

F13, Jämförelse mellan två populationer.

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Normalfördelade variabler med okänd varians

Antag att variablerna X_1 och X_2 är normalfördelade med okända väntevärden μ_1 och μ_2 , respektive. Antag vidare att (de okända) varianserna är lika, dvs $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Två slumpmässiga stickprov av storlek n_1 och n_2 dras oberoende av varandra från populationerna. 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

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

dvs t-fördelad med $(n_1 + n_2 - 2)$ frihetsgrader. Ett $(1 - \alpha)$ 100 %-igt k.i. för $\mu_1 - \mu_2$ ges då av

$$(\bar{x}_1 - \bar{x}_2) \pm t(n_1 + n_2 - 2) \text{sqrt}\left(s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Samma exempel som ovan där σ_1^2 och σ_2^2 skattas med $s_1^2 = 6.5$ och $s_2^2 = 5.2$, respektive. Den poolade variansen blir då

$$s_p^2 = \frac{(10 - 1)6.5 + (15 - 1)5.2}{10 + 15 - 2} = 5.71,$$

och ett 95 %-igt k.i. ges av

$$(15 - 14) \pm 2.07 \text{sqrt}\left(5.71\left(\frac{1}{10} + \frac{1}{15}\right)\right) = 1 \pm 2.02$$

Med 95 % tillförlitlighet ligger skillnaden i medelinkomst i intervallet $(-1.0; 3.0)$.

Stora stickprov

Om fördelningen för $X_i, i = 1, 2$, är okänd, utnyttjar vi att fördelningen för $\bar{X}_i$ är approximativt normalfördelad om n_i är stort ($n_i \geq 30$). Alltså

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

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

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

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

Ett $(1 - \alpha)$ 100 %-igt k.i. för $\mu_1 - \mu_2$ ges då av

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \text{sqrt}\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right).$$

I exemplet ovan låter vi nu fördelningarna för $X_i, i = 1, 2$, vara okända samt σ_i^2 vara okända. Vi drar två slumpmässiga oberoende stickprov av storlek $n_1 = 30$ och $n_2 = 45$. Stickprovsvarianserna skattas till $s_1^2 = 6.5$ och $s_2^2 = 5.2$. Ett 95 %-igt k.i. för $\mu_1 - \mu_2$ ges då av

$$(15 - 14) \pm 1.96 \text{sqrt}\left(\frac{6.5}{30} + \frac{5.2}{45}\right) = 1 \pm 1.13.$$

Med 95 % tillförlitlighet ligger skillnaden i medelinkomst i intervallet $(-0.1; 2.1)$.

Skillnader mellan andelar

Låt p_1 vara andelen EMU-positiva i en pop. och p_2 vara andelen EMU-positiva i en annan pop. Vi vill jämföra andelarna som är positiva till EMU i de två populationerna. Två slumpmässiga stickprov av storlek n_1 och n_2 dras oberoende av varandra från populationerna.

Om följande villkor är uppfyllda

$$1. \quad n_1 p_1 q_1 > 5$$

$$2. \quad n_2 p_2 q_2 > 5$$

har vi att

$$\hat{P}_1 \sim \text{approx } N\left(p_1; \frac{p_1(1-p_1)}{n_1}\right)$$

och

$$\hat{P}_2 \sim \text{approx } N\left(p_2; \frac{p_2(1-p_2)}{n_2}\right)$$

Då följer (enligt satsen om linjära kombinationer) att

$$(\hat{P}_1 - \hat{P}_2) \sim approx N \left(p_1 - p_2; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right)$$

eftersom väntevärdet blir

$$E(\hat{P}_1 - \hat{P}_2) = E(\hat{P}_1) - E(\hat{P}_2) = p_1 - p_2,$$

och variansen blir

$$\begin{aligned} V(\hat{P}_1 - \hat{P}_2) &= V(\hat{P}_1) + V(\hat{P}_2) - 2Cov(\hat{P}_1, \hat{P}_2) \\ &= V(\hat{P}_1) + V(\hat{P}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}, \end{aligned}$$

där $Cov(\hat{P}_1, \hat{P}_2) = 0$ eftersom stickproven dras oberoende av varandra.

Dessutom är då

$$\frac{(\hat{P}_1 - \hat{P}_2) - p_1 - p_2}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \sim approx N(0; 1).$$

Ett $(1 - \alpha)$ 100 %-igt k.i. för $p_1 - p_2$ ges då av

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

Om $\hat{p}_1 = 0.36$, $\hat{p}_2 = 0.27$, $n_1 = 150$ och $n_2 = 250$ har vi att

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

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

Ett 90 %-igt k.i. för $p_1 - p_2$ ges då av

$$\begin{aligned} &(0.36 - 0.27) \pm \\ &1.64 \sqrt{\frac{0.36(1-0.36)}{150} + \frac{0.27(1-0.27)}{250}} \\ &= 0.09 \pm 0.079. \end{aligned}$$

Med 90 % tillförlitlighet ligger differensen av andelen EMU-positiva (i populationen) i intervallet (0.01; 0.17).

Parvisa observationer

I Körners exempel 10 sid (173) har vi nio slumpmässigt valda personer som får fylla i ett formulär före och efter de sett en film. Man vill sedan undersöka om filmen ökat kunskapsnivån. Man mäter kunskapsnivån på personerna före och efter de sett filmen. Vi drar alltså bara ett stickprov men får istället två mätvärden för varje individ. Vi har inte längre två oberoende stickprov.

Låt Y vara mätvärden innan filmen och X mätvärden efter filmen där

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

Låt $X_1, \dots, X_9$ vara oberoende observationer från X och $Y_1, \dots, Y_9$ oberoende observationer från Y . Vi skulle kunna undersöka effekten av filmen på kunskapsnivån genom att skatta $\mu_x - \mu_y$ med $\bar{X} - \bar{Y}$ på vanligt sätt. Tyvärr har vi ett beroende mellan X och Y eftersom vi har mätt kunskapsnivån före och efter filmen på samma individer. Detta gör att variansen $V(\bar{X} - \bar{Y})$ blir komplicerad. Istället bildar vi differenserna $D_1 = X_1 - Y_1, \dots, D_9 = X_9 - Y_9$ 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 k.i. bildas på vanligt sätt.

I exempel 10 har vi följande data:

Individ	1	2	...	9
y	42	49	...	39
x	62	67	...	51
d	20	18	...	12

Vi har att $n = 9$, $\sum d = 20 + 18 + \dots + 12 = 162$, $\sum d^2 = 20^2 + 18^2 + \dots + 12^2 = 3036$ och att stickprovsmedelvärdet och stickprovsvariansen är

$$\begin{aligned}\bar{d} &= \frac{\sum d}{n} = \frac{162}{9} = 18 \\ s^2 &= \frac{\sum (d_i - \bar{d})^2}{n - 1} = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n - 1} \\ &= \frac{3036 - \frac{162^2}{9}}{8} = 15.\end{aligned}$$

Ett 95 %-igt k.i. för μ ges av

$$\begin{aligned}&\bar{d} \pm t_{0.025}(9 - 1) \frac{s}{\sqrt{n}} \\ &= 18 \pm 2.31 \sqrt{15} = 18 \pm 3.0.\end{aligned}$$

Med 95 % tillförlitlighet ligger den genomsnittliga ökningen i kunskapsnivån μ_D i intervallet (15; 21).