STK1110 Høsten 2021

Hypotesetesting og konfidensintervaller for to utvalg

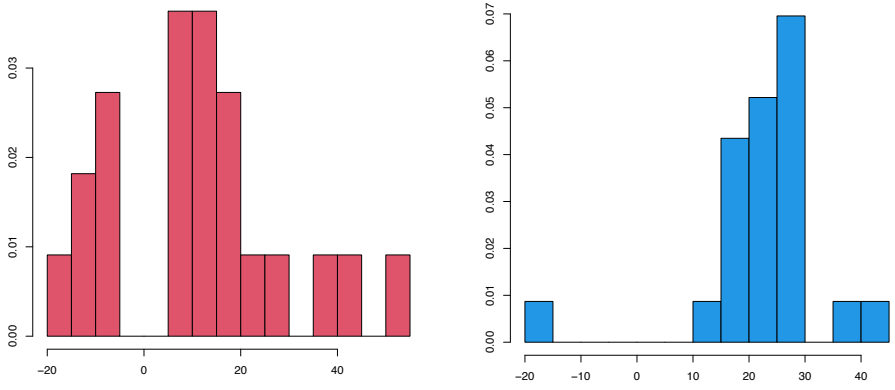
Tilsvarer Avsnitt 10.1 og 10.2

Ingrid Hobæk Haff
Matematisk institutt
Universitetet i Oslo

Eksempel

I et forsøk ble 22 rotter utsatt for ozon (behandlede), mens 23 andre ikke ble det (kontroll). Så registrerte en rottenes vektøkning i gram i løpet av en uke.

Behandlede rotter **Kontrollgruppe**



- Kan vi med rimelig grad av sikkerhet si at det er en forskjell i vektøkning mellom behandlede og ubehandlede rotter?
- Kan vi gi et anslag på forskjellen i vektøkning?
- Og hvor sikkert er dette anslaget?

Hypotesetester og konfidensintervaller for $\mu_1 - \mu_2$

- Vi antar at $X_1, \ldots, X_m$ er uif med forventning μ_1 og varians σ_1^2, $Y_1, \ldots, Y_n$ er uif med forventning μ_2 og varians σ_2^2, og at X_i-ene og Y_i-ene er uavhengige.
- Vi er interessert i å lage konfidensintervaller for og teste hypoteser knyttet til $\mu_1 - \mu_2$.
- Vi skal se på tre forskjellige tilfeller: normalfordelte data med kjente varianser, store utvalg med ukjente varianser og små normalfordelte utvalg med ukjente varianser.

Konfidensintervall for $\mu_1 - \mu_2$, normalfordelte data med kjent varians

- Anta at $X_1, \ldots, X_m \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$, med σ_1 og σ_2 kjent.
- En naturlig estimator for $\mu_1 - \mu_2$ er $\bar{X} - \bar{Y}$.
- Vi har

$$\mathrm{E}(\bar{X} - \bar{Y}) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = \mu_1 - \mu_2$$

$$\mathrm{V}(\bar{X} - \bar{Y}) \overset{uavh.}{=} \mathrm{V}(\bar{X}) + (-1)^2 \cdot \mathrm{V}(\bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

.

- Da $\bar{X}$ og $\bar{Y}$ er normalfordelt, må $\bar{X} - \bar{Y}$ også være det, slik at

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

og

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1).$$

Normalfordelte data med kjente varianser (forts.)

- Vi får:

$$\mathrm{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

slik at

$$\mathrm{P}\left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \leq \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right)$$
$$= 1 - \alpha.$$

- Et $100 \cdot (1 - \alpha)\%$ konfidensintervall for $\mu_1 - \mu_2$ er da

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}.$$

Konfidensintervall for $\mu_1 - \mu_2$, store utvalg med ukjente varianser

- Anta at $X_1, \ldots, X_m$ er uif med forventning μ_1 og varians σ_1^2, $Y_1, \ldots, Y_n$ er uif med forventning μ_2 og varians σ_2^2, med σ_1 og σ_2 ukjent.
- Anta videre at m og n er så store at vi ikke trenger å anta normalfordeling for X_i-ene og Y_i-ene.
- Vi trenger å estimere σ_1^2 og σ_2^2, og forventningsrette estimatorer for disse er

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \quad \text{og} \quad S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Store utvalg med ukjente varianser (forts.)

- Nå er
$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \overset{\text{tiln.}}{\sim} N(0,1).$$

- Et tilnærmet $100 \cdot (1-\alpha)\%$ konfidensintervall for $\mu_1 - \mu_2$ er da
$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

Eksempel

Eks. 10.5 i boka.

Konfidensintervall for $\mu_1 - \mu_2$, normalfordelte data med ukjente, like varianser

- Anta at $X_1, \ldots, X_m$ er uif med forventning μ_1 og varians σ_1^2, $Y_1, \ldots, Y_n$ er uif med forventning μ_2 og varians σ_2^2, med σ_1 og σ_2 ukjent.
- Anta videre at m og n ikke er store nok til å bruke resultatene for store utvalg.
- Da trenger vi i tillegg å anta normalfordeling, altså $X_1, \ldots, X_m \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$.
- Først antar vi at $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- La

$$S_p^2 = \frac{m-1}{m+n-2} S_1^2 + \frac{n-1}{m+n-2} S_2^2.$$

Normalfordelte data med ukjente, like varianser (forts.)

- $\bar{X} - \bar{Y}$ og S_p^2 er uavhengige, S_p^2 er forventningsrett for σ^2 og $\frac{(m+n-2)}{\sigma^2} S_p^2 \sim \chi^2_{m+n-2}$.
- Da er

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}.$$

- Det gir

$$\mathrm{P}\left(-t_{\alpha/2,m+n-2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \leq t_{\alpha/2,m+n-2} \right) = 1-\alpha.$$

- Et $100 \cdot (1-\alpha)\%$ konfidensintervall for $\mu_1 - \mu_2$ er da

$$\bar{x} - \bar{y} \pm t_{\alpha/2,m+n-2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Normalfordelte data med ukjente, ulike varianser

- Anta nå at $\sigma_1^2 \neq \sigma_2^2$.
- Vi lar

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} = \frac{(\bar{X} - \bar{Y} - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2/m + \sigma_2^2/n}}{\sqrt{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right) / \left(\sigma_1^2/m + \sigma_2^2/n\right)}} = \frac{Z}{\sqrt{U/\nu}},$$

der $Z = (\bar{X} - \bar{Y} - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2/m + \sigma_2^2/n} \sim N(0,1)$ og $U/\nu = \left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right) / \left(\sigma_1^2/m + \sigma_2^2/n\right)$ er uavhengige, men U er ikke χ_ν^2-fordelt.

Normalfordelte data med ukjente, ulike varianser (forts.)

- Vi kan imidlertid tilnærme U med en χ^2_ν-fordeling, der parameteren ν justeres slik at U/ν får riktig forventning og varians.
- Det gjør vi ved å la

$$\nu = \frac{\left(\sigma_1^2/m + \sigma_2^2/n\right)^2}{\left(\frac{\sigma_1^2}{m}\right)^2/(m-1) + \left(\frac{\sigma_2^2}{n}\right)^2/(n-1)}.$$

- Da σ_1^2 og σ_2^2 er ukjent, tilnærmes denne med

$$\nu = \frac{\left(s_1^2/m + s_2^2/n\right)^2}{\left(\frac{s_1^2}{m}\right)^2/(m-1) + \left(\frac{s_2^2}{n}\right)^2/(n-1)}. \qquad (1)$$

Normalfordelte data med ukjente, ulike varianser (forts.)

- Dermed er

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \overset{\text{tiln.}}{\sim} t_\nu,$$

  der ν er gitt ved (1).

- Et tilnærmet $100 \cdot (1 - \alpha)\%$ konfidensintervall for $\mu_1 - \mu_2$ er da

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

Eksempel

Vektøkning for rotter.

Hypotesetester ang. $\mu_1 - \mu_2$, normalfordelte data med kjent varians

- Vi ønsker å teste hypoteser av typen
  - $H_0 : \mu_1 - \mu_2 \leq \Delta_0$ mot $H_a : \mu_1 - \mu_2 > \Delta_0$
  - $H_0 : \mu_1 - \mu_2 \geq \Delta_0$ mot $H_a : \mu_1 - \mu_2 < \Delta_0$
  - $H_0 : \mu_1 - \mu_2 = \Delta_0$ mot $H_a : \mu_1 - \mu_2 \neq \Delta_0$.
- Som regel er $\Delta_0 = 0$.
- Anta at $X_1, \ldots, X_m \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$, med σ_1 og σ_2 kjent.
- Vi begynner med å se på tester av typen $H_0 : \mu_1 - \mu_2 \leq \Delta_0$ mot $H_a : \mu_1 - \mu_2 > \Delta_0$.
- Da forkaster vi H_0 dersom $\bar{X} - \bar{Y} \geq k$, som tilsvarer

$$Z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \geq \frac{k - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \tilde{k}.$$

Normalfordelte data med kjent varians (forts.)

- Det gjenstår å finne $\tilde{k}$ slik at testen får signifikansnivå α.
- Vi får

$$\mathrm{P}(\text{Feil av type I}) \leq \mathrm{P}(Z \geq \tilde{k} | \mu_1 - \mu_2 = \Delta_0) = 1 - \Phi(\tilde{k}) = \alpha.$$

- Det gir $\tilde{k} = \Phi^{-1}(1-\alpha) = z_\alpha$, og vi forkaster dermed H_0 dersom $Z \geq z_\alpha$.
- En test for $H_0 : \mu_1 - \mu_2 \geq \Delta_0$ mot $H_a : \mu_1 - \mu_2 < \Delta_0$ med signifikansnivå α får vi tilsvarende ved å forkaste H_0 dersom $Z \leq -z_\alpha$.
- En test for $H_0 : \mu_1 - \mu_2 = \Delta_0$ mot $H_a : \mu_1 - \mu_2 \neq \Delta_0$ med signifikansnivå α får vi ved å forkaste H_0 dersom $Z \leq -z_{\alpha/2}$ eller $Z \geq z_{\alpha/2}$.

Normalfordelte data med kjent varians (forts.)

- Vi går tilbake til testen for $H_0 : \mu_1 - \mu_2 \leq \Delta_0$ mot $H_a : \mu_1 - \mu_2 > \Delta_0$. Da er styrkefunksjonen gitt ved:

$$\gamma(\Delta) = 1 - \Phi\left(z_\alpha - \frac{\Delta - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right), \quad \Delta = \mu_1 - \mu_2.$$

- Vi ønsker nå å bestemme m og n slik at sannsynligheten for feil av type II blir høyst β for $\Delta = \Delta' > \Delta_0$.
- Vi skal altså løse $\beta = 1 - \gamma(\Delta') = \Phi\left(z_\alpha - \frac{\Delta' - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right)$ for m og n.
- Hvis vi lar $m = n$, får vi $m = n = \left(\frac{z_\alpha + z_\beta}{\Delta' - \Delta_0}\right)^2 (\sigma_1^2 + \sigma_2^2)$.

Normalfordelte data med kjent varians (forts.)

- For tester av typen $H_0 : \mu_1 - \mu_2 \geq \Delta_0$ mot $H_a : \mu_1 - \mu_2 < \Delta_0$ blir styrkefunksjonen:

$$\gamma(\Delta) = \Phi\left(-z_\alpha - \frac{\Delta - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right).$$

- De m og n, med $m = n$, som er slik at sannsynligheten for feil av type II blir høyst β for $\Delta = \Delta' < \Delta_0$ blir gitt ved samme formel som på forrige foil.

Normalfordelte data med kjent varians (forts.)

- For tester av typen $H_0 : \mu_1 - \mu_2 = \Delta_0$ mot $H_a : \mu_1 - \mu_2 \neq \Delta_0$ blir styrkefunksjonen:

$$\begin{aligned}\gamma(\Delta) =& \Phi\left(-z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right) \\ &+ 1 - \Phi\left(z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}\right).\end{aligned}$$

- For å sannsynligheten for feil av type II blir høyst β for $\Delta = \Delta' \neq \Delta_0$ setter en $m = n = \left(\frac{z_{\alpha/2}+z_\beta}{\Delta'-\Delta_0}\right)^2(\sigma_1^2 + \sigma_2^2)$.

Hypotesetester ang. $\mu_1 - \mu_2$, store utvalg med ukjente varianser

- Anta at $X_1, \ldots, X_m$ er uif med forventning μ_1 og varians σ_1^2, $Y_1, \ldots, Y_n$ er uif med forventning μ_2 og varians σ_2^2, med σ_1 og σ_2 ukjent.
- Anta videre at m og n er så store at vi ikke trenger å anta normalfordeling for X_i-ene og Y_i-ene.
- For å teste hypoteser ang. $\mu_1 - \mu_2$ kan vi da bruke samme tester som for normalfordelte data med kjente varianser, bare at σ_1^2 og σ_2^2 byttes ut med S_1^2 og S_2^2.
- For å finne m og n slik at sannsynlighet for feil av type II blir høyst β for $\Delta = \Delta'$ i samsvar med H_a bruker en samme formler som for normalfordelte data med kjente varianser, der en bruker anslag for σ_1^2 og σ_2^2, f.eks. basert på en pilotstudie eller lignende studier.

Eksempel

Eks. 10.4 fra boka.

Hypotesetester ang. $\mu_1 - \mu_2$, normalfordelte data med ukjente varianser

- Anta nå at m og n ikke er store nok til å bruke resultatene for store utvalg.
- Da trenger vi å anta normalfordeling, altså $X_1, \ldots, X_m \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$.
- Vi vil se på de to tilfellene $\sigma_1^2 \neq \sigma_2^2$ og $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- I begge tilfeller har vi en observator på formen $T = \frac{\bar{X} - \bar{Y} - \Delta_0}{V}$, der $T \overset{(tiln.)}{\sim} t_\nu$ når $\mu_1 - \mu_2 = \Delta_0$.
- I begge tilfeller får vi følgende tester med (tilnærmet) signifikansnivå α:
  - $H_0 : \mu_1 - \mu_2 \leq \Delta_0$ mot $H_a : \mu_1 - \mu_2 > \Delta_0$: Forkast H_0 dersom $T \geq t_{\alpha,\nu}$
  - $H_0 : \mu_1 - \mu_2 \geq \Delta_0$ mot $H_a : \mu_1 - \mu_2 < \Delta_0$: Forkast H_0 dersom $T \leq -t_{\alpha,\nu}$
  - $H_0 : \mu_1 - \mu_2 = \Delta_0$ mot $H_a : \mu_1 - \mu_2 \neq \Delta_0$: Forkast H_0 dersom $T \leq -t_{\alpha/2,\nu}$ eller $T \geq t_{\alpha/2,\nu}$.

Normalfordelte data med ukjente varianser (forts.)

- Når $\sigma_1^2 \neq \sigma_2^2$, lar vi $V = \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$.
- Når $\mu_1 - \mu_2 = \Delta_0$, er da T tilnærmet t_ν-fordelt med $\nu = \frac{\left(s_1^2/m + s_2^2/n\right)^2}{\left(\frac{s_1^2}{m}\right)^2/(m-1) + \left(\frac{s_2^2}{n}\right)^2/(n-1)}$.
- Når $\sigma_1^2 = \sigma_2^2 = \sigma^2$, lar vi $V = S_p\sqrt{\frac{1}{m} + \frac{1}{n}}$, med $S_p^2 = \frac{m-1}{m+n-2}S_1^2 + \frac{n-1}{m+n-2}S_2^2$.
- Når $\mu_1 - \mu_2 = \Delta_0$, er da $T \sim t_{m+n-2}$.

Eksempel

Vektøkning for rotter.

Normalfordelte data med ukjente varianser (forts.)

- For å beregne styrkefunksjonen til testene når $\sigma_1^2 = \sigma_2^2 = \sigma^2$ trenger vi fordelingen til $T = \frac{\bar{X} - \bar{Y} - \Delta_0}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$ når $\mu_1 - \mu_2 \neq \Delta_0$.
- Dette er en såkalt ikke-sentral t-fordeling, som vi ikke skal gå nærmere inn på i dette kurset.
- I stedet kan vi bruke R-funksjonen `power.t.test()`.
- Denne kan gi oss styrkefunksjonen, sannsynlighet for feil av type II, samt m og n slik at sannsynligheten for feil av type II blir høyst β for en Δ i samsvar med H_a.