Kap 10, Toutvalgsdata, forts.

$X_1, \dots, X_m \sim f(x; \theta_1)$ uavh. Er $\theta_1 = \theta_2$

$Y_1, \dots, Y_n \sim f(x; \theta_2)$

Spesielt interessant forvent. $\mu_1 = E(X_i)$ og $\mu_2 = E(Y_i)$

Betrakter $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{Var(\bar{X} - \bar{Y})}} \sim N(0,1)$ Eksakt m. normalf. data. Tiln. via CLT generelt

Må estimere $Var(\bar{X} - \bar{Y})$ ved

$$\widehat{Var}(\bar{X} - \bar{Y}) = \begin{cases} \frac{s_1^2}{m} + \frac{s_2^2}{n} & \text{når } \sigma_1^2 = Var(X_i) \neq \sigma_2^2 = Var(Y_i) \\ s_p^2 \left(\frac{1}{m} + \frac{1}{n}\right) & \text{når } \sigma_1^2 = \sigma_2^2 = \sigma^2 \end{cases}$$

der $s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$

Dessuten $t =$ " Z innsatt $\widehat{Var}(\bar{X} - \bar{Y})$ " $\sim t_{m+n-2}$ når $\sigma_1^2 = \sigma_2^2$, eller t_ν der ν har komplisert formel.

Teste på denne bakgrunn $H_0: \mu_1 = \mu_2$ og beregne KI for $\mu_1 - \mu_2$

okt 3-12:14

Styrkebetraktninger

Med kjente σ_1 og σ_2 blir styrkefunksjonen for $H_0: \mu_1 = \mu_2$ mot $H_a: \mu_1 > \mu_2$ med nivå α, $\Delta = \mu_1 - \mu_2$

$$\gamma(\Delta) = P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{Var(\bar{X} - \bar{Y})}} > z_\alpha \mid \mu_1 - \mu_0 = \Delta\right)$$

$$= P(\text{Forkaste } H_0 \mid \mu_1 - \mu_2 = \Delta)$$

$$= P\left(\frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{Var(\bar{X} - \bar{Y})}} > z_\alpha - \frac{\Delta}{\sqrt{Var(\bar{X} - \bar{Y})}}\right)$$

$$= \Phi\left(\frac{\Delta}{\sqrt{Var(\bar{X} - \bar{Y})}} - z_\alpha\right)$$

Med fiksert Δ og $n \to \infty$, $m \to \infty$ så vil $Var(\bar{X} - \bar{Y}) \to 0$

og $\gamma(\Delta) \to 1$ når $\Delta > 0$

Med σ_1 og σ_2 ukjente kreves et numerisk integral for å finne styrkefunksjon.

Bruk software: power.t.test : R

okt 3-12:25

Hva innebærer det at $\theta_1 \neq \theta_2$.

Eks) Utv. 1 Temp 2001 – 2016

Utv. 2 Temp 1961 – 1990

Eks) Utv. 1 Røykere

Utv. 2 Ikke røykere

Er forskjellene gitt av årsaker, kausale

eller bare assosiasjoner – faktorer på noen måte knyttet til årsaker

Observasjonelle studier – må passivt observere data

Planlagt eksperiment – kan aktivt velge verdi av faktor for hver studieenhet og bør ha kontroll.

Særlig for observasjonelle studier er det stor risiko for at sammenhenger er spuriøse, skyldes

- konfunderende = sammenblandede effekter.
- lurkende (lurking) variable

okt 3-12:33

Ex) Lungekreft og røyking

Ex) Livmorhalskreft og HPV, røyking, alkohol

Ex) Fødselsvekt og mors utdannelse, røyking

Ex) Stork og fødselstall

For at vi skal ha spuriøse sammenhenger må vår variabel være knyttet/korrelert med de reelle årsakene.

Hvis de konfunderende variablene er observert, så kan vi "justere" for dem i multiple regresjonsanalyse (Kap. 12)

For et planlagt eksperiment derimot kan man randomisere

dvs. trekke tilfeldig hvem som får ulike behandlinger.

Dermed oppnås uavhengighet mellom behandling og lurkende variabler og $\theta_1 \neq \theta_2$ signifikant kan forskjellen tilskrives årsak.

okt 3-12:45

Parede data, 10.3

X_i = respons med beh 1 — Begge for studieenhet $i = 1, \dots, n$

Y_i = — " — 2

Så (X_i, Y_i) være avhengige

Er $\mu_D = \mu_1 - \mu_2 = E(Y_i) - E(X_i)$

lik eller ulik null.

Ex) X_i = ant myggstikk arm 1 med myggmiddel A

Y_i = — " — B

Bør randomisere om høyre/venstre hånd får A/B.

Ex) X_i = temperamentsproblemer barn A om mor i med(lutt) angst/svangerskapet

Y_i = — " — B — " —

Barn A og B søsken, men kan ikke randomisere

Altså naturlige å studere $D_i = X_i - Y_i$

f.ex med antagelse $D_i \sim N(\mu_D, \sigma_D^2)$

Selv om dette er sammenligning av to faktorer kan vi nå klart å transformere til ettutvalgsdata

$D_1, D_2, \dots, D_n$

Inferens følger direkte ~~fra~~ metodene i Kap 7-9.

okt 3-13:13

Vi estimerer μ_D med $\hat{\mu}_D = \bar{D} = \bar{X} - \bar{Y}$

og $s_D^2 = \frac{1}{n-1} \sum (D_i - \bar{D})^2$

Videre får vi et $(1-\alpha)100\%$ KI for μ_D ved

$$\bar{D} \pm t_{\alpha/2} \frac{s_D}{\sqrt{n}} \quad (t_{\alpha/2} \text{ fra } t_{n-1} \text{ ford.})$$

og tester $H_0: \mu_D = 0$ mot $H_a: \mu_D \neq 0$

ved $t = \frac{\bar{D}}{s_D} \sqrt{n} \sim t_{n-1}$ under H_0.

Betrakt to studieopplegg (design)

1. n personer i utv. 1 – myggmiddel A

   n — " — 2 – " B

Analyseres med vanlig toutvalgsmetodikk.

2. Paret studie. For samme individ får

   Arm 1 middel A X_i – "

   Arm 2 middel B Y_i – "

Vi antar ved(?) design at $Var(X_i) = Var(Y_i) = \sigma^2$

Ved oppsett 1 blir $Var(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2}{n}\sigma^2$

Ved oppsett 2 blir $Var(X_i - Y_i) = Var(X_i) + Var(Y_i) - 2\,Cov(X_i, Y_i)$

$= \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho)$

Dermed blir $Var(\bar{X} - \bar{Y}) = \frac{2}{n}(\sigma^2 - \rho) < \frac{2}{n}\sigma^2$ hvis $\rho > 0$

$\quad\quad = Var(\bar{D})$

Dermed er parede studier mer informative enn vanlig toutvalg med positiv ρ.

Hvis $\rho = 0$ taper vi frihetsgrader, fra $2n-2$ til $n-1$

okt 3-13:30

To binomiske variable, 10.4

Utv. 1 m forsøk, X = ant suksesser $\sim \text{Bin}(m, p_1)$

Utv. 2 n " , Y = — " — $\sim \text{Bin}(n, p_2)$

Utvalgene er uavhengige

Ønsker KI for $p_1 - p_2$ og test av $H_0: p_1 = p_2$

Estimatorer $\hat{p}_1 = \frac{X}{m}$ og $\hat{p}_2 = \frac{Y}{n}$ for p_1 og p_2.

Da er $E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2$

og $\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$

$p_1 = p_2 = p \rightarrow \quad = p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)$

og $\hat{p}_1 - \hat{p}_2 \overset{\text{tiln}}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}\right)$

Test for $H_0: p_1 = p_2$ gis ved testobs.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\text{Var}(\hat{p}_1 - \hat{p}_2)}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{\text{tiln}}{\sim} N(0,1)$$

og $\hat{p} = \frac{X+Y}{n+m}$ estimat for $p = p_1 = p_2$ under H_0.

okt 3-13:43

Forkaster f.eks. H_0 mot alt. $H_a: p_1 \neq p_2$, på nivå α når

$$|Z| > z_{\alpha/2} = (1-\alpha/2)100 \text{ persentil i } N(0,1).$$

og et tiln 95% KI for $p_1 - p_2$ gis ved

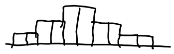
$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\hat{p}_1(1-\hat{p}_1)/m + \hat{p}_2(1-\hat{p}_2)/n}$$

Her har $p_1 \neq p_2$, bør ha generell variansestimator.

Software: binom.test ; R

Merk at utskrift fra denne kanne avvike litt fra formlene over pga.

- Kontinuitetskorreksjon
- Kan tillate eksakte metoder (Eksakt Fisher test)



okt 3-13:53