STK1100 våren 2018

Konfidensintevaller

Svarer til avsnitt 8.2 i læreboka

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Konfidensintervall for μ i store utvalg

Anta at de stokastiske variablene $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte med forventningsverdi μ og standardavvik σ

Sentralgrensesetningen gir da at

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

er tilnærmet standardnormalfordelt når n er tilstrekkelig stor

Men hvis σ ikke er kjent (som vanligvis er tilfellet), kan vi ikke bruke dette til å lage et konfidensintervall for μ

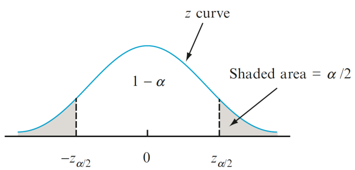
En forventningsrett estimator for σ^2 er

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

En kan vise at også

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

er tilnærmet standardnormalfordelt når n er tilstrekkelig stor (ofte nok at n er minst 40)



Derfor er

$$P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{S / \sqrt{n}} \le z_{\alpha/2}\right) \approx 1 - \alpha$$

Ved å omforme ulikhetene gir dette at

$$P\left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha$$

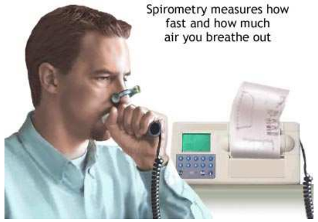
Når vi setter inn observerte verdier $x_1, x_2, \ldots, x_n$ for de stokastiske variablene $X_1, X_2, \ldots, X_n$ får vi et tilnærmet 100(1-α)% konfidensintervall for μ

$$\left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}\right]$$

Dette gjelder uansett fordeling for X_i - ene så sant n er tilstrekkelig stor

Eksempel: Måling av lungefunksjon

Et mål på lungefunksjon er FEV1 (forced expiratory volume in 1 second).



I en studie i Hordaland på 1990-tallet ble FEV1 målt for 1642 ikke-røykende, friske menn i alder 30-34 år

For FEV1-målingene var $\bar{x} = 4.48$ og $s = \sqrt{s^2} = 0.60$

Et tilnærmet 95% konfidensitervall for forventet FEV1-verdi for 30-34 år gamle menn er

$$\left[4.48 - 1.96\frac{0.60}{\sqrt{1642}}\;,\;\; 4.48 + 1.96\frac{0.60}{\sqrt{1642}}\right]$$

dvs 4.48 ± 0.03

Eksempel: Dødsulykker i trafikken i Norge og Sverige

I 2015 døde 117 personer i trafikken i Norge, mens 259 personer døde i trafikken i Sverige

Hva kan vi si om risikoen for dødsulykker i de to landene?

Er er en reell forskjell på risikoen for dødsulykker i Norge og Sverige?

For å si noe om risikoen for dødsulykker og kunne sammenligne landene, må vi ta hensyn til størrelsen av befolkningene

1. januar 2015 bodde det 5165 802 mennesker i Norge og 9 747 355 mennesker i Sverige

Vi vil anta

X = antall dødsulykker

er Poisson-fordelt med forventningsverdi

$$E(X) = \lambda w$$

der

$w =$ antall innbyggere / 100 000

λ er forventet antall dødsulykker per 100 000 personer

Generelt har vi følgende situasjon:

X er Poisson fordelt, og $E(X) = \lambda w$ der w er en kjent størrelse

Vi vil estimere λ og bestemme et konfidensintervall

En forventningsrett estimator for λ er

$$\hat{\lambda} = \frac{X}{w}$$

Standardfeilen er

$$\sigma_{\hat{\lambda}} = \sqrt{V(\hat{\lambda})} = \sqrt{\frac{\lambda}{w}}$$

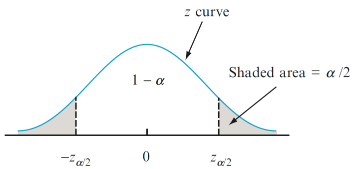
Hvis $E(X)$ er tilstrekkelig stor, er

$$\frac{X - \lambda w}{\sqrt{\lambda w}} = \frac{\hat{\lambda} - \lambda}{\sqrt{\lambda / w}} = \frac{\hat{\lambda} - \lambda}{\sigma_{\hat{\lambda}}}$$

tilnærmet standardnormalfordelt

Derfor er

$$P\left(-z_{\alpha/2} \le \frac{\hat{\lambda}-\lambda}{\sqrt{\lambda / w}} \le z_{\alpha/2}\right) \approx 1-\alpha$$



Ved å omforme ulikhetene gir dette (detaljer på forelesningen)

$$P\left(\hat{\lambda}+\frac{z_{\alpha/2}^2}{2w}-z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{w}+\frac{z_{\alpha/2}^2}{4w^2}} \le \lambda \le \hat{\lambda}+\frac{z_{\alpha/2}^2}{2w}+z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{w}+\frac{z_{\alpha/2}^2}{4w^2}}\right) \approx 1-\alpha$$

Et tilnærmet 100(1-α)% konfidensintervall for λ er:

$$\hat{\lambda}+\frac{z_{\alpha/2}^2}{2w} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{w}+\frac{z_{\alpha/2}^2}{4w^2}}$$

Når w er stor, kan vi bruke intervallet: $\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{w}}$

Norge: $x = 117$ $w = 51.658$

Estimat: $\hat{\lambda} = \frac{117}{51.658} = 2.26$

95% konfidensintervall: $[1.89\ ,\ 2.71]$

Enkelt konfidensintervall: $[1.85\ ,\ 2.68]$

Sverige: $x = 259$ $w = 97.474$

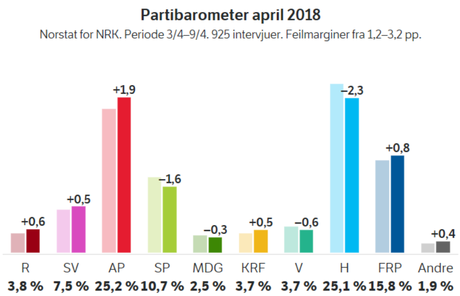
Estimat: $\hat{\lambda} = \frac{259}{97.474} = 2.66$

95% konfidensintervall: $[2.35\ ,\ 3.00]$

Enkelt konfidensintervall: $[2.33\ ,\ 2.98]$

Meningsmåling

Spør et tilfeldig utvalg på 925 personer hva de ville ha stemt hvis det hadde vært valg



233 ville ha stemt Ap

La p være andelen i befolkningen som ville ha stemt Ap hvis det hadde vært stortingsvalg

Et estimat for p er $\hat{p} = \frac{233}{925} = 0.252$

Vi vil finne et 95 % konfidensintervall for p

Generelt har vi følgende situasjon:

Vi har observert verdien y av en stokastisk variabel Y som er binomisk fordelt med n forsøk og «suksessannsynlighet» p

Vi antar at np og $n(1-p)$ begge er minst lik ti, slik at vi kan bruke tilnærmelsen til normalfordelingen (jf. sidene 189-190 og 302)

Vi vil bestemme et tilnærmet 100(1-α)% konfidensintervall for p

En estimator for p er $\hat{p} = \dfrac{Y}{n}$

Vi har at

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

er tilnærmet standardnormalfordelt

Det gir at

Løs ulikheten

$$P\left(-z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \le z_{\alpha/2}\right) \approx 1 - \alpha$$

Hvis vi løser de ulikhetene (jf. forelesningen) får vi at

$$P\left(\tilde{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n} \le p \le \tilde{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}\right) \approx 1 - \alpha$$

der $\tilde{p} = \dfrac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n}$

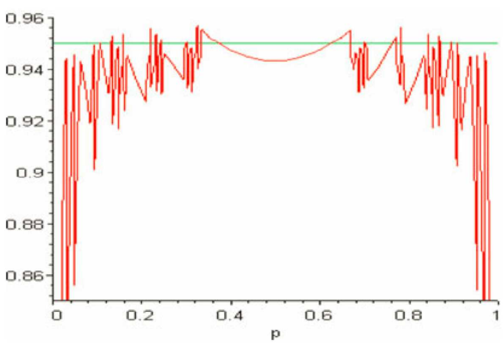
Et tilnærmet 100(1-α)% konfidensintervall for p er

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}$$

Hvis n er stor nok, er det vanlig å bruke det enklere intervallet

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad (*)$$

Men dette intervallet har dårligere egenskaper enn det på forrige slide når n er «moderat stor»



Faktisk konfidens-koeffisient for (*) for ulike verdier av p når n = 100

For meningsmålingen gir formelen nederst på slide 14 følgende 95% konfidensintervall for Ap's oppslutning:

$$[0.225\ ,\ 0.281]$$

mens det enkle intervallet (*) gir

$$0.252 \pm 0.028$$

dvs.

$$[0.224\ ,\ 0.280]$$

Konfidensintervall for binomisk p

$Y \sim \text{binomisk}(n, p) \qquad \hat{p} = \frac{Y}{n}$

Normalfordeling:

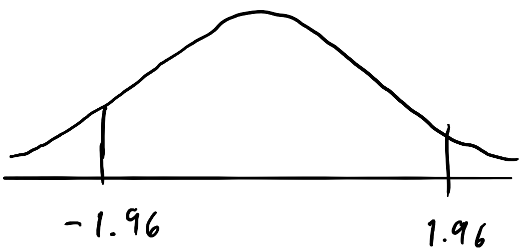
$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Har også at

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{\text{tiln}}{\sim} N(0,1)$$

Har dermed at

$$P\left(-1.96 \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le 1.96\right) \approx 0.95$$



Løser ulikheten og får

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

Tilnærmet 95% konfidensintervall

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For nettingsavdelingen fikk vi $\hat{p} = 0.252$ og $n = 225$

95% konf. intervall blir

$$0.252 \pm 1.96 \sqrt{0.25 \cdots}$$