NTNU

Det skapende universitet

ST0103 Brukerkurs i statistikk
Forelesning 26, 18. november 2016
Kapittel 8: Sammenligning av grupper

Bo Lindqvist

Institutt for matematiske fag

Kapittel 8: Sammenligning av grupper

Situasjon: Vi ønsker å sammenligne *to* populasjoner med populasjonsgjennomsnitt henholdsvis μ_1 og μ_2.

Vi trekker da *ett utvalg fra hver populasjon.*

To muligheter: Vi kan ha *avhengige* eller *uavhengige* utvalg.

Avhengige (parede) utvalg: De samme kilder (person, gjenstand, etc.) brukes for å få data fra de to populasjonene.

Uavhengige (uparede) utvalg: Det trekkes ett utvalg fra hver populasjon, og kildene for dataene fra de to populasjonene har ingen sammenheng med hverandre.

Eksempel

Undersøk om et nytt treningsprogram påvirker det fysiske nivået til elevene ved en videregående skole.

Populasjon 1: Alle elevene før de gjennomgår programmet.

Populasjon 2: Alle elevene etter at de har gjennomgått programmet.

Spørsmål: Er populasjon 2 i bedre form enn populasjon 1?

Uavhengige utvalg:

- Trekk 6 elever som ennå ikke har gjennomgått treningsprogrammet og test dem.
- Trekk 6 elever som har gjennomgått treningsprogrammet og test dem.

Elevene i de to utvalgene er forskjellige.
Dataene er *et sett med 6 verdier for hvert utvalg*.

Avhengige utvalg:

- Trekk 6 elever. Test dem før de gjennomgår treningsprogrammet, la dem så gjennomgå programmet og test *de samme* elevene etterpå.

Elevene i de to utvalgene er de samme. Dataene er *to verdier for hver av de 6 elevene* (såkalte pardata).

Eksempel med avhengige (parede) utvalg

Sammenligner to typer dekk A og B med hensyn på dekkslitasje.
På 6 biler monteres ett bildekk av hver type på forhjulene.
Dekkslitasje etter kjøring en viss lengde måles:

Bil nr. i	1	2	3	4	5	6
Dekk A (X_i)	133	65	103	37	102	115
Dekk B (Y_i)	125	64	94	38	90	106
Pardifferanse ($D_i = X_i - Y_i$)	8	1	9	-1	12	9

Vil basere analysen på *differansene* D_i.

Fordel: Observasjonene varierer mye, da de er påvirket av mange faktorer: Bilens tyngde, type kjøring, førerens kjørevaner etc. Slike effekter elimineres i høy grad ved å basere analysen på D_i-ene.
Dette er essensen i bruk av avhengige (parede) utvalg.

Inferens om forskjell i forventning ved bruke av avhengige utvalg og paret T-test (8.3.2)

Har nå pardata, X_i og Y_i, for hvert av n utvalgte par.
Vi ønsker å finne ut om det er forskjell på forventningsverdiene μ_1 og μ_2 i de to populasjonene. For dette ser vi på:

Pardifferansene: $D_i = X_i - Y_i$ beregnet for hvert av de n parene

Antagelser: Parene (X_i, Y_i) for $i = 1, \ldots, n$ er uavhengige av hverandre, mens de to variablene X_i og Y_i innen ett og samme par typisk er *avhengige.* Da er D_i-ene uavhengige, og vi antar at de er tilnærmet normalfordelte med

$$\begin{aligned} E(D_i) &= \mu_D \equiv \mu_1 - \mu_2 \\ SD(D_i) &= \sigma_D \end{aligned}$$

Tilbake til dekk-eksemplet: På 6 biler monteres ett bildekk av hver type på forhjulene. Dekkslitasje etter kjøring en viss lengde måles:

Bil nr. i	1	2	3	4	5	6
Dekk A (X_i)	133	65	103	37	102	115
Dekk B (Y_i)	125	64	94	38	90	106
Pardifferanse ($D_i = X_i - Y_i$)	8	1	9	-1	12	9

Beregninger:
$\bar{D} = 6.33$ (punktestimat for μ_D),
$S_D = 5.13$ (utvalgsstandardavvik for D_i-ene; punktestimat for σ_D)

For statistisk inferens om μ_D sitter vi dermed med kun ett utvalg (av D_i-er), og vi er dermed tilbake til situasjonen i Kap. 6.

Konfidensintervall og tester for forventet forskjell μ_D ved avhengige utvalg

Konfidensintervall og testing er basert på $T = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}}$, som er Student-$t$-fordelt med $df = n - 1$ frihetsgrader.

Regel 8.3: Et $(1 - \alpha)$ konfidensintervall for μ_D er gitt ved

$$\bar{D} \pm t_{\alpha/2,n-1} \cdot \frac{S_D}{\sqrt{n}}$$

I eksempel (95%):
$6.33 \pm t_{0.025,5} \cdot \frac{5.13}{\sqrt{6}} = 6.33 \pm 2.571 \cdot \frac{5.13}{\sqrt{6}} = 6.33 \pm 5.38$

Tester $H_0 : \mu_D = 0$ mot $H_1 : \mu_D > 0$. Testobservator er da:

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{6.33}{5.13/\sqrt{6}} = 3.03$$

og vi forkaster H_0 hvis $\alpha = 0.05$. Kritisk verdi: $t_{0.025,5} = 2.015$.

Inferens om forskjell i forventning ved bruk av uavhengige utvalg og uparet T-test (8.3.1)

Populasjon 1:	*Populasjon 2*
μ_1 forventning (populasjonsgjennomsnitt)	μ_2 forventning (populasjonsgjennomsnitt)
σ_1 populasjonsstandardavvik	σ_2 populasjonsstandardavvik
n_1 antall observasjoner	n_2 antall observasjoner
X_i observasjoner	Y_i observasjoner
$\bar{X}$ utvalgsgjennomsnitt	$\bar{Y}$ utvalgsgjennomsnitt
S_1 utvalgsstandardavvik	S_2 utvalgsstandardavvik

- Vi er nå interessert i $\mu_1 - \mu_2$,
- som har punktestimat $\bar{X} - \bar{Y}$

Fordeling for $\bar{X} - \bar{Y}$

Antagelse: Uavhengige utvalg av størrelse n_1 og n_2 trekkes tilfeldig fra normalfordelte populasjoner.

Da er $\bar{X} - \bar{Y}$ normalfordelt med

1. forventning

$$\mu_1 - \mu_2$$

2. varians

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Dette betyr at

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

er standard normalfordelt og kan brukes til inferens om $\mu_1 - \mu_2$ hvis σ_1 og σ_2 er kjente.

Hvis σ_1 og σ_2 er ukjente, kan disse erstattes med S_1 og S_2. Da blir dette tilnærmet Student-t-fordelt med et antall df gitt ved en komplisert formel. Klassisk teori (og boka) antar at $\sigma_1 = \sigma_2$ og estimerer den felles σ^2 ved den såkalte interpolerte varians (Definisjon 8.1), S_p^2. Dette leder til at

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

er (eksakt) Student-t-fordelt med $df = n_1 - n_2 - 2$ frihetsgrader (se neste side).

Definisjon 8.1. Interpolert varians

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

(Vi ser at dette er et vektet gjennomsnitt mellom S_1^2 og S_2^2.)

For å teste om det er forskjell i forventningene kan vi teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } \mu_1 \neq \mu_2$$

Da er testobservatoren for den såkalte **to-utvalgs T-test**:

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Student-t-fordelt med $df = n_1 - n_2 - 2$ frihetsgrader *hvis H_0 gjelder.*

Vi forkaster H_0 hvis $T < -t_{\alpha/2,n_1+n_2-2}$ eller $T > t_{\alpha/2,n_1+n_2-2}$.

Fra eksamen i ST0202 24. mai 2003

Oppgave 1

Vekta (i kilogram) til forsvarsspillerne, x, og til angrepsspillerne, y, i Molde Fotballklubbs A-stall (MFK) er slik:

x	79	83	88	89	78	84
y	80	80	77	78	72	

Det oppgis at $\sum x = 501$, $\sum x^2 = 41935$, $\sum y = 387$ og $\sum y^2 = 29997$.

a) Finn utvalgsmiddelverdiene og utvalgsstandardavvikene for de to utvalgene.

Anta at vi kan betrakte forsvarsspillerne og angrepsspillerne i MFK som uavhengige tilfeldige utvalg fra henholdsvis populasjonen av alle forsvarsspillere og populasjonen av alle angrepsspillere på høyt nivå.

b) Foreslå en testmetode for å undersøke om det er noen forskjell i gjennomsnittsvekta til forsvarsspillere og angrepsspillere på høyt nivå. Gjør greie for antakelsene for testmetoden.

c) Utfør testen med signifikansnivå $\alpha = 0{,}10$.

Løsning:

μ_1 er forventet vekt for forsvarsspiller

μ_2 er forventet vekt for angrepsspiller

a) $\bar{X} = 501/6 = 83.5$, $\bar{Y} = 387/5 = 77.4$

$$S_1 = \sqrt{\frac{(\sum X_i^2) - (\sum X_i)^2/n_1}{n_1 - 1}} = \sqrt{\frac{41935 - (501)^2/6}{6 - 1}} = 4.51$$

$$S_1 = \sqrt{\frac{(\sum Y_i^2) - (\sum Y_i)^2/n_2}{n_2 - 1}} = \sqrt{\frac{29997 - (387)^2/5}{5 - 1}} = 3.29$$

Interpolert standardavvik:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{5 \cdot 4.51^2 + 4 \cdot 3.29^2}{9}} = 4.01$$

b) Bruker uparet T-test. Utvalgene må være uavhengige og tilfeldige, fra normalfordelte populasjoner (viser seg rimelig for vekt). Antar også $\sigma_1 = \sigma_2 = \sigma$.
Tester $H_0 : \mu_1 - \mu_2 = 0$ mot $H_1 : \mu_1 - \mu_2 \neq 0$

c) Testobservator

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{83.5 - 77.4}{4.01\sqrt{\frac{1}{6} + \frac{1}{5}}} = 2.51$$

Hvis H_0 gjelder er T Student-t-fordelt med $df = 9$.

Siden oppgaven sier at $\alpha = 0.10$ forkaster vi H_0 hvis
$T < -t_{\alpha/2,n_1+n_2-2} = -t_{0.05,9} = -1.833$ (tabell), eller hvis
$T > t_{0.05,9} = 1.833$.

Vi forkaster altså H_0 og påstår H_1 siden 2.51 > 1.833.

Vi ville også forkastet hvis $\alpha = 0.05$, siden $t_{0.025,9} = 2.262$. Men vi ville ikke forkastet hvis $\alpha = 0.02$, siden $t_{0.01,9} = 2.821$. p-verdien *(det minste signifikansnivå som vil gi forkastning)* er derfor mellom 0.02 og 0.05 (0.033 ifølge MINITAB).

Konfidensintervall for $\mu_1 - \mu_2$ ved uavhengige utvalg

Et $(1-\alpha)$ *konfidensintervall for* $\mu_1 - \mu_2$ *er gitt ved*

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

I det siste eksemplet gir dette følgende 95% konfidensintervall:

$$83.5 - 77.4 \pm 2.262 \cdot 4.01 \sqrt{\frac{1}{6} + \frac{1}{5}} = 6.10 \pm 5.49$$

dvs. (0.61,11.59)