TMA4240 Statistikk H2016 [16]

Bolk 2: Statistisk inferens
Fordeling til gjennomsnittet og sentralgrenseteoremet [8.4]
Fordeling til utvalgsvarians [8.5] og khikvadratfordelingen [6.7]
Estimering [9.1-9.3]

Mette Langaas

Institutt for matematiske fag, NTNU

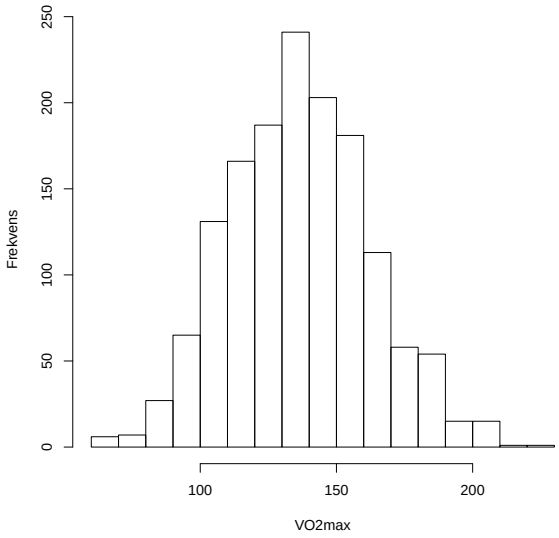
wiki.math.ntnu.no/emner/tma4240/2016h/start/

Parkslirenke: spiringsgrad

Bilde tilgjengelig fra www.anpc.ab.ca/wiki/index.php?title=Fallopia_japonica.

- I artikkelen *Seed Germinability and Its Sesonal Onset of Japanese Knotweed* (Bram og McNair, 2004, Weed Science Society of America) ser man blant annet på springsgrad av frø.
- I en populasjon av denne typen frø - hva er spiringsgraden p?
- I ett av eksperimentene i artikkelen plantet man 1200 frø og observerte at 987 av dem spiret i løpet av en tidsperiode på ca tre måneder.
- Kan vi bruke dette eksperimentet til å si noe om spiringsgraden p?

Fysisk kondisjon målt ved maksimalt oksygenopptak

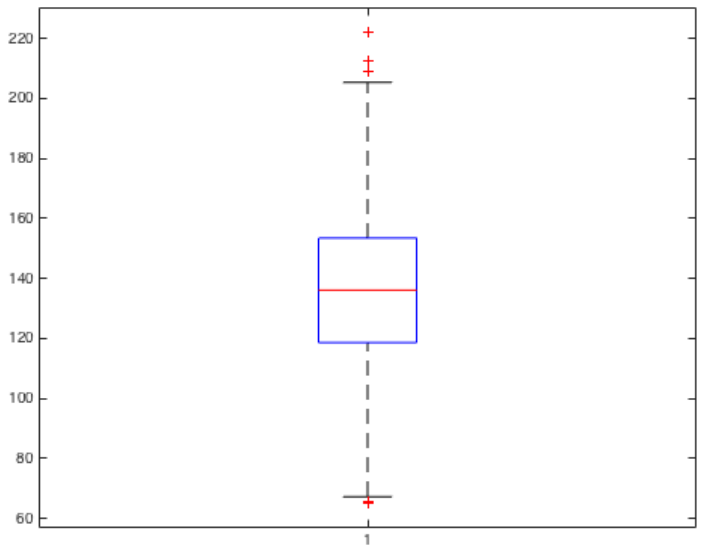


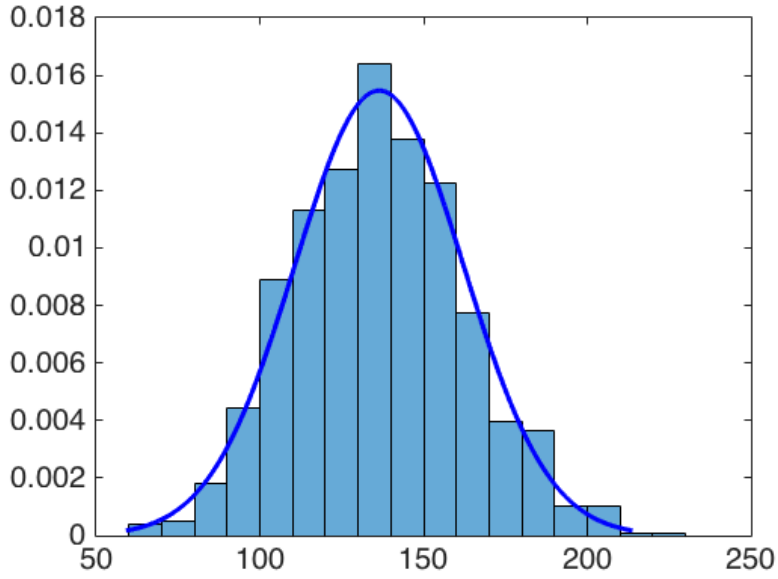
- Maksimalt oksygenopptak (VO2max) kan måles ved å løpe på tredemølle med oksygenmaske til man ikke klarer mer.
- VO2max angis med benevning $\text{ml/kg}^{0.75}\text{/minutt}$.
- Vi studerer menn i Nord-Trøndelag, og lurer på hva forventet VO2max for disse er.
- VO2max kan antas normalfordelt i denne populasjonen, men med ukjent forventning og varians.
- Vi har data på VO2max for 1471 menn i Kondisprosjektet i HUNT 3.
- Kan vi bruke disse dataene til å si noe om forventning og varians for VO2max i Nord-Trøndelag?

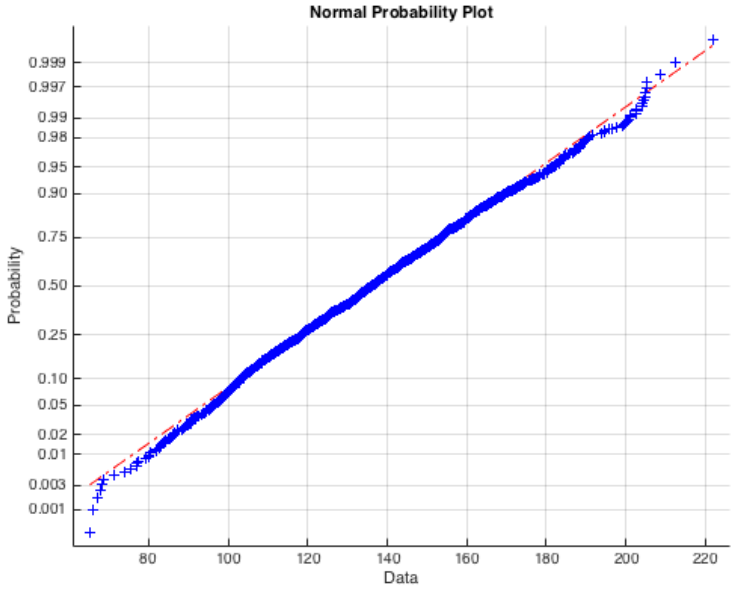
Statistisk inferens

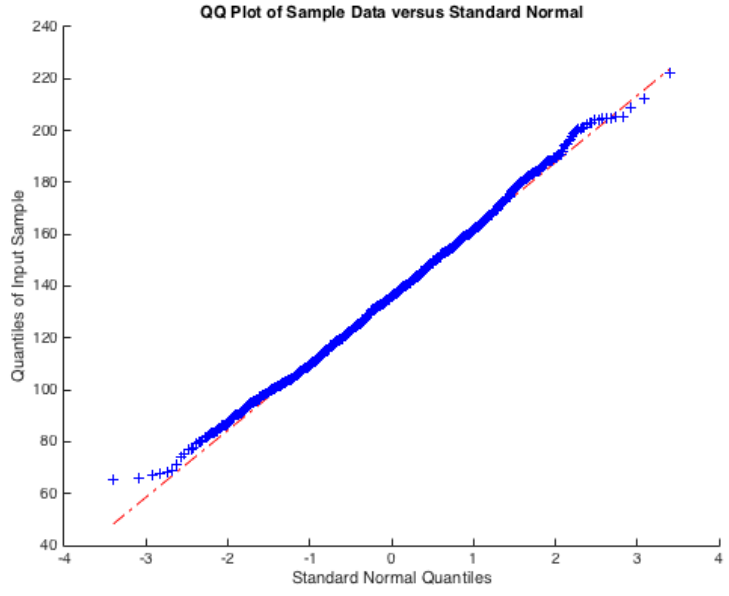
- Vi ønsker å si noe generelt om en **populasjon** basert på et innsamlet **tilfeldig utvalg** fra populasjonen.
- Fra innsamling, bearbeiding, analyse og fortolkning av numeriske data og målinger: **trekke slutninger utover det man har observert.**
- Bakgrunn: vår kunnskap i sannsynlighetsregning.

VO2max-dataene: normalfordelt? [8.8]









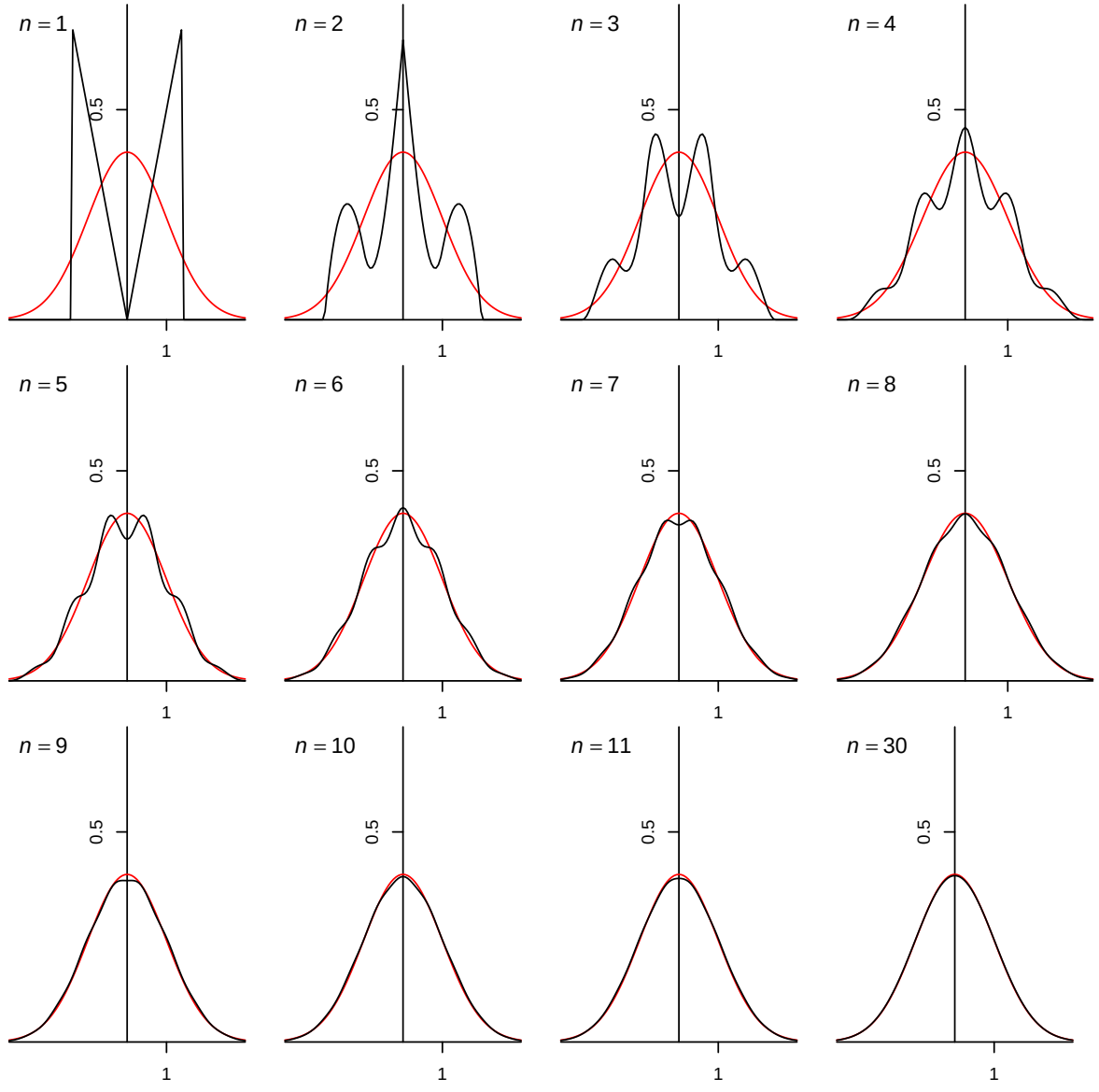
Fordeling til gjennomsnittet $\bar{X}$ [8.4]

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

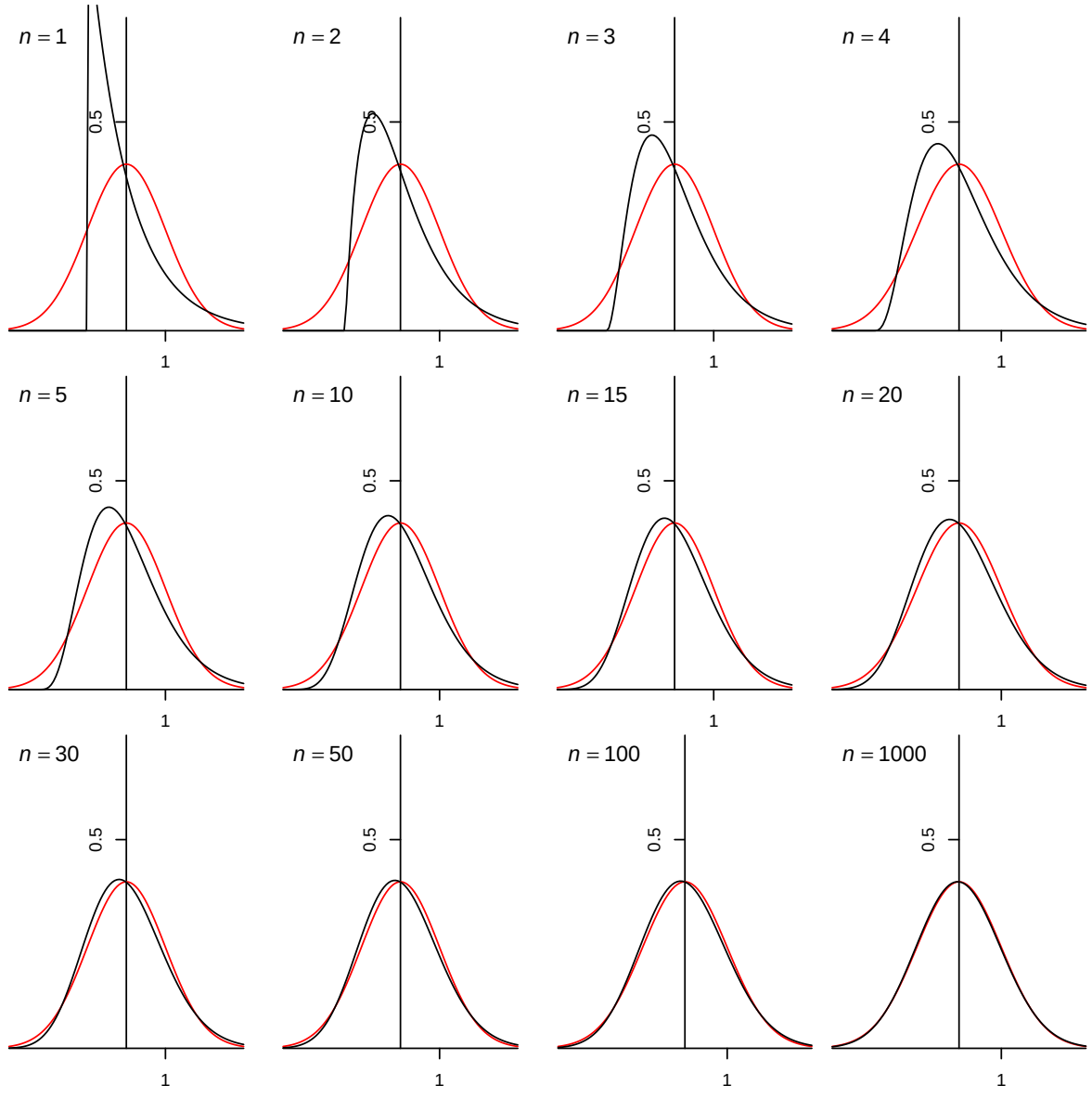
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

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

Sentralgrenseteoremet-for en symmetrisk fordeling



Sentralgrenseteoremet-for eksponensialfordeling



Fra Tabeller og formeler i statistikk (s 38)

Sentralgrenseteoremet

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

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

være tilnærmet standard normalfordelt.
Mer presist:

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

der Z er standard normalfordelt.

Forventingsrett estimator

DEF 9.1: En observator $\widehat{\theta}$ er en **forventningsrett** estimator for parameteren θ hvis

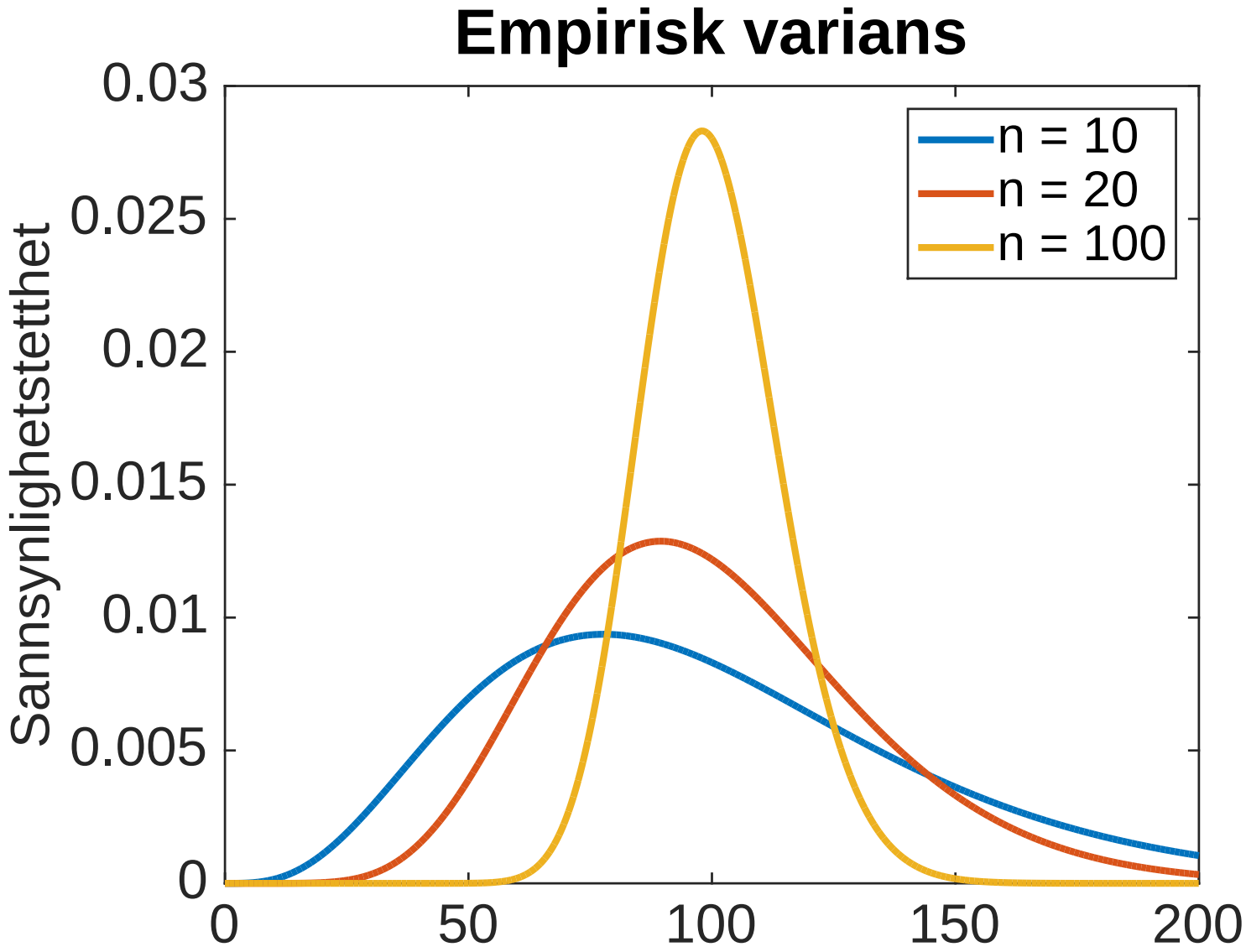
$$\mathrm{E}(\widehat{\theta}) = \theta.$$

$\mathrm{E}(S^2)$ (se også læreboka)

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 = \sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + n(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 - 2(\bar{X}-\mu)n(\bar{X}-\mu) + n(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(S^2) &= \mathrm{E}[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2] = \frac{1}{n-1}\{\mathrm{E}[\sum_{i=1}^{n}(X_i-\mu)^2] - E[n(\bar{X}-\mu)^2]\} \\
&= \frac{1}{n-1}[\sum_{i=1}^{n}\mathrm{Var}(X_i) - n\mathrm{Var}(\bar{X})] = \frac{1}{n-1}[\sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n}] \\
&= \frac{1}{n-1}(n\sigma^2-\sigma^2) = \sigma^2
\end{aligned}
$$

Fordelingen til S^2 for $\sigma^2 = 100$ for ulike n



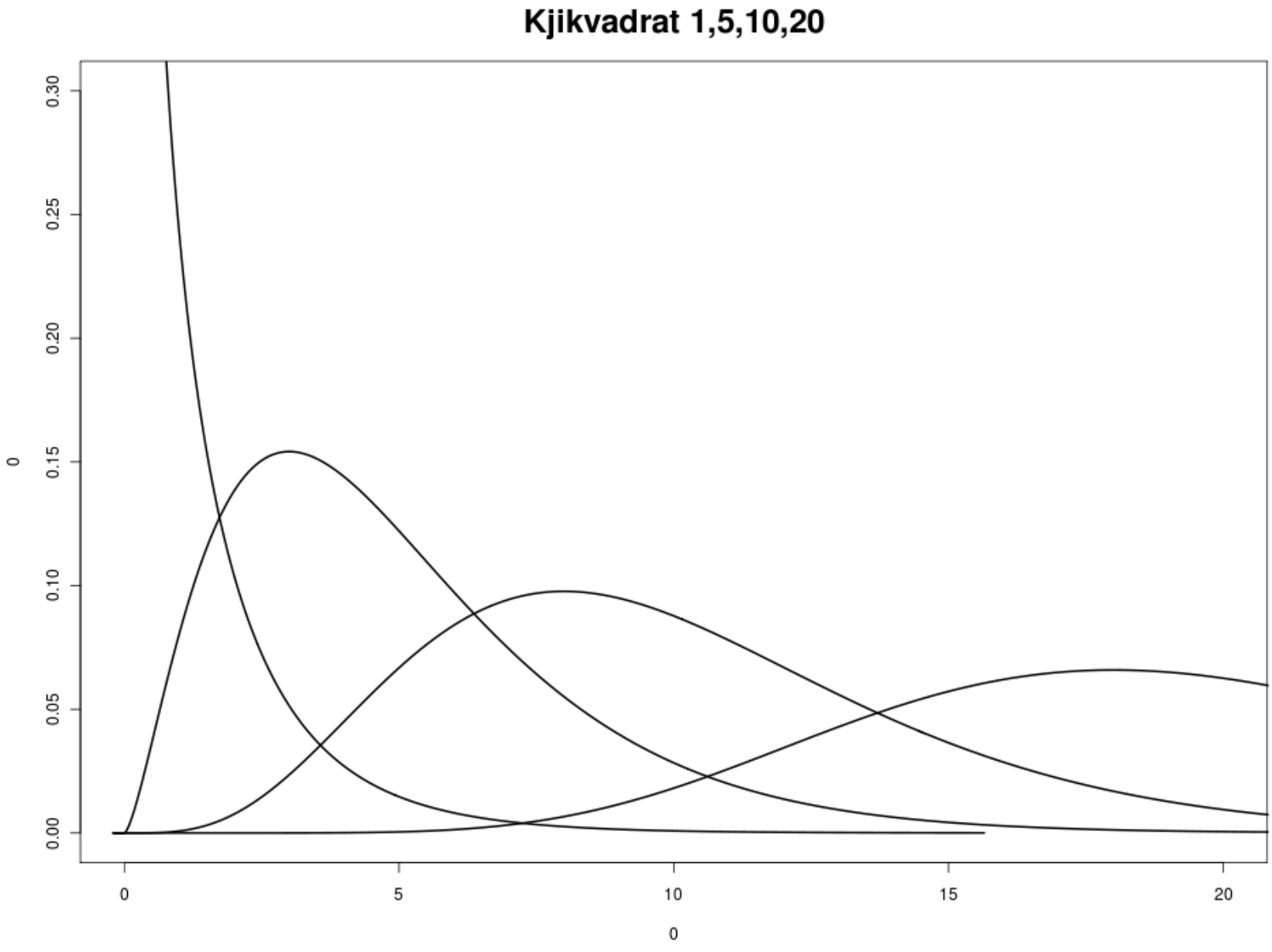
Khikvadrat fordelingen [6.8]

En kontinuerlig stokastisk variabel X er khikvadrat fordelt med parameter ν (kalt frihetgrader), hvis sannsynlighetstettheten er gitt ved

$$f(x;\nu) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, & x > 0 \\ 0 & \text{ellers.} \end{cases}$$

hvor ν er et positivt heltall.

Khikvadrat fordelingen [6.8]



Z er $N(0,1)$ og Z^2 er khikvadrat med parameter 1

- Hvis Z har fordeling $n(x;0,1)$, så vil
- Z^2 ha en (for oss ny) fordeling som heter khikvadrat med parameter 1.

Bevis: kapittel 7, funksjoner av stokastiske variabler.

Khikvadrat fordelingen med $\nu = 1$

$$f(x; \nu = 1) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}, & x > 0 \\ 0 & \text{ellers.} \end{cases}$$

Kjikvadrat og khikvadrat brukes på norsk. Chi-squared på engelsk. Noteres ofte χ^2_1.

Fra Tabeller og formeler i statistikk (s 34)

Fordeling til lineærkombinasjoner

La $X_1, \ldots, X_n$ være uavhengige variabler.

Dersom X_i er normalfordelt med forventning μ_i og varians σ_i^2 vil $Y = \sum_{i=1}^n a_i X_i$ være normalfordelt med forventning $\sum_{i=1}^n a_i \mu_i$ og varians $\sum_{i=1}^n a_i^2 \sigma_i^2$.

Dersom X_i er binomisk fordelt med parametre m_i og p vil $Y = \sum_{i=1}^n X_i$ være binomisk fordelt med parametre $\sum_{i=1}^n m_i$ og p.

Dersom X_i er Poissonfordelt med parameter μ_i vil $Y = \sum_{i=1}^n X_i$ være Poissonfordelt med parameter $\sum_{i=1}^n \mu_i$.

Dersom X_i er χ^2-fordelt med ν_i frihetsgrader vil $Y = \sum_{i=1}^n X_i$ være χ^2-fordelt med $\sum_{i=1}^n \nu_i$ frihetsgrader.

Fordelingen til S^2 [8.5]

- Resultat: $V = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - \bar{Z}^2$ er kjikvadrat-fordelt med $\nu = n - 1$ frihetsgrader. Fordi:

  i) $X_1, ..., X_n$ u.i.f. normal, $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$.

  ii) $Z_i = \frac{X_i - \mu}{\sigma}$ er standard normalfordelt, og $\bar{Z} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ er standard normalfordelt.

  iii) $Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2$ er kjikvadrat-fordelt med 1 frihetsgrad.

  $\bar{Z}^2 = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ er kjikvadrat-fordelt med 1 frihetsgrad.

  iv) $\sum_{i=1}^{n} Z_i^2$ er kjikvadratfordelt med n frihetsgrader.

  v) $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$, og dermed $V = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - \bar{Z}^2$

  vi) S^2 og $\bar{Z}$ er uavhengige.

χ^2-fordeling (kjikvadratfordeling)

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \ , \quad x \geq 0 \ , \quad \nu = 1, 2, \ldots .$$

$$\mathrm{E}(X) = \nu \ , \quad \mathrm{Var}(X) = 2\nu \ , \quad M_X(t) = \left(\frac{1}{1-2t} \right)^{\nu/2} \text{ for } t < \frac{1}{2}.$$

Kommentar: Dersom $X_1, \ldots, X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2 har vi at

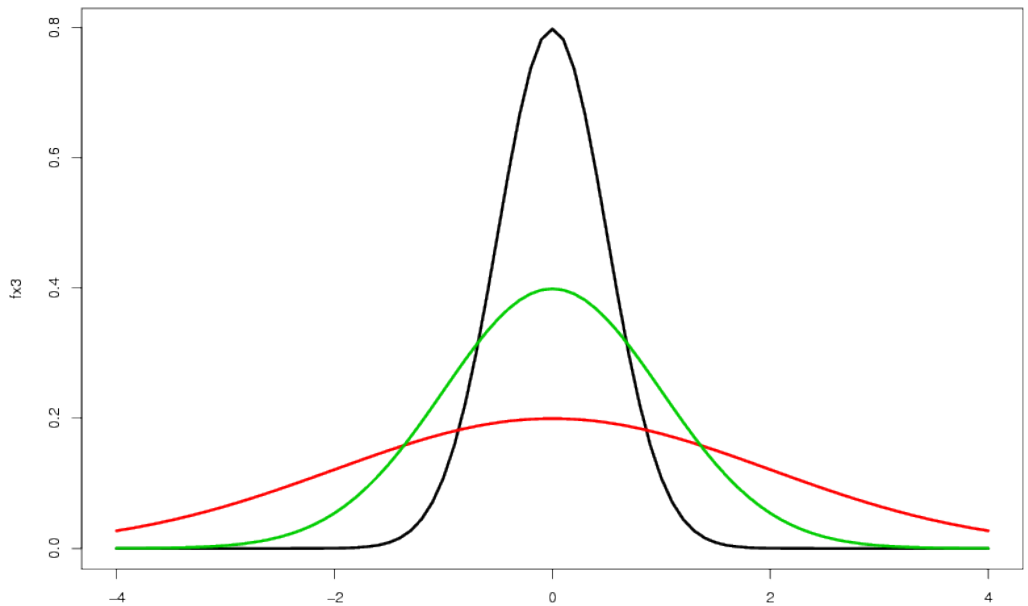
$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n-1 \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ og } \bar{X} \text{ er uavhengige.}$$

Mest effektive estimator

DEF 9.2: Hvis vi ser på alle mulige forventningsrette estimatorene for en parameter θ, kaller vi den med minst varians for den **mest effektive estimatoren** til θ.



Eksamen, juni 2004, 1c

Et av spørsmålene som regnes å være av delvis sensitiv natur er "hvor gammel er du?". En gruppe på n personer ble stilt dette spørsmålet, deretter ble svarene registrert og sammenlignet med informasjon i offentlige registre. La X være en stokastisk variabel som angir antall personer som lyver blant n personer, og la p være sannsynligheten for at en person lyver.
Under hvilke antagelser vil X være binomisk fordelt?
Vi antar nå at p er ukjent. For å estimere p er det foreslått to estimatorer,

$$\hat{p} = \frac{X}{n} \quad \text{og} \quad p^* = \frac{X}{n-1}.$$

Finn forventningsverdi og varians til hver av estimatorene $\hat{p}$ og p^*.
Hvilke to egenskaper kjennetegner en god estimator?
Hvilken av estimatorene $\hat{p}$ og p^* vil du foretrekke? Begrunn svaret.

Estimering

- Mål: finne "sannheten" om et fenomen i en populasjon.
- "Sannheten" knytter vi til en ukjent parameter, θ, i en valgt fordeling.
- Vi trekker et tilfeldig utvalg fra populasjonen; $X_1, X_2, ..., X_n$ (u.i.f.).
- En estimator gir et anslag for den ukjente parameteren og er en funksjon av stokastiske variabler, $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$.
- Hvilke egenskaper bør en god estimator ha?
  - Estimatoren bør være forventningsrett, dvs. $\mathrm{E}(\hat{\theta}) = \theta$.
  - Estimatoren bør ha minst mulig varians, $\mathrm{Var}(\hat{\theta})$, og variansen bør avta når antall observasjoner, n, øker.
- Hvordan kan vi finne estimatorer?
  - ved intuisjon,
  - ved matematisk metode.