STK1110 Høsten 2026

Konfidensintervaller

Tilsvarer Kapittel 8
samt 6.3 og 6.4

Denne og forrige uker

Uke 35: Avsnitt 7.4

- Egenskaper MLE i store utvalg
- Variansestimering for MLE i store utvalg Fisher informasjon
- Cramér-Rao-nedre grense for forventningsrette estimatorer
- Tilnærmet effisiens for MLE

Denne og forrige uke: Kapittel 8 - Konfidensintervaller

- Repetisjon STK1100: Avsnitt 8.1 (Uke 36)
- Avsnitt 6.3 og 6.4 om fordelinger relatert til normalfordelingen (Uke 36)
- Avsnitt 8.2: konfidensintervall for forventning med normalfordelte data med ukjent varians (Uke 36, litt igjen i dag)
- Avsnitt 8.4: konfidensintervall for standardavvik med normalfordelte data (Uke 37, i dag)
- Avsnitt 8.5: bootstrap av konfidensintervall (Uke 37, i dag)
- Avsnitt 8.3: konfidensintervall for populasjonandeler (Uke 37, i dag)

Intervallestimering

Vi har funnet metoder for å *punktestimere* parametre θ i statistiske modeller:

- Intuitive, f.eks. $\hat{p} = \frac{Y}{n}$ for $Y \sim \text{Bin}(n, p)$.
- Momentestimatorer
- Maximum Likelihood estimatorer

Vi har også funnet uttrykk for varians/standardavik for (noen av) estimatorene og kan da anslå estimatorene standardfeil.

Mer generelt ønsker vi å estimere intervall der vi med rimelighet vil si at den faktiske parameterverdien θ befinner seg.

Dette krever formalisering av hva som presist menes og vi skal i dag se på det vanligste intervallet, nemlig konfidensintervall.

Det finnes alternative måter å formalisere usikkerhetsintervall på og i forbindelse med Kap. 15 skal vi kort komme inn på Bayesianske *kredibilitetsintervall*

Kvantifisering av usikkerhet

- Standardfeil til estimator $\hat{\theta}$
  - Gir uttryk for variasjonen vi venter for $\hat{\theta}$ over gjentatte eksperimenter

- Konfidensintervall
  - Et intervall som ved gjentatte eksperimenter vil dekke den sanne parameter med en gitt, vanligvis høy, sannsynlighet

Eks 8.1 DBC: Tensile (strekk) styrke i titanlegeringer

- n=31
- $\bar{x} = 1064$
- $\sigma = 55$ (antatt kjent)
- standardfeil til $\bar{x}$ er $\sigma/\sqrt{n} = 9.88 \approx 10$

Eksempel: Normalfordelte data med kjent varians

- Anta vi har observasjoner $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu, \sigma)$.
- Vi ønsker å konstruere et konfidensintervall for μ, som er ukjent, og vi antar først at σ^2 er kjent. (i praksis urealistisk)
- Vi vet at $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, slik at $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- For å lage et 95% konfidensintervall for μ, bruker vi at

$$\mathrm{P}\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

- Ved omforming av ulikhetene får vi da

$$\mathrm{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Konfidensintervall for μ

- Da inneholder det stokastiske intervallet

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right),$$

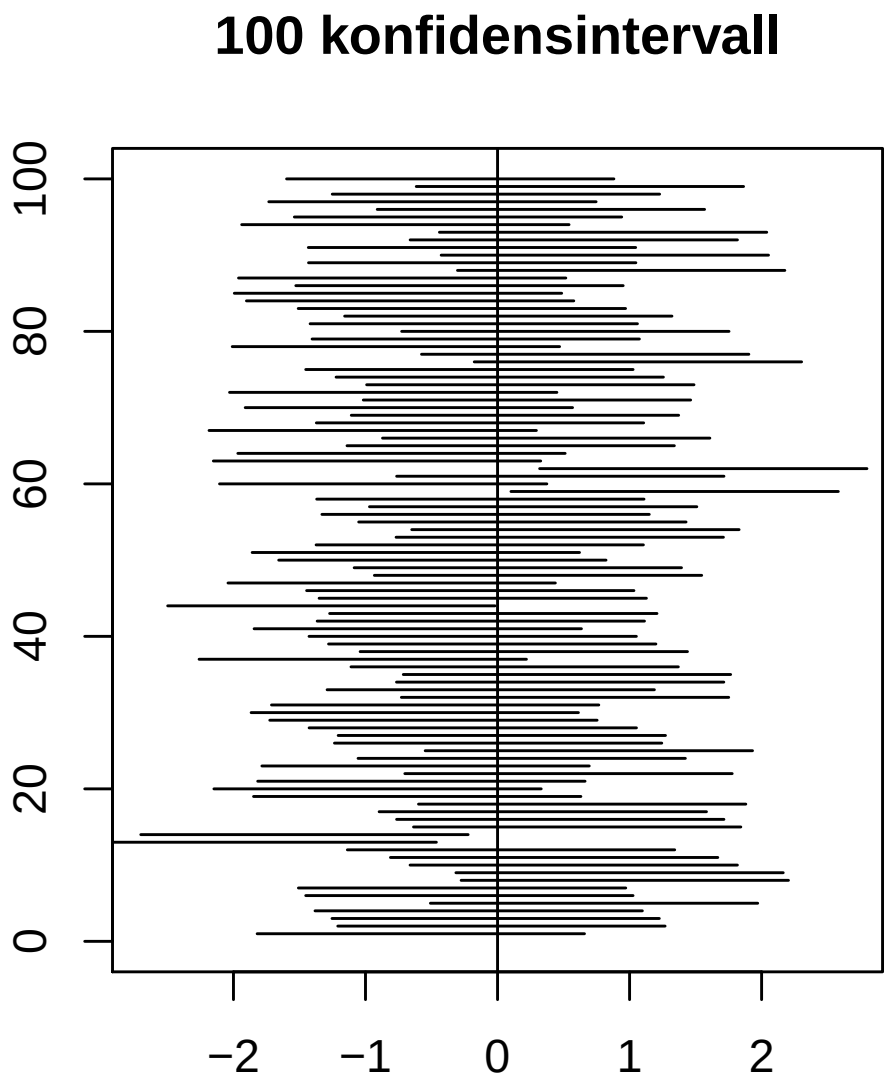
  μ med 95% sannsynlighet.
- Når vi setter inn de observerte verdiene $x_1, \ldots, x_n$, får vi et 95% konfidensintervall for μ, gitt ved:

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right),$$

- Vi kan ikke si at μ ligger innenfor intervallet med 95% sannsynlighet, da μ ikke er en stokastisk variabel!
- I stedet må vi tolke det slik: dersom vi gjentar øvelsen over for mange tilsvarende observerte datasett av samme størrelse, vil rundt 95% av de resulterende konfidensintervallene inneholde den sanne verdien av μ.

Simulerte konfidensintervall for μ

Simulerer $X_i \sim \mathrm{N}(\mu, \sigma), i = 1, \ldots, n = 10$ med $\mu = 0$ og $\sigma = 2$.
Gjentar simuleringen 100 ganger og plotter konfidensintervallene.



Detaljert utledning av konfidensintervall $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ for μ

når $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu, \sigma)$. Siden $Z = \frac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0,1)$ har vi

$$\mathrm{P}\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

Men dette er ekvivalent med at

$$\mathrm{P}\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

som igjen er ekvivalent med

$$\mathrm{P}\left(-\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

og etter at vi multiplisere ulikhetene med -1 ender vi opp med

$$\mathrm{P}\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

hvilket vi ønsket å komme fram til.

Konfidensintervall for μ (forts.)

- Konfidensintervaller med andre konfidensnivåer, som 90%, 99%, eller mer generelt $100 \cdot (1 - \alpha)\%$, får vi ved å bytte ut $1.96 = z_{0.05/2}$ med $z_{\alpha/2}$, der z_α er øvre $\alpha/2$-kvantilen i standard normalfordeling:

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

Bestemmelse av utvalgsstørrelse

- En viktig problemstilling i statistikk er planlegging av forsøk
- Spesielt er det ofte kostbart å samle inn mange observasjoner.
- Det er derfor fornuftig å planlegge forsøket slik at estimeringen er presis nok
- og samtidig ikke for dyr og tidkrevende
- Ønsker derfor å bestemme hvor mange observasjoner n som er nødvendig for at et konfidensintervall blir smalt nok.
- Lengden av et konfidensintervall: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ gis ved

$$w = 2z_{\alpha/2}\sigma/\sqrt{n}$$

- Gitt σ og α får vi da at vi trenger

$$n = \left(\frac{z_{\alpha/2}\sigma}{w/2}\right)^2$$

for å oppnå et konfidensintervall med ønsket lengde w.

Generell framgangsmåte for å konstruere konfidensintervall

- Anta at vi har stokastiske variable $X_1, \ldots, X_n$, og at vi vil finne et $100 \cdot (1-\alpha)\%$ konfidensintervall for en parameter θ.
- Anta videre at vi har en størrelse $h(X_1, \ldots, X_n; \theta)$, som avhenger av θ, men hvis fordeling ikke avhenger av θ.
  - For normalfordeling, $Z = h(X_1, ..., X_n, \mu) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- La a og b være henholdvis øvre $(1-\alpha/2)$- og $\alpha/2$-kvantilen i fordelingen til $h(X_1, \ldots, X_n; \theta)$.
- Da har vi:

$$P(a \leq h(X_1, \ldots, X_n; \theta) \leq b) = 1 - \alpha.$$

Generell framgangsmåte for å konstruere konfidensintervall (forts.)

Basert på

$$\mathrm{P}(a \leq h(X_1, \ldots, X_n; \theta) \leq b) = 1 - \alpha.$$

- så omformer vi ulikhetene:

  $$a \leq h(X_1, \ldots, X_n; \theta) \leq b \iff L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n).$$

- Da er $\mathrm{P}(L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)) = 1 - \alpha.$

- For å få et $100 \cdot (1 - \alpha)\%$ konfidensintervall for θ setter vi inn de observerte verdiene og får:

  $$(L(x_1, \ldots, x_n), U(x_1, \ldots, x_n)).$$

Eksempel

$X_1, \ldots, X_n \overset{uif}{\sim} N(\mu, \sigma^2)$, der μ er kjent. Vi ønsker å lage et konfidensintervall for σ^2. Kommer tilbake til eksempelet.

Stort utvalg uif data med ukjent varians

- Vi antar at vi har observasjoner $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n$, som er uif med forventning μ og varians σ^2, som begge er ukjent.
- Men X_i-ene er ikke antatt å være normalfordelte!
- I henhold til sentralgrenseteoremet er likevel $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{tiln.}{\sim} N(0,1)$ når n er tilstrekkelig stor.
- Dersom σ^2 er kjent, kan vi lage et *tilnærmet* konfidensintervall for μ på samme måte som før, dvs. $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- Ellers kan vi bruke at (det kan vises) $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{tiln.}{\sim} N(0,1)$, med $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, når n er stor nok.

Stort utvalg uif data med ukjent varians (forts.)

- Vi har da:

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

- Ved omforming av ulikhetene får vi

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

- Det betyr at et *tilnærmet* $100 \cdot (1-\alpha)\%$ konfidensintervall for μ er gitt ved:

$$\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

Tilnærmet konfidensintervall for MLE

For $X_i \sim f(x;\theta), i = 1, \ldots, n$ uavhengige fant vi at MLE

$$\hat{\theta} \sim \mathrm{N}(\theta, 1/\sqrt{nI(\theta)}) \quad \text{tilnærmet}$$

der $I(\theta) = -\mathrm{E}(\frac{\partial^2 \ln(f(X_i;\theta))}{\partial\theta^2}) =$ informasjonen i en X_i.

Vi får dermed også, med standardfeil $se = 1/\sqrt{nI(\hat{\theta})}$ for $\hat{\theta}$, at

$$\hat{\theta} \pm 1.96se$$

er et tilnærmet 95% konfidensintervall for θ når n er tilstrekkelig stor.

Eksempel tilnærmet KI for MLE

La $X_i, i = 1, \ldots, n$ være uavhengige og binære med $p = \mathrm{P}(X_i = 1)$.

Da er

- $Y = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$
- $\hat{p} = \frac{Y}{n}$ MLE for p er tilnærmet $\mathrm{N}(p, se)$
- der standardfeilen $se = \sqrt{\hat{p}(1-\hat{p})/n} = 1/\sqrt{nI(\hat{p})}$
- og $1/I(p) = p(1-p)$ er invers informasjon for en X_i.

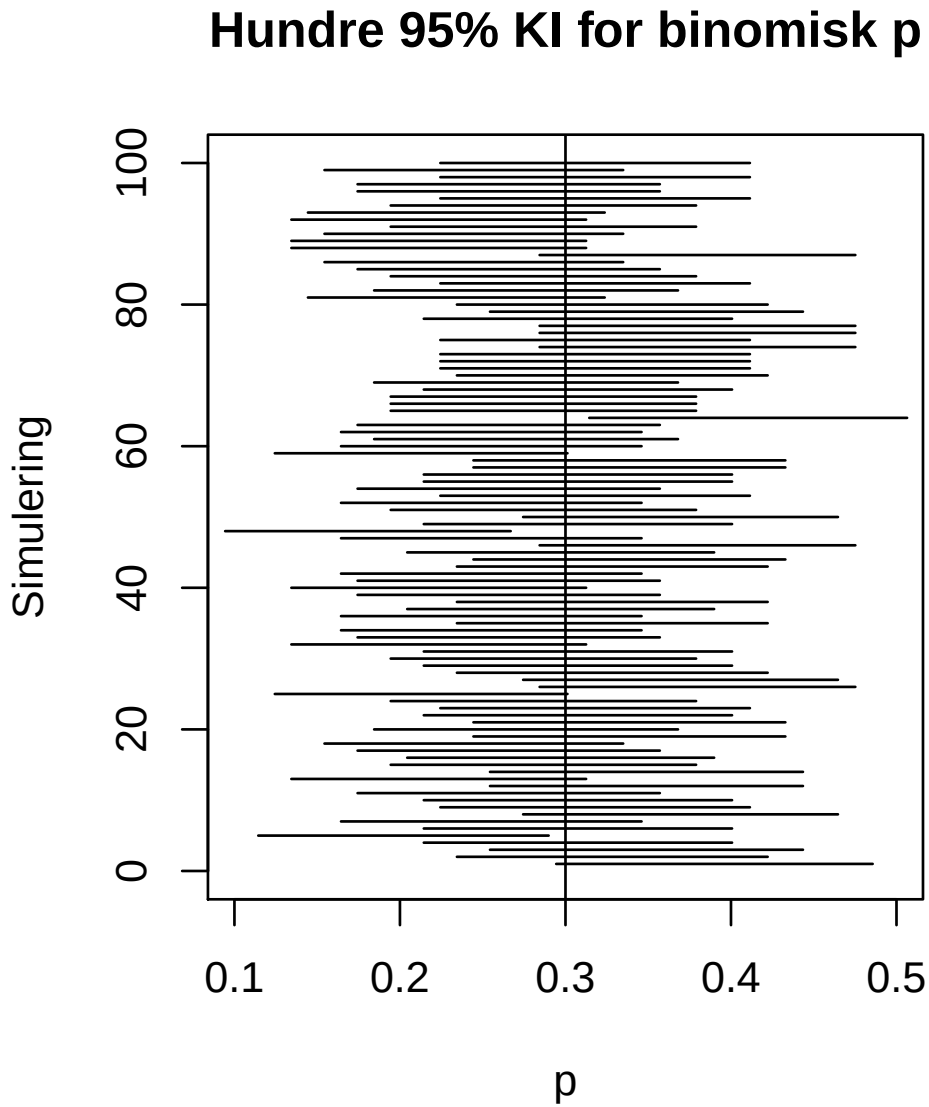
Et tilnærmet 95% konfidensintervall for p gis ved

$$\hat{p} \pm 1.96se$$

Vi skal senere (Avsn. 8.3) se på et noe bedre konfidensintervall for binomisk p.

Simulering binomiske konfidensintervall

Simulerer $Y \sim \text{Bin}(100, 0.3)$ og beregner tilnærmet 95% konfidensintervall totalt 100 ganger som illustrert i flg. plott:



Konfidensintervall for μ i små utvalg uif data og ukjent σ

- Vi antar at vi har observasjoner $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n$, som er uif med forventning μ og varians σ^2, som begge er ukjent.

- Vil se vi her på det tilfellet at n er for lav til at sentralgrenseteoremet slår inn.

- For å komme noen vei antar vi i tillegg at $X_i \sim N(\mu, \sigma)$, $i = 1, \ldots, n$.

- Vi ønsker å konstruere et konfidensintervall for μ, og da trenger vi estimatorer for både μ og σ^2, og vi kan bruke $\bar{X}$ og S^2.

Konfidensintervall for μ

- Vi har observatoren
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

- Når n er stor, er T tilnærmet standard normalfordelt. Det gjelder imidlertid ikke for mindre n.
- Hva kan vi si om T i slike tilfeller?
- Trenger en liten omvei om fordelinger fra kap 6.4 og 6.3

- Merk først at $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ involverer ledd X_i^2 når kvadratene ekspanderes.
- Ledd av type X_i^2 og $Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2$ vil være viktige framover.

Plan for utledning av fordeling for $T = (\bar{X} - \mu)/(S/\sqrt{n})$

når $X_i \sim \mathrm{N}(\mu, \sigma), i = 1, \ldots, n$ og uif.

- Definere χ^2-fordelingene og påvise at de er gammafordelinger
- Vise at $\bar{X}$ and S^2 er uavhengige
- Påvise $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$, med $n-1$ frihetsgrader
  Også sentralt for konfidensintervall for σ (Avsn. 8.4)
- Definere t-fordelingene: $T = \frac{Z}{\sqrt{U/\nu}}$ er t-fordelt med ν frihetsgrader når $Z \sim \mathrm{N}(0,1)$ og $U \sim \chi^2_\nu$ er uavhengige
- Vise at

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

  er t-fordelt med $n-1$ frihetsgrader

Avsnitt 6.3 og 6.4 nevner også F-fordelinger, men disse trenger vi ikke før Kap. 10-12.

Minner først om Gamma-fordelingen

- X er Gamma fordelt hvis

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

- Egenskaper:

$$E[X] = \mu = \alpha\beta$$
$$V[X] = \sigma^2 = \alpha\beta^2$$
$$M_X(t) = \frac{1}{(1-\beta t)^{\alpha}}$$

- Hvis X_1 og X_2 uif $Gamma(\alpha,\beta)$, så er $X = X_1 + X_2$ Gamma$(2\alpha,\beta)$.

χ^2-fordelingen (kji-kvadrat fordelingen)

Anta $Z_1, ..., Z_\nu \overset{uif}{\sim} N(0,1)$. Da er (per def.)

$$X = Z_1^2 + \cdots + Z_\nu^2$$

χ^2-fordelt med ν frihetsgrader (skriver $X \sim \chi^2_\nu$).

Vi skal på de neste slidene vise at χ^2_ν fordelingen er en gammafordeling med $\alpha = \nu/2$ og $\beta = 2$.

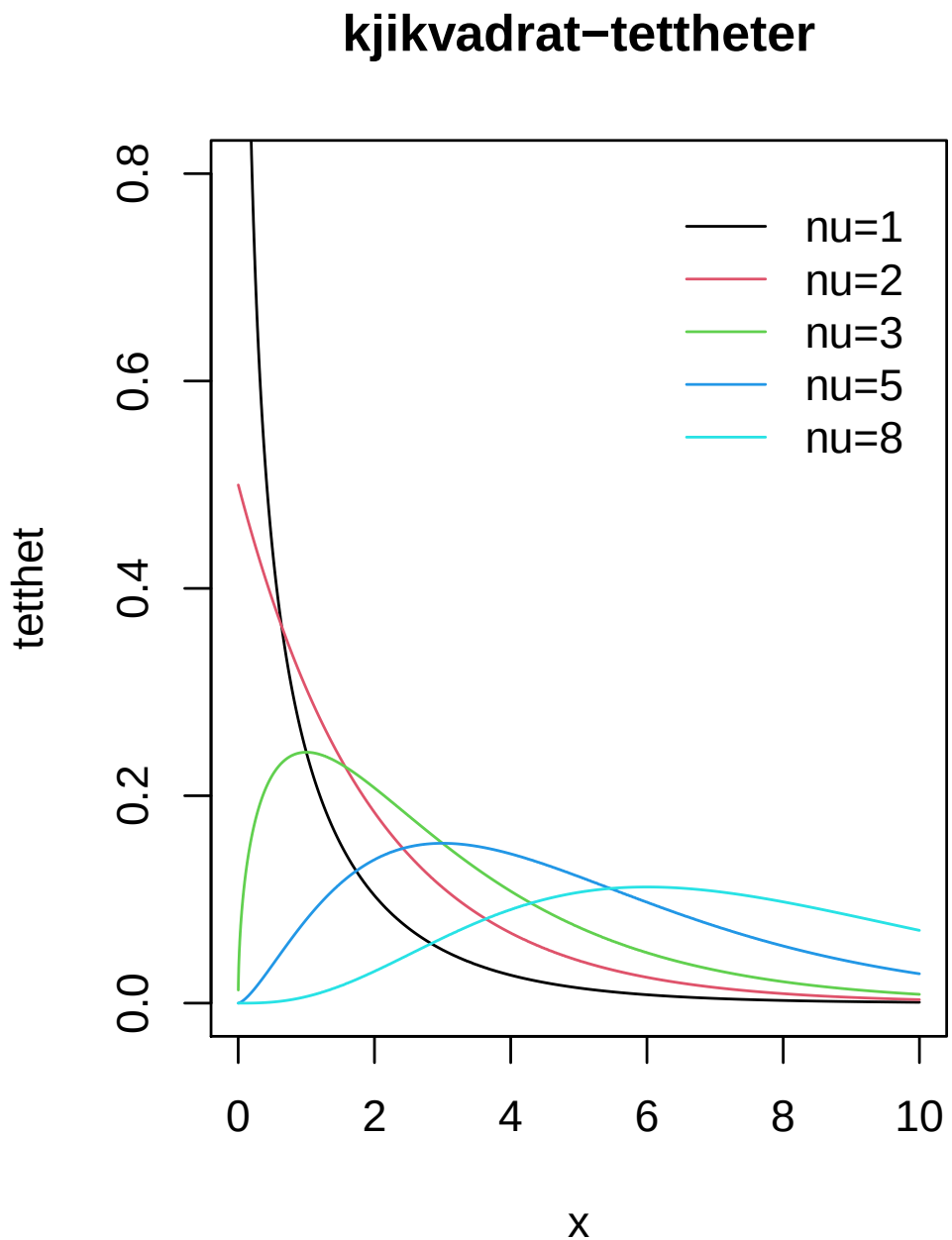
Men siden $X \sim \chi^2_\nu$ er en gamma$(\nu/2, 2)$ fordeling får vi

$$\mathrm{E}(X) = \alpha\beta = \nu \quad \text{og} \quad \mathrm{V}(X) = \alpha\beta^2 = 2\nu$$

Videre er 95-%-persentilen i χ^2_1 lik (ca.) $3.84 \approx 1.96^2$ (siden $Z \sim \mathrm{N}(0,1)$ betyr $Z^2 \sim \chi^2_1$)

Dessuten er χ^2_2-fordelingen en eksponensialfordeling med tetthet $\frac{1}{2}\exp(-\frac{x}{2})$ og 95-persentil 5.99.

Tettheter for noen χ^2_ν-fordelinger



Tetthet for χ^2_1-fordelingen, altså $\nu = 1$

La $\Phi(z)$ og $\phi(z)$ være hhv. kumulativ og tetthet for $Z_1 \sim \mathsf{N}(0,1)$.

Får da kumulativ fordelingsfunksjon for $X = Z_1^2 \sim \chi^2_1$ ved

$$\begin{aligned} F(x) &= \Pr(X \le x) = \Pr(Z_1^2 \le x) = \Pr(-\sqrt{x} \le Z_1 \le \sqrt{x}) \\ &= \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = \Phi(\sqrt{x}) - [1 - \Phi(\sqrt{x})] = 2\Phi(\sqrt{x}) - 1 \end{aligned}$$

Dette gir tetthet for $X = Z_1^2$ ved

$$\begin{aligned} f(x) = \quad F'(x) &= 2\Phi'(\sqrt{x})\tfrac{1}{2\sqrt{x}} = \phi(\sqrt{x})\tfrac{1}{\sqrt{x}} \\ &= \tfrac{1}{\sqrt{2\pi}} e^{-(\sqrt{x})^2/2} \tfrac{1}{\sqrt{x}} = \tfrac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \\ &= \tfrac{1}{2^{1/2}\Gamma(1/2)} x^{-1/2} e^{-x/2} \end{aligned}$$

som (jammen) er Gamma$(1/2, 2)$-tettheten!

χ^2-fordelingen (forts)

- Hvis $Z \sim N(0,1)$, så er altså $V = Z^2 \sim \chi^2_1$
- som er en Gamma(1/2,2)-fordeling

- Men summer av uavhengige gammafordelte variable med samme β-parameter er igjen gammafordelt med summen av α-parametre.
- Derfor med $Z_1, ..., Z_\nu \overset{uif}{\sim} N(0,1)$, så blir
  $V = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2 \sim \chi^2_\nu$
  altså Gamma$(\nu/2, 2)$-fordelt

- χ^2_ν-fordelingen kan utvides til ikke-heltallige ν (via Gamma$(\nu/2, 2)$-fordelingen), men vi trenger ikke dette i STK1110.

Ytterligere viktig resultat om $\chi 2$-fordelinger

Anta følgende tre betingelser

- V_1, V_2 er uavhengige,
- $V_3 = V_1 + V_2$ og
- $V_j \sim \chi^2_{\nu_j}$ for $j = 1, 3$

Da holder det at $V_2 \sim \chi^2_{\nu_3 - \nu_1}$

Bevis: Siden V_1 og V_2 er uavhengige, $V_j \sim \chi^2_{\nu_j}$ for $j = 1, 3$ og $V_3 = V_1 + V_2$ har vi

$$\begin{aligned} M_{V_3}(t) &= \frac{1}{(1-2t)^{\nu_3/2}} = M_{V_1+V_2}(t) = M_{V_1}(t) M_{V_2}(t) \\ &= \frac{1}{(1-2t)^{\nu_1/2}} M_{V_2}(t) \end{aligned}$$

og det følger at $M_{V_2}(t) = \frac{1}{(1-2t)^{(\nu_3-\nu_1)/2}}$, og altså $V_2 \sim \chi^2_{\nu_3-\nu_1}$

Resultater for normalfordelte utvalg

- Anta $X_1, ..., X_n \overset{uif}{\sim} N(\mu, \sigma)$
- Teorem: $\overline{X}$ og S^2 er uavhengige
- Bevis:

$$\begin{aligned}
\text{Cov}[X_i - \overline{X}, \overline{X}] &= \text{Cov}[X_i, \overline{X}] - \text{Cov}[\overline{X}, \overline{X}] \\
&= \text{Cov}[X_i, \tfrac{1}{n}\textstyle\sum_{k=1}^{n} X_k] - V[\overline{X}] \\
&= \tfrac{1}{n} V[X_i] + \tfrac{1}{n}\text{Cov}[X_i, \textstyle\sum_{k \neq i} X_k] - V[\overline{X}] \\
&= \tfrac{1}{n}\sigma^2 + 0 - \tfrac{1}{n}\sigma^2 = 0
\end{aligned}$$

- For (simultant) normalfordelte variable er ukorrelerthet ekvivalent med uavhengighet
- Det følger at S^2 som er en funksjon av $X_i - \overline{X}, i = 1, ..., n$ dermed er uavhengig av $\overline{X}$

Resultater for normalfordelte utvalg - forts

- Anta $X_1, ..., X_n \overset{uif}{\sim} N(\mu, \sigma)$
- Da er $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$

Bevis:

$$\begin{aligned}\sum_i (X_i - \mu)^2 &= \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2 \\ &= \sum_i (X_i - \overline{X})^2 + \sum_i (\overline{X} - \mu)^2 + 2\sum_i (X_i - \overline{X})(\overline{X} - \mu) \\ &= \sum_i (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2\end{aligned}$$

som gir

$$\underbrace{\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\chi^2_n} = \underbrace{\sum_i \left(\frac{X_i - \overline{X}}{\sigma}\right)^2}_{\frac{(n-1)S^2}{\sigma^2}} + \underbrace{\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2}_{\chi^2_1}$$

Og dermed $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Resultater for normalfordelte utvalg - forts

- Størrelse av interesse: $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$
- Har så langt
  - $\overline{X} - \mu \sim N(0, \sigma/\sqrt{n})$
  - $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
  - $\overline{X}$ og S^2 er uavhengige

- Trenger nå *t*-fordelingen
  - Hvis $Z \sim N(0,1)$ og $U \sim \chi^2_\nu$ er uavhengige så er

$$T = \frac{Z}{\sqrt{U/\nu}} \sim t_\nu \quad \text{dvs. t-fordelt med } \nu \text{ frihetsgrader}$$

- Får dermed

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t_{n-1}$$

som er meget nyttig for inferens om normalfordelte data

Mer om t-fordelinger

Vi definerte altså *t*-fordelingen på følgende måte:

Hvis $Z \sim N(0,1)$ og $U \sim \chi^2_\nu$ er uavhengige så er

$$T = \frac{Z}{\sqrt{U/\nu}} \sim t_\nu \quad \text{dvs. t-fordelt med } \nu \text{ frihetsgrader}$$

Det kan vises (DBC s.383, utelates) at tettheten i t_ν-fordelingen gis ved:

$$f(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Denne tettheten er symmetrisk om $t = 0$ og entoppet, men med tyngre haler en normalfordelingen.

Spesielt får vi for $\nu = 1$ at denne tettheten blir gitt som

$$f(t,1) = \frac{1}{\pi}\frac{1}{1+t^2},$$

en Cauchy-fordeling der hverken varians forventning eksisterer!

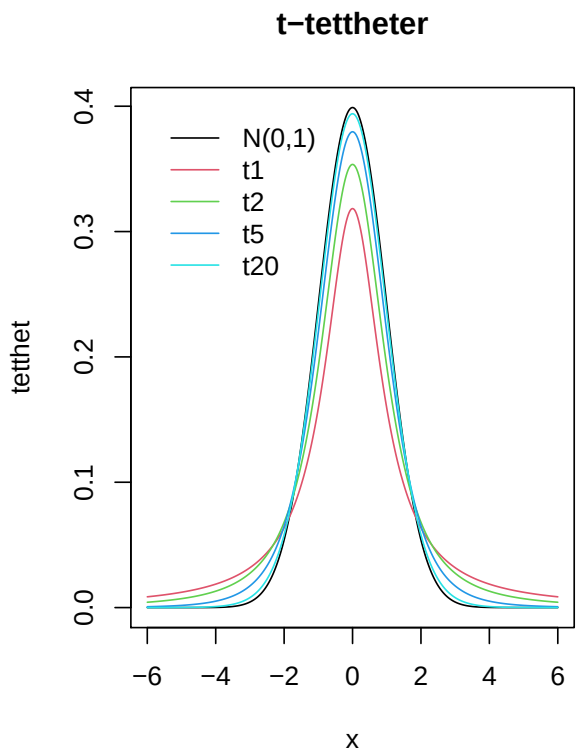
Ytterligere om t-fordelinger

Videre har vi for $T \sim T_\nu$ at

$$\begin{aligned} \mathrm{E}(T) &= 0 \quad \text{hvis } \nu > 1 \\ \mathrm{V}(T) &= \frac{\nu}{\nu-2} \quad \text{hvis } \nu > 2 \end{aligned}$$

Dermed har t-fordelingene i grensen samme forventning og varians som $\mathrm{N}(0,1)$ når $\nu \to \infty$.

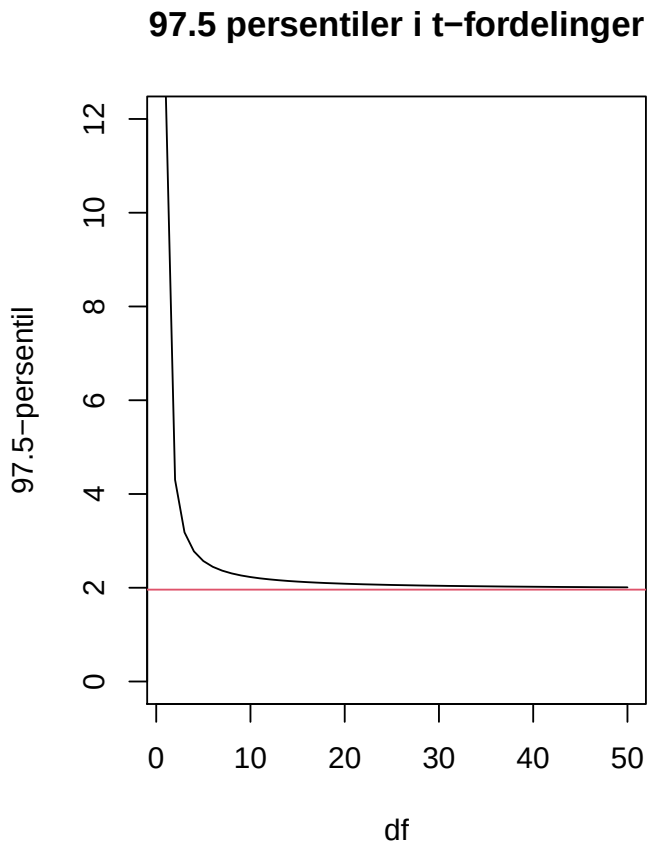
Det kan vises at t-fordelingene går mot N(0,1)-fordelingen, men har tyngre haler.



Kvantiler (persentiler) i t-fordelingene

Kvantiler t_α i t-fordelingene avviker mer fra null enn kvantiler i N(0,1)-fordelingen, men vil konverger relativt raskt mot kvantilene i denne når antall frihetsgrader $\nu \to \infty$.

Figuren under viser spesielt hvordan 97.5%-persentilen i t-fordelingene konvergerer mot 1.96.



Konfidensintervall for μ (forts.)

$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

- La $t_{\alpha,\nu}$ være øvre α-kvantil i t_ν-fordelingen.
- Vi har eksakt at

$$\mathrm{P}\left(-t_{\alpha/2,n-1} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1-\alpha.$$

- Det kan omformes til

$$\mathrm{P}\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Konfidensintervall for μ (forts.)

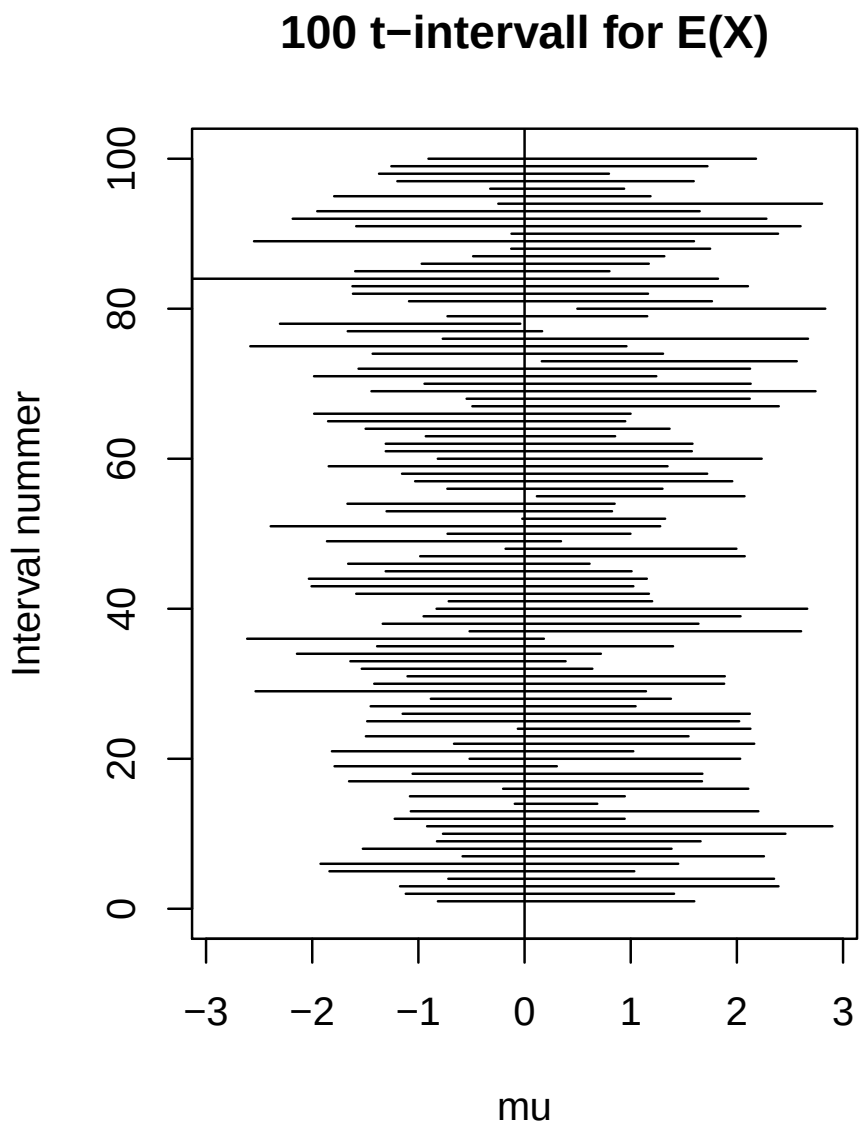
- Dermed inneholder det stokastiske intervallet

$$\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right),$$

  μ med $100 \cdot (1-\alpha)\%$ sannsynlighet.
- Ved å sette inn observerte verdier får vi et $100 \cdot (1-\alpha)\%$ konfidensintervall for μ:

$$\left(\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right).$$

Dekningsgrad for t-intervaller for μ



Bestemmelse av utvalgsstørrelse:

Hvor stor n for lengde w med konfidensgrad α:

- For kjent σ:

$$n = \left(\frac{z_{\alpha/2}\sigma}{w/2} \right)^2$$

- For ukjent σ:

$$n = \left(\frac{t_{\alpha/2;n-1}s}{w/2} \right)^2$$

  To problemer:

  - s ukjent. Kan anslås fra tidligere lignende eksperimenter
  - n inngår på begge sider. Kan løses numerisk, evnt bruke $z_{\alpha/2}$ istedet.

Prediksjonsintervall for en ny observasjon

- Generelt har vi denne situasjonen: $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu, \sigma)$ og vi vil predikere verdien til en ny observasjon X_{n+1}.
- En prediktor for X_{n+1} er $\overline{X}$.
- Prediksjonsfeilen er da $\overline{X} - X_{n+1}$ og oppfyller

$$\overline{X} - X_{n+1} \sim N\left(0, \sigma\sqrt{1 + \frac{1}{n}}\right) \text{ og } Z = \frac{\overline{X} - X_{n+1}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0, 1).$$

- Videre er $\overline{X} - X_{n+1}$ og S^2 uavhengige, slik at

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}.$$

Prediksjonsintervall for en ny observasjon (forts.)

- Ved å følge samme framgangsmåte som for konfidensintervallet for μ, får vi et $100 \cdot (1-\alpha)\%$ prediksjonsintervall for X_{n+1} gitt ved

$$\left(\overline{X} - t_{\alpha/2,n-1} s \sqrt{1+\frac{1}{n}},\ \overline{X} + t_{\alpha/2,n-1} s \sqrt{1+\frac{1}{n}} \right).$$

- Prediksjonsintervallet for en ny observasjon er mye videre enn konfidensintervallet for forventningen.
- Konfidensintervallet vil kollapse mot verdien μ når $n \to \infty$ da all usikkerhet om parameterne forsvinner.
- Prediksjonsintervallet, derimot, går mot $\mu \pm z_{\alpha/2}\sigma$ da det alltid er en usikkerhet i den nye verdien en vil predikere, også når parameterne er kjent.

Konfidensintervall for σ^2 og σ (Avsn. 8.4)

- Vi antar at vi har observasjoner $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu, \sigma)$, der både μ og σ er ukjent.
- Vi ønsker nå å lage et $100 \cdot (1-\alpha)\%$ konfidensintervall for σ^2 og σ.
- Fra Kap 6.4 vet vi at $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- La $\chi^2_{\alpha,\nu}$ være øvre α-kvantil i χ^2_ν-fordelingen.
- Da har vi:

$$\mathrm{P}\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha.$$

Konfidensintervall for σ^2 og σ, forts.

- Det kan omformes til

$$\mathrm{P}\left(S^2 \frac{n-1}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq S^2 \frac{n-1}{\chi^2_{1-\alpha/2,n-1}} \right) = 1-\alpha$$

$$\mathrm{P}\left(S\sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}} \leq \sigma \leq S\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}} \right) = 1-\alpha.$$

- Det gir $100 \cdot (1-\alpha)\%$ konfidensintervall for:

$$\sigma^2 : \left(s^2 \frac{n-1}{\chi^2_{\alpha/2,n-1}}, s^2 \frac{n-1}{\chi^2_{1-\alpha/2,n-1}} \right)$$

$$\sigma : \left(s\sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}}, s\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}} \right).$$

Kommentarer konfidensintervall for μ og σ

Med data $X_1, \ldots, X_n$ uif med $\mu = \mathrm{E}(X_i)$ og $\mathrm{V}(X_i) = \sigma^2$ har vi at

$$\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$$

er et tilnærmet 95% konfidensintervall for μ når n stor.

Med eksakt normalfordeling for X_i-ene har vi videre et eksakt 95%-konfidensintervall for μ med

$$\bar{X} \pm t_{0.025} \frac{s}{\sqrt{n}}$$

der $t_{0.025} = 97.5\%$-persentilen i t-fordelingen med $n-1$ frihetsgrader - også når n er liten.

Dette intervallet er en fornuftig forbedring også når X_i-ene ikke er normalfordelt.

Likevel kan det være nyttig å sammenligne intervallet med bootstrap-intervaller!

Kommentarer konfidensintervall for μ og σ

Vi fant også konfidensintervall for σ og σ^2 under forutsetning $X_i \sim \mathrm{N}(0, \sigma)$ uif basert på kvantiler i χ^2_{n-1}-fordelingen.

For disse intervallene er imidlertid antagelsen om normalfordeling essensiell! og intervallene må brukes med forsiktighet.

Hvis man trenger konfidensintervall for σ anbefales derfor bootstrap på det sterkeste!

Bootstrap er dessuten særlig nyttig for intervallestimering der vi ikke har noen enkel varianssformel:

- Eks: Median og andre persentiler

Konfidensintervaller basert på bootstrap

- Vi antar at vi har observasjoner $x_1, \ldots, x_n$ av de stokastiske variablene $X_1, \ldots, X_n$, som er uif med en fordeling som avhenger av en ukjent parameter θ.
- Vi ønsker å estimere θ, samt lage et $100 \cdot (1 - \alpha)\%$ konfidensintervall for θ.
- Anta videre at vi har estimatoren $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, som gir estimatet $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$.
- Vi skal nå bruke ikke-parametrisk bootstrap til å lage konfidensintervall.

Konfidensintervaller basert på bootstrap (forts.)

- Vi begynner med å bootstrappe $\hat{\theta}$:
  1. For $b = 1, \ldots, B$
  2. Trekk $x^*_{b1}, \ldots, x^*_{bn}$ fra $x_1, \ldots, x_n$ med tilbakelegging.
  3. La $\hat{\theta}^*_b = \hat{\theta}(x^*_{b1}, \ldots, x^*_{bn})$.
- Basert på $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ kan en lage konfidensintervall på flere måter.
- Dersom disse er tilnærmet normalfordelt med forventning θ og standardavvik $\sigma_{\hat{\theta}}$, kan en estimere $\sigma_{\hat{\theta}}$ med $s_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \bar{\hat{\theta}}^*)^2}$, med $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b$.
- En bruker da at $\frac{\hat{\theta}-\theta}{s_{boot}} \overset{tiln.}{\sim} N(0,1)$.
- Da er $\hat{\theta} \pm z_{\alpha/2} s_{boot}$ et tilnærmet $100 \cdot (1-\alpha)\%$ konfidensintervall for θ.

Konfidensintervaller basert på bootstrap (forts.)

- Når normalfordelingsantakelsen ikke holder, kan vi i stedet bruke **persentilintervallet**.

- Det beregnes på følgernde måte:
  1. Sorter bootstrap-estimatene, slik at $\hat{\theta}^*_{(1)} \leq \ldots \leq \hat{\theta}^*_{(B)}$.
  2. Persentilintervallet for θ blir da $\left(\hat{\theta}^*_{(B\alpha/2)}, \hat{\theta}^*_{(B(1-\alpha/2))}\right)$.

Konfidensintervaller basert på bootstrap (forts.)

- I visse tilfeller er det behov for å gjøre noen korreksjoner av persentilintervallet.
- F.eks. kan det være skjevhet i bootstrapfordelingen sammenlignet med det opprinnelige estimatet $\hat{\theta}$.
- Denne skjevheten kan en vurdere ved å beregne $\hat{\theta} - \bar{\hat{\theta}}^*$.
- Et annet problem kan være at $\sigma_{\hat{\theta}}$ avhenger av θ.
- I slike situasjoner kan en i stedet bruke BCa-intervallet (står for "bias corrected and accelerated"), som prøver å korrigere problemene over ved å bruke andre persentiler enn $\alpha/2$ og $1 - \alpha/2$.

Eksempel

Tipsdata

Avsnitt 8.3: Konfidensintervall for binomisk p

Med $Y \sim \text{bin}(n, p)$ er MLE for p like $\hat{p} = \frac{Y}{n}$.

På Slide 16 fant vi at $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ er et tilnærmet 95% konfidensintervall for p.

Intervallet er basert på at $Z_0(p) = \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} \sim N(0,1)$ tilnærmet når $np > 10$ og $n(1-p) > 10$.

For små np eller $n(1-p)$ kan intervallet være utilfredstillende. Vi kan da finne et bedre intervall basert på tilnærmingen for

$$Z_1(p) = \frac{\hat{p}-p}{\sqrt{p(1-p)}}\sqrt{n} \sim \text{N}(0,1),$$

mer presist på at $-1.96 < Z_1(p) < 1.96$ som er ekvivalent med

$$(\hat{p}-p)^2 < 1.96^2 \frac{p(1-p)}{n}$$

som er en 2. gradsulikhet i p og dermed kan løses eksplisitt ved formelen på s. 476 i Læreboka.

Ytterligere forbedret KI for p

Vi kan videre benytte at tilnærming av diskret Y til kontinuerlig fordeling forbedres ved å "smøre" punktsannsynlighetene $P(Y = y)$ utover intervallet $(y - \frac{1}{2}, y + \frac{1}{2})$, ved *halvkorreksjon*

$P(Y \leq y) \approx \Phi(\frac{y+0.5-np}{\sqrt{np1-p}})$ (bedre enn $P(Y \leq y) \approx \Phi(\frac{y-np}{\sqrt{np1-p}})$)

