NTNU

Kunnskap for ei betre verd

# Observatorar og utvalsfordeling

Torstein Fjeldstad

Institutt for matematiske fag, NTNU

25.02.2019

# I dag

- Til no i emnet
- Observatorar
- Utvalsfordelingar
- Sentralgrenseteoremet

# Til no i emnet

- definisjon av hendingar, sannsyn og stokastiske variable

# Til no i emnet

- definisjon av hendingar, sannsyn og stokastiske variable
- definisjon sannsynsfordeling, forventning, varians, osb.

# Til no i emnet

— definisjon av hendingar, sannsyn og stokastiske variable

— definisjon sannsynsfordeling, forventning, varians, osb.

— viktige fordelingar

- $b(x; n, p)$ binomisk
- $p(x; \lambda)$ Poisson
- $n(x; \mu, \sigma)$ normal
- $f(x; \lambda)$ eksponential

# Til no i emnet

- definisjon av hendingar, sannsyn og stokastiske variable
- definisjon sannsynsfordeling, forventning, varians, osb.
- viktige fordelingar
  - $b(x; n, p)$ binomisk
  - $p(x; \lambda)$ Poisson
  - $n(x; \mu, \sigma)$ normal
  - $f(x; \lambda)$ eksponential

Typiske antakingar i spel (kast terning, kortspel, osb.)

# Eksempel (levetider)

- Undersøk levetida $X$ til elektroniske komponentar
- Anta at levetida er eksponentialfordelt ($E(X) = 1/\lambda$, $\lambda$ ukjend)
- Test $n$ komponenter

$$\begin{aligned} X_1 &\sim f(x_1;\lambda) \\ X_2 &\sim f(x_2;\lambda) \\ &\vdots \\ X_n &\sim f(x_n;\lambda) \end{aligned}$$

Viktig: anta $X_1, X_2, \ldots, X_n$ uavhengig

# Innhald resten av emnet

1. Punktestimator: eit "godt" gjett $\hat{\lambda}$ på verdien til den ukjende parameteren $\lambda$
2. Intervallestimat/konfidensintervall: finn intervall $[\hat{\lambda}_{\text{nedre}}, \hat{\lambda}_{\text{øvre}}]$ som dekker den sanne verdien med eit gitt sannsyn[1]

[1] Dette er ei grov forenkling som me vil sjå seinare

# Statstisk inferens I

## Populasjon

Ein populasjon består av alle moglege observasjonar me kan gjere.

## Utval

Eit utval er ei delmengde av ein populasjon.

## Tilfeldig utval

Tilfeldig utval (eng: random sample): einingane som trekkes frå populasjonen velges tilfeldig og uavhengig av kvarandre.

**Mål**: Trekke konklusjonar om eigenskapar til ein populasjon.

$$\begin{tikzcd}
\text{Populasjon} \arrow[r, bend left, "\text{sannsyn}"] & \text{Stokastisk utval} \arrow[l, bend left, "\text{statistisk inferens}"]
\end{tikzcd}$$

## Obervator

Ein **observator** (eng: statistic) er ein funksjon av stokastiske variablar i eit tilfeldig utval.

## Obervator

Ein **observator** (eng: statistic) er ein funksjon av stokastiske variablar i eit tilfeldig utval.

Vil finne ein funksjon $u(X_1, X_2, \ldots, X_n)$ som me kan nytte til å seie noko om $\lambda$.

## Obervator

Ein **observator** (eng: statistic) er ein funksjon av stokastiske variablar i eit tilfeldig utval.

Vil finne ein funksjon $u(X_1, X_2, \ldots, X_n)$ som me kan nytte til å seie noko om $\lambda$.

## Utvalsfordeling

Fordelinga til ein observator kallast utvalsfordelinga (eng: sampling distribution).

## Obervator

Ein **observator** (eng: statistic) er ein funksjon av stokastiske variablar i eit tilfeldig utval.

Vil finne ein funksjon $u(X_1, X_2, \ldots, X_n)$ som me kan nytte til å seie noko om $\lambda$.

## Utvalsfordeling

Fordelinga til ein observator kallast utvalsfordelinga (eng: sampling distribution).

Sidan $u(X_1, X_2, \ldots, X_n)$ er ein funksjon av SV er denne og ein SV og har difor ei (utvals)fordeling.

# Er dataa mine normalfordelt?

# Fysisk kondisjon ved maksimalt oksygenopptak

- Helseundersøkinga i Nord-Trøndelag (HUNT) er ei av verdas største folkehelseundersøkinger. HUNT 3 pågjekk i perioden 2006-2008
- Me skal sjå på data frå 1471 menn frå *Kondisprosjektet i HUNT 3*
- Prosjektet ville kartlege samanhengen mellom kondisjon, fysisk aktivitet, smerte, karfunksjon osb.

Maksimalt oksygenopptak (VO2max) måles ved å springe på ei tredemølle med oksygenmaske til ein ikkje klarer meir.

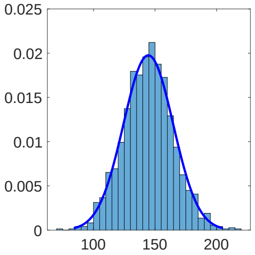


Figure: Histogram

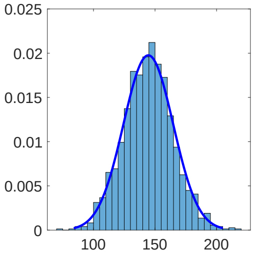


Figure: Histogram

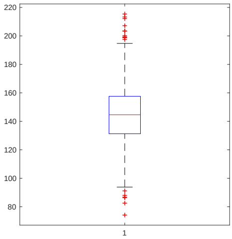


Figure: Boksplott

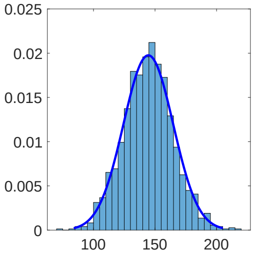


Figure: Histogram

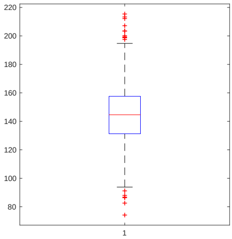


Figure: Boksplott

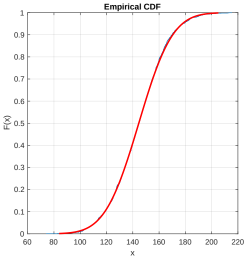


Figure: Kumulativ fordeling

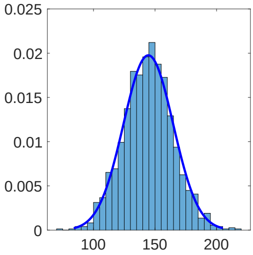


Figure: Histogram

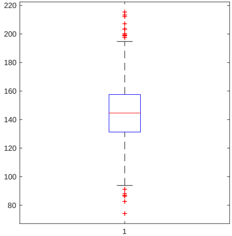


Figure: Boksplott

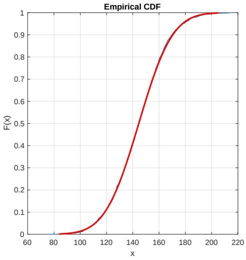


Figure: Kumulativ fordeling

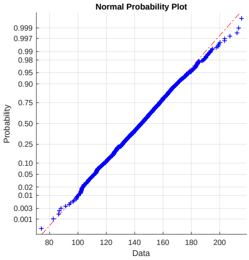


Figure: Normalplott

# Sentralgrenseteoremet

# Sentralgrenseteoremet I

## Teorem

La $X_i$ vere uavhengig identisk fordelte (u.i.f.) stokastiske variablar med $E(X_i) = \mu_X$ og $Var(X_i) = \sigma_X^2 < \infty$ for $i = 1, \ldots, n$. La $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Då vil fordelinga til

$$Z = \frac{\bar{X}_n - E(\bar{X}_n)}{SD(\bar{X}_n)} = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{X}_n - \mu_X}{\sigma_X}\right)$$

gå mot ei standard normalfordeling når $n \to \infty$.

# Eksempel sentralgrenseteoremet

1. Anta $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ er eit tilfeldig utval frå ei kjend fordeling $g(x)$
2. Rekn ut $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
3. Gjenta steg 1 og 2 $k(= 5, 000, 000)$ gonger. Sjå på utvalsfordelinga til $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$

# Sentralgrenseteoremet for normalfordelinga ($n(x; 0, 1)$)

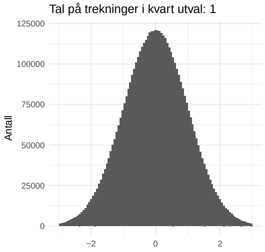


Figure: $n = 1$

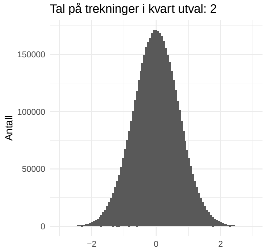


Figure: $n = 2$

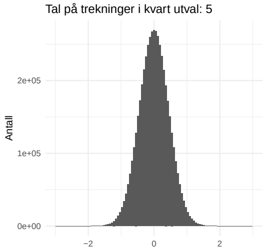


Figure: $n = 5$

# Sentralgrenseteoremet for uniformfordelinga ($f(x; 0, 1)$)

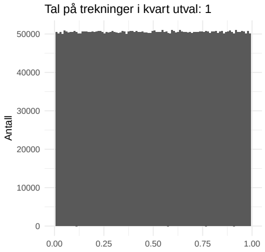


Figure: $n = 1$

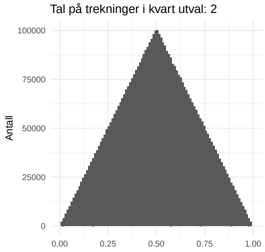


Figure: $n = 2$

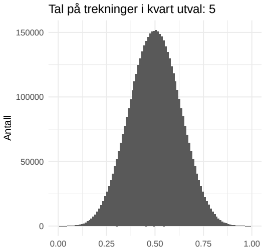


Figure: $n = 5$

# Sentralgrenseteoremet for eksponentialfordelinga ($f(x;1)$)

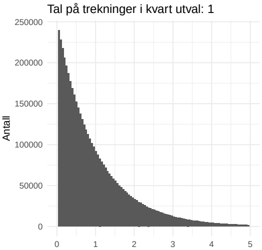


Figure: $n = 1$

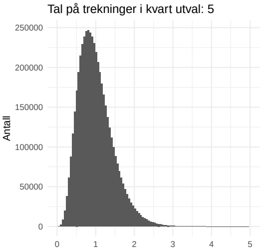


Figure: $n = 5$

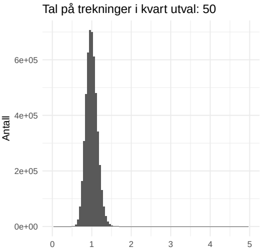


Figure: $n = 50$

# Torsdag

— Estimatorar