

TMA4240 Statistikk H2017 [15]

Del 2: Statistisk inferens

Populasjon og utvalg [8.1]

Observatorer og utvalgsfordelinger [8.2-8.3]

Fordeling til gjennomsnittet og sentralgrenseteoremet [8.4]

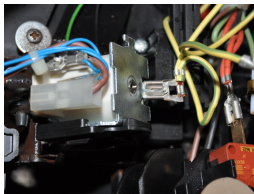
Normalplott [8.8]

Mette Langaas

Institutt for matematiske fag, NTNU

09.10 2017

Utfordring



- ▶ Ved en bedrift produseres en elektrisk komponent.
 - ▶ Komponenten må oppfylle/etterleve visse kvalitetskrav.
-
- ▶ Bedriften har i en kort periode fått endel retur av defekte komponenter, flere enn tidligere,
 - ▶ og er bekymret for at “noe” kan ha skjedd (dårligere kvalitet av råvarer, endringer i produksjonsprosessen).
 - ▶ Bedriften har ikke mulighet til å kvalitetssjekke alle komponenter som produseres (destruktiv testing?).
 - ▶ Hva vil du råde bedriften til å gjøre?

Statistisk inferens

- ▶ Vi ønsker å si noe generelt om en **populasjon** basert på et innsamlet **tilfeldig utvalg** fra populasjonen.
- ▶ Fra innsamling, bearbeiding, analyse og fortolkning av numeriske data og målinger: **trekke slutninger utover det man har observert.**
- ▶ Bakgrunn: vår kunnskap i sannsynlighetsregning.



Fysisk kondisjon målt ved maksimalt oksygenopptak

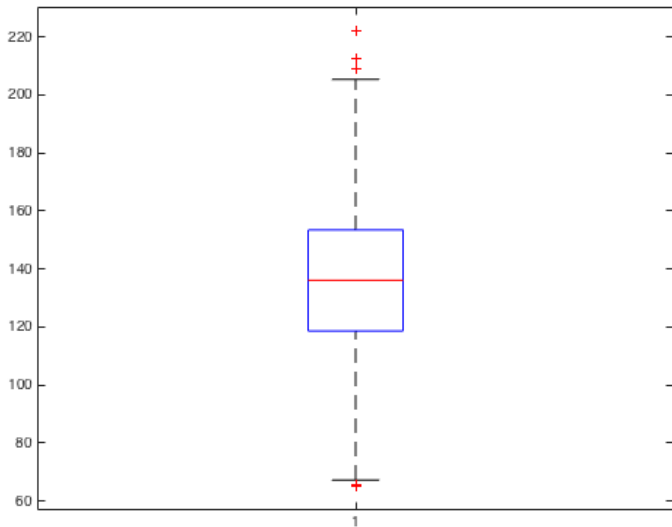


- ▶ Helseundersøkelsen i Nord-Trøndelag (HUNT), er en av verdens største folkehelseundersøkelser. HUNT 3 foregikk 2006-2008.
- ▶ Vi skal se på data fra 1471 menn som er innsamlet i *Kondisprosjektet i HUNT 3*.
- ▶ Prosjektet inngår i en større satsning på å kartlegge sammenhengen mellom kondisjon, fysisk aktivitet, smerte, karfunksjon, livskvalitet og kjente faktorer som påvirker forekomsten av hjerte/kar, og ulike typer kreftsykdommer.

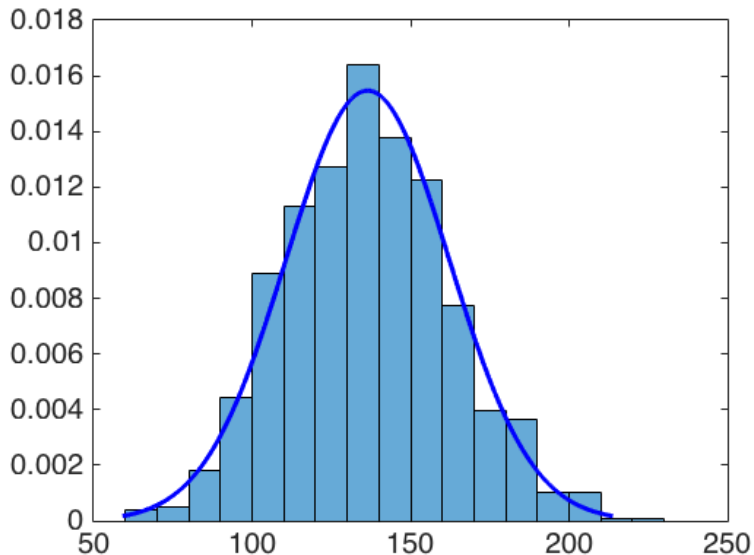
Maksimalt oksygenopptak (VO_{2max}) kan måles ved å løpe på tredemølle med oksygenmaske til man ikke klarer mer. VO_{2max} angis med benevning $ml/kg^{0.75}/minutt$.

Vil du vite mer: se <http://www.ntnu.no/ansatte/anja.bye>

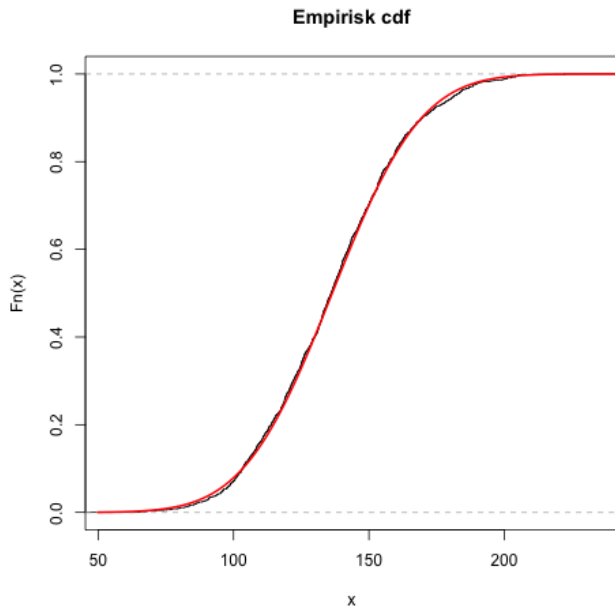
VO2max-dataene: boxplot



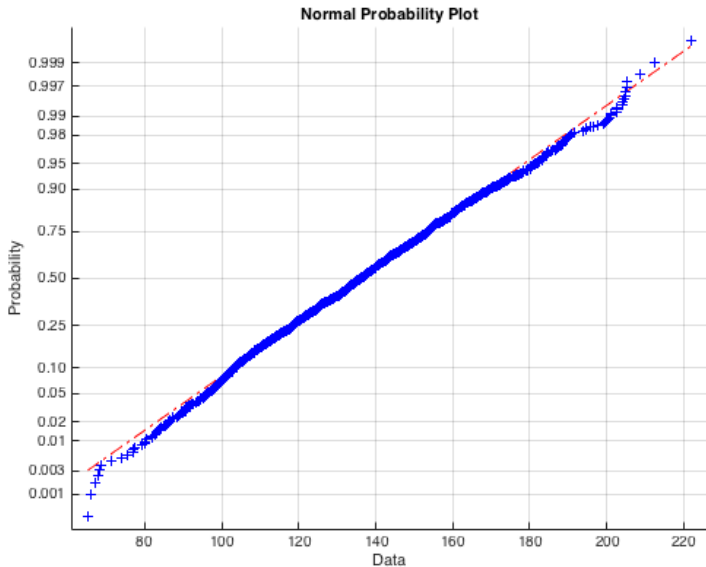
VO2max-dataene: histogram



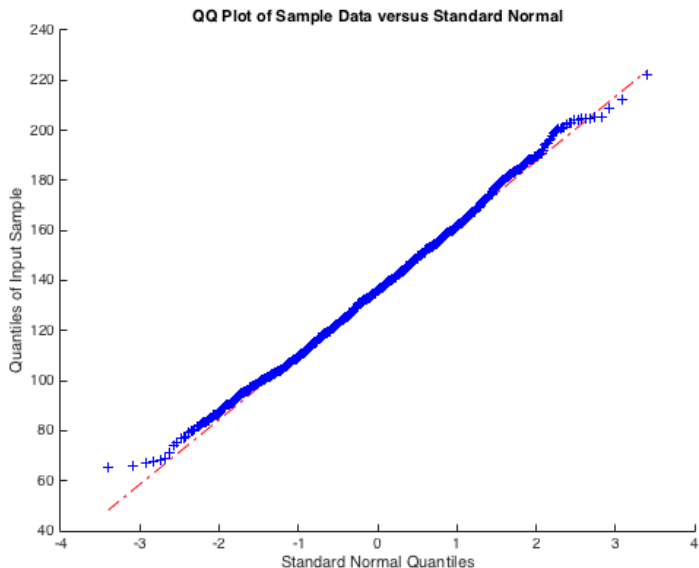
VO2max-dataene: empirisk cdf



VO2max-dataene: Normalplott



VO2max-dataene: Normal qq-plott



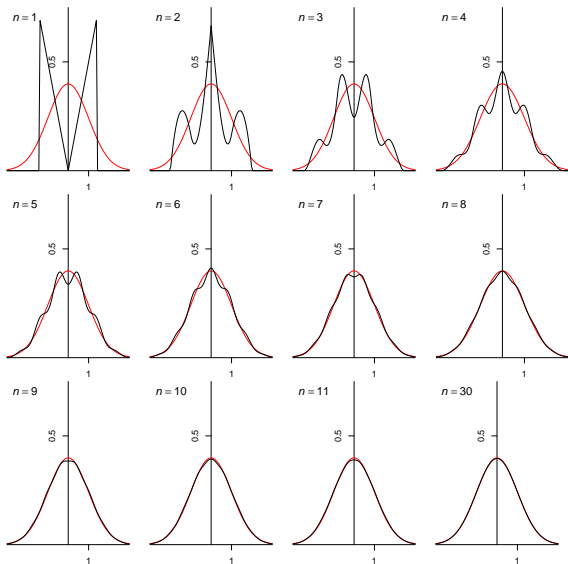
Fordeling til gjennomsnittet \bar{X} [8.4]

TEO 8.2: Sentralgrenseteoremet La X_1, X_2, \dots, X_n være et tilfeldig utvalg fra en fordeling med forventning μ og varians σ^2 . Da har vi at sannsynlighetsfordelingen til

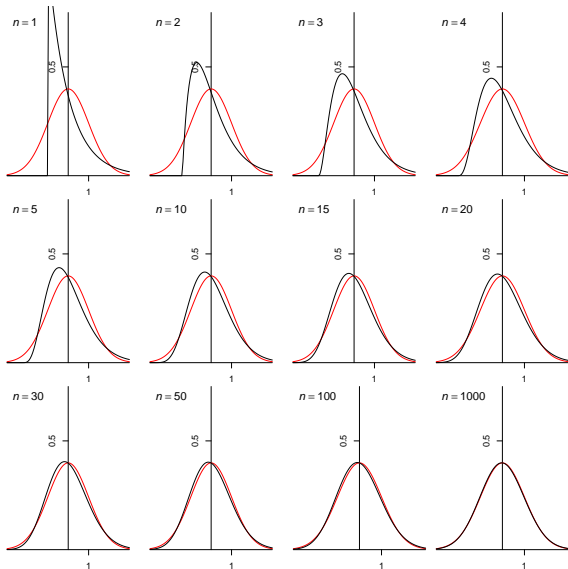
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

går mot standard normalfordelingen, $n(z; 0, 1)$, når $n \rightarrow \infty$.

Sentralgrenseteoremet-for en symmetrisk fordeling



Sentralgrenseteoremet-for eksponensialfordeling



Fra Tabeller og formeler i statistikk (s 38)

Sentralgrenseteoremet

La X_1, X_2, \dots, X_n være en følge av uavhengige identisk fordelte stokastiske variabler med $E(X_i) = \mu$ og $0 < \text{Var}(X_i) = \sigma^2 < \infty$. La $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Da vil for store n

$$\frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

være tilnærmet standard normalfordelt.

Mer presist:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

der Z er standard normalfordelt.

Oppsummering

- ▶ Populasjon vs utvalg
- ▶ Sannsynlighetsregning vs statistisk inferens
- ▶ X_1, X_2, \dots, X_n er et tilfeldig utvalg fra en fordeling med forventning μ og varians σ^2 .
- ▶ Fordeling til gjennomsnittet $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
 - ▶ $E(\bar{X}) = \mu, \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- ▶ Hvis X_i -ene er normalfordelt så er også \bar{X} normalfordelt.
- ▶ Selv om ikke X_i -ene er normalfordelt, så vil allikevel \bar{X} være normalfordelt når n blir stor=*sentralgrenseteoremet*.