# STK1100 våren 2023

## Mere om konfidensintervaller

Svarer til avsnitt 8.2 i læreboka

Matematisk institutt
Universitetet i Oslo

## Konfidensintervall for $\mu$ i store utvalg

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

Sentralgrensesetningen gir da at

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

er tilnærmet standardnormalfordelt når $n$ er tilstrekkelig stor

Men hvis $\sigma$ ikke er kjent (som vanligvis er tilfellet), kan vi ikke bruke dette til å lage et konfidensintervall for $\mu$

En forventningsrett estimator for $\sigma^2$ er

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

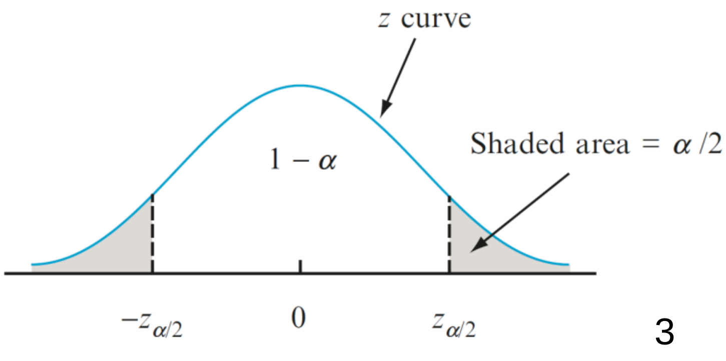
En kan vise at også

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

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

Derfor er

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



Ved å omforme ulikhetene gir dette at

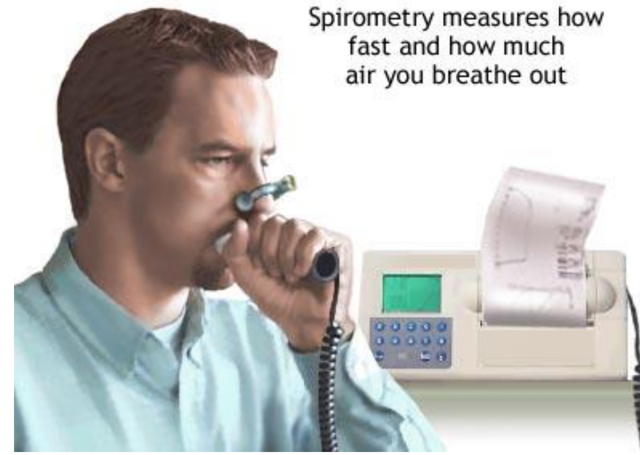
$$P\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1-\alpha$$

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

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

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 utført i Bergen 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

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

dvs $4.48 \pm 0.03$

## Eksempel: Dødsulykker i trafikken i Norge og Sverige

I 2017 døde 107 personer i trafikken i Norge, mens 251 personer døde i trafikken i Sverige

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

Er det 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 på befolkningene

1. januar 2017 bodde det 5.26 millioner mennesker i Norge og 10.00 millioner mennesker i Sverige

Vi vil anta at

$X$ = antall dødsulykker

er Poisson-fordelt med forventningsverdi

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

der

$w =$ antall innbyggere / 100 000

$\lambda$ 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 $\lambda$ og bestemme et konfidensintervall

En forventningsrett estimator for $\lambda$ er

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

Standardfeilen er

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

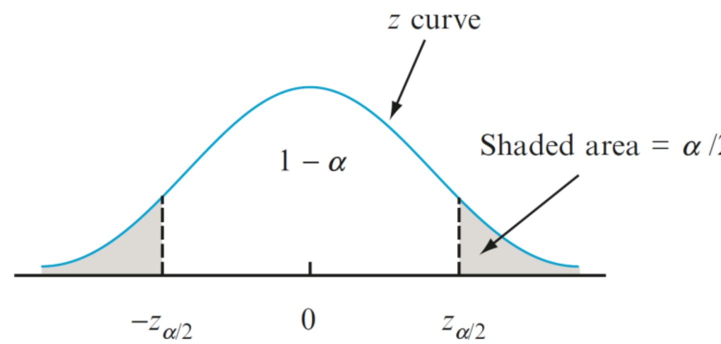
Hvis $E(X)$ er tilstrekkelig stor, er

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

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-$\alpha$)% konfidensintervall for $\lambda$ 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:** $x = 107$ $w = 52.6$

Estimat: $\hat{\lambda} = \frac{107}{52.6} = 2.03$

95% konfidensintervall: $[1.68\ ,\ 2.46]$

Enkelt konfidensintervall: $[1.65\ ,\ 2.42]$

**Sverige:** $x = 251$ $w = 100.0$

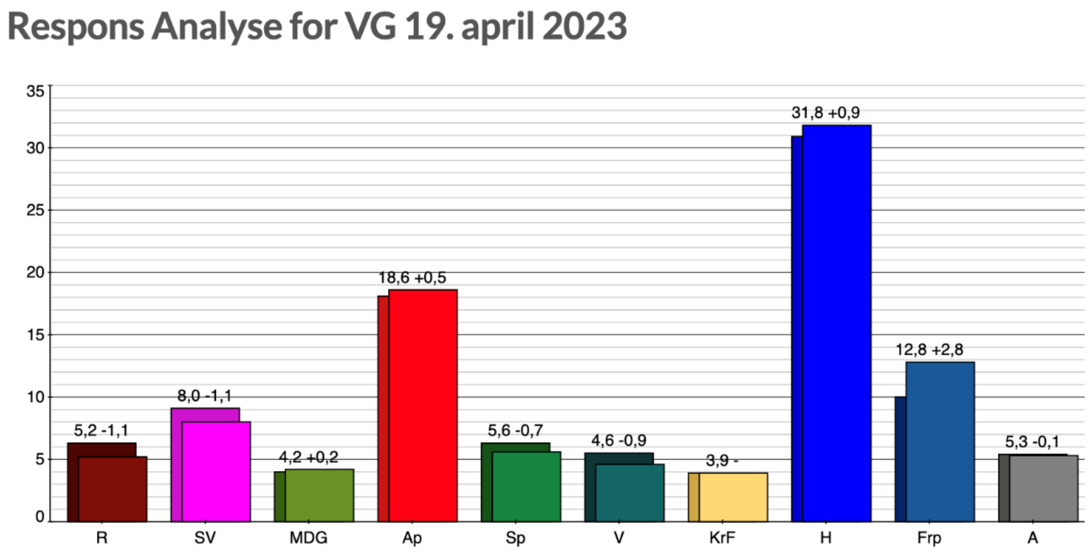
Estimat: $\hat{\lambda} = \frac{251}{100.0} = 2.51$

95% konfidensintervall: $[2.22\ ,\ 2.84]$

Enkelt konfidensintervall: $[2.20\ ,\ 2.82]$

# Meningsmåling

Spør et tilfeldig utvalg på 680 personer hva de ville ha stemt hvis det hadde vært valg




38 ville ha stemt Sp

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

Et estimat for *p* er $\hat{p} = \frac{38}{680} \approx 0.056$

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 10, slik at vi kan bruke tilnærmingen til normalfordelingen (jf. sidene 223-225 og 375)

Vi vil bestemme et tilnærmet $100(1-\alpha)\%$ 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

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

Hvis vi løser ulikhetene, får vi at (dropper utledning)

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

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

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

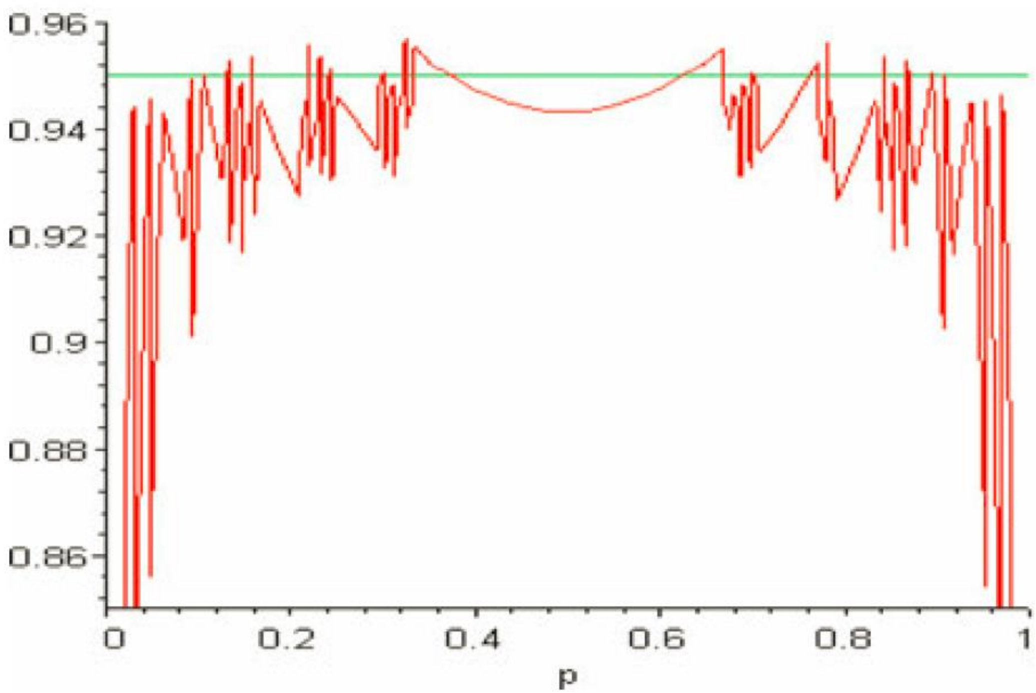
Et tilnærmet 100(1-$\alpha$)% konfidensintervall for *p* er

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

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

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

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 foil 14 følgende 95% konfidensintervall for Sp's oppslutning:

$$[0.041, 0.076]$$

mens det enkle intervallet (*) gir

$$0.056 \pm 0.017$$

dvs.

$$[0.039, 0.073]$$

En «feilmargin» for estimatet for Sp’s oppslutning er 1.7 prosentpoeng