicerad. Istället bildar vi differenserna $D_1 = X_1 - Y_1, \dots, D_5 = X_5 - Y_5$ som är oberoende observationer från

$$D \sim N(\mu_D, \sigma_D^2).$$

Då är

$$\bar{D} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

och hypotestest görs på vanligt sätt.

Vi vill testa hypoteserna att medicinen ger en sänkning med i genomsnitt tre enheter, mot alternativet att trycket sjunker med mer än tre enheter, dvs

$$\begin{aligned} H_0 &: \mu_D = 3 \\ H_1 &: \mu_D > 3. \end{aligned}$$

Vi har följande data

Individ	Före	Efter	Diff
1	70	60	10
2	70	68	2
3	63	61	2
4	80	71	9
5	90	80	7

som ger

$$\bar{d} = \frac{\sum d}{n} = \frac{10 + 2 + 2 + 9 + 7}{5} = 6$$

och

$$\begin{aligned} s^2 &= \frac{\sum (d - \bar{d})^2}{n - 1} = \frac{(10 - 6)^2 + \dots + (7 - 6)^2}{5 - 1} \\ &= 14.5. \end{aligned}$$

Eftersom σ_D^2 är okänd och måste skattas utnyttjar vi att

$$\frac{\bar{D} - \mu_D}{\text{sqrt}\left(\frac{s^2}{n}\right)} \sim t(n-1),$$

vilket ger följande testvariabel

$$\frac{\bar{D} - \mu_{D0}}{\text{sqrt}\left(\frac{s^2}{n}\right)} \sim t(n-1) \text{ då } H_0 \text{ är sann.}$$

Det observerade värdet är

$$t_{\text{obs}} = \frac{\bar{d} - \mu_{D0}}{\text{sqrt}\left(\frac{s^2}{n}\right)} = \frac{6 - 3}{\text{sqrt}\left(\frac{14.5}{5}\right)} = 1.76.$$

Då p -värdet = 0.04 förkastas H_0 . Det kan anses statistiskt påvisat att den nya medicinen sänker blodtrycket med igenomsnitt mer än tre enheter för alla $\alpha > 0.04$.

Jämförelser av populationsvarianser

Vi vill undersöka om det finns skillnader mellan populationsvarianserna σ_1^2 och σ_2^2 i population 1 respektive population 2. Nollhypotesen formuleras som

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

och mothypotesen som en av

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

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

$$H_1 : \sigma_1^2 > \sigma_2^2.$$

Alternativt formuleras hypoteserna som

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1.$$

Låt X_1 och X_2 vara normalfördelade variabler med okända varianser σ_1^2 och σ_2^2 i population 1 respektive population 2. Då är

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1),$$

dvs F -fördelad med $n_1 - 1$ och $n_2 - 1$ frihetsgrader.

Testvariabeln ges då av

$$\frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1) \text{ då } H_0 \text{ är sann.}$$

Exempel 3 Körner (sid 380):

Legeringar framställs med två metoder. Hårdhet hos de framställda metallstyckena med metod 1 och metod 2 varierar som normalfördelade variabler med medelvärde μ_1 och μ_2 och okända varianser σ_1^2 och σ_2^2 , respektive. Man vill testa om det föreligger någon skillnad i variationerna i hårdhet hos de framställda metallstyckena. Dvs

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Två slumpmässiga stickprov av storlek $n_1 = 25$ och $n_2 = 21$ dras. Testvariabeln

$$\frac{s_1^2}{s_2^2} \sim F(24, 20) \text{ då } H_0 \text{ är sann.}$$

Stickprovsvarianserna är $s_1^2 = 150.4$ och $s_2^2 = 60.4$.
Det observerade testvärdet är

$$F_{obs} = \frac{s_1^2}{s_2^2} = \frac{150.4}{60.4} = 2.49$$

Nollhypotesen förkastas för $\alpha = 0.05$, men accepteras för $\alpha = 0.02$. Det kan anses statistiskt påvisat att variationerna i hårdhet skiljer sig mellan de två metoderna för $\alpha = 0.05$.