HG
Revidert april 2013

Oversikt over konfidensintervall i Econ 2130

Merk at denne oversikten ikke er ment å leses istedenfor framstillingen i Løvås, men som et supplement. Den inneholder tabeller med formler for konfidensintervaller for situasjoner som er aktuelt å kjenne til i dette kurset. Den inneholder også noen eksempler på bruk av formlene. Utfordringen for studentene i oppgaver blir således derfor først og fremst å kunne gjenkjenne situasjonen i oppgaven og derfor plukke ut korrekt formel for konfidensintervallet. Oversikten inneholder også noen detaljer som jeg ikke rekker å snakke om på de få forelesningene som gjenstår.

1 Generell innledning med noen presiseringer og et regneeksempel

La θ være en ukjent parameter (populasjons-størrelse) i en statistisk modell. Uttrykket "ukjent parameter" betyr at den "sanne" verdien av θ i populasjonen er ukjent. Når vi setter opp en statistisk modell (som representerer populasjonen vi trekker data fra og trekningsprosedyren), antar vi i utgangspunktet at modellen er sann for en viss (ukjent) verdi av parameteren θ og usann for alle andre verdier. Anførselstegnene rundt "sann" ovenfor skyldes at begrepet *sann parameterverdi* kun gir god mening i relasjon til populasjonen dersom forutsetningene som er foretatt i modellen er realistiske forutsetninger om populasjonen og måten data er trukket på.

La $\hat{\theta}$ være en aktuell estimator for θ, og $SE = SE(\hat{\theta})$ står for en eller annen estimert versjon av standardfeilen til $\hat{\theta}$. (**Husk (NB!) at hvis $\hat{\theta}$ er forventningsrett, er standardfeilen til $\hat{\theta}$ ikke noe annet enn standardavviket til $\hat{\theta}$, nemlig $\sqrt{\text{var}(\hat{\theta})}$.**)

Det viser seg at alle konfidensintervall (KI) i pensum (inklusivt regresjonsanalysen) - med et unntak i tabell 2 – koker ned til samme form:

$$\hat{\theta} \pm c \cdot SE(\hat{\theta})$$

der *c* er en kvantil bestemt av den valgte konfidensgraden. Denne kvantilen er som oftest fra $N(0,1)$-fordelingen og noen ganger fra *t*-fordelingen (se situasjon 2 i tabell 1 (. Med konfidensgraden $1-\alpha$, er for eksempel $c = z_{\alpha/2}$ (dvs $\alpha/2$-kvantilen i $N(0,1)$) i situasjon 1 og 3 i tabell 1 og i alle situasjoner i tabell 3.

Årsaken til at denne typen av KI er så vanlig er at det ofte finnes teoremer (som for eksempel sentralgrenseteoremet og regel 5.20 og andre lignende) som viser at estimatoren $\hat{\theta}$ er tilnærmet (i noen få tilfeller eksakt) normalfordelt, $\hat{\theta} \overset{\text{tilnærmet}}{\sim} N(\theta, \sqrt{\text{var}(\hat{\theta})}) = N(\theta, SE(\hat{\theta}))$. Dette innebærer (jfr. regel R1 i notat til kap. 5 om normalfordelingen) at

(*) $$\frac{\hat{\theta}-\theta}{SE(\hat{\theta})} \overset{\text{tilnærmet}}{\sim} N(0,1)$$

(*) gjelder bare hvis utvalgsstørrelsen (antall observasjoner) ikke er for liten (se eksempel 1 under). Løvås viser side 226 hvordan utsagnet (*) leder til konfidensintervallet formulert i **regel 6.7.** Dessverre er **regel 6.7** unødvendig snevert formulert hos Løvås med få anvendelser (det er få situasjoner der $\hat{\theta}$ er eksakt normalfordelt, men mange situasjoner der $\hat{\theta}$ er tilnærmet normalfordelt). Vi blir derfor nødt til å gi en modifisert reformulering av regel 6.7 for å gjøre den mer anvendelig:

Regel 6.7 (Løvås side 225) modifisert. (Konfidensintervall basert på normalfordelingen).

(a) Hvis estimatoren $\hat{\theta}$ er forventningsrett og *tilnærmet* normalfordelt med standardfeil $SE(\hat{\theta})$, vil følgende intervall være et *tilnærmet* $100(1-\alpha)\%$ konfidensintervall for θ

(**) $[\hat{\theta} - z_{\alpha/2} \cdot SE(\hat{\theta}),\ \hat{\theta} + z_{\alpha/2} \cdot SE(\hat{\theta})]$

Hvis $\dfrac{\hat{\theta}-\theta}{SE(\hat{\theta})}$ i (*) er eksakt normalfordelt, $N(0,1)$, vil intervallet ha eksakt konfidensgrad $1-\alpha$ (eller $100(1-\alpha)\%$).

(b) Videregående teoremer i sannsynlighetsteori viser at i situasjoner der standardfeilen, $SE(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})}$ er ukjent (dvs avhenger av ukjente parametre i modellen) så vil under generelle betingelser konfidensintervallet fortsatt ha konfidensgrad tilnærmet $100(1-\alpha)\%$ om standardfeilen byttes ut med en estimert versjon. Med andre ord, utsagnet (*) - og dermed (**) - gjelder fortsatt om $SE(\hat{\theta})$ nå står for estimert standardfeil.

(a) begrunnes[1] som i Løvås side 226 der den eneste forskjellen er at det første likhetstegnet byttes ut med $\approx$:

$$1-\alpha \approx P\left(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{SE(\hat{\theta})} \le z_{\alpha/2}\right) = \cdots \text{ som i Løvås side 226 (gjør det selv!) } \cdots = P\left(\hat{\theta} - z_{\alpha/2} \cdot SE(\hat{\theta}) \le \theta \le \hat{\theta} + z_{\alpha/2} \cdot SE(\hat{\theta})\right)$$

(b) bygger på videregående sannsynlighetsteori (delvis tatt opp i Stat2-kurset og mye brukt i økonometrisk teori) og som ikke behandles her i Stat1.

For øvrig: *Når det gjelder tolkningen av begrepet konfidensgrad, les diskusjonen til figur 6.3 i Løvås side 211.*

Regneeksempel 1 **(Basert på oppgave 5.6 i Løvås med ny problemstilling)**. En spesialpedagog skal undersøke læreevnen til $n = 900$ tilfeldig utvalgte elever. I oppgave 5.6 antas at andelen av alle skolebarn (populasjonen) som har lærevansker, er $p = 0{,}15$,

[1] Dette er en viktig manipulasjon som vi krever at studentene behersker og forstår.

altså kjent. Vi skal nå i stedet anta at *p* er ukjent og at 0.15 er et estimat for *p* basert på utvalget av 900 elever. Vi er interessert i å beregne usikkerheten ved dette anslaget uttrykt ved et 95% konfidensintervall for *p*. For å komme noen vei, trenger vi en statistisk modell for populasjonen og utvalgsmetoden.

Modell. *La X være antall barn med lærevansker i et rent tilfeldig utvalg på* $n = 900$ *elever trukket fra populasjonen av alle skolebarn. Anta* $X \sim \text{bin}(n, p)$ *der p er andelen av skolebarn i populasjonen med lærevansker og antas ukjent.*

> **Merknad til modellen.** Merk at utvalget er forutsatt representativt. Dette ligger i forutsetningen om at utvalget er "rent tilfeldig" som ideelt sett (sjelden eksakt oppfylt i praksis, men ofte akseptabelt bra oppfylt) betyr at alle mulige utvalg på 900 fra populasjonen har samme sannsynlighet for å bli trukket ut. *X* er, under denne forutsetningen, strengt tatt hypergeometrisk fordelt, men, siden populasjonen er stor, kan vi uten vesentlig tap av realisme anta at *X* er binomisk fordelt - som gir en enklere modell.

Anta at pedagogen fant 135 barn med lærevansker i utvalget. Tallet 135 er nå å oppfatte som en observasjon av den stokastiske variabelen, *X*. Den vanlige estimatoren i denne modellen er $\hat{p} = \frac{X}{n}$. Estimatet (dvs den observerte verdien $\hat{p}_{obs}$ basert på data) får vi ved å sette data inn i estimatoren, $\hat{p}_{obs} = \frac{135}{900} = 0.15$. Oppgaven er altså å beregne et konfidensintervall for den ukjente *p* med konfidensgrad (tilnærmet) 95%:

Om estimatoren $\hat{p}$ vet vi følgende ut fra teorien som er etablert i kurset til nå:

(a) $\hat{p}$ er forventningsrett.

[**Begrunnelse:** $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$ (jfr. regel 5.3)]

(b) Standardfeilen for $\hat{p}$ er $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

[**Begrunnelse:** $\text{var}(\hat{p}) = \text{var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$. Dermed $SE(\hat{p}) = \sqrt{\text{var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$]

(c) $\hat{p}$ er tilnærmet normalfordelt, $\hat{p} \overset{\text{tilnærmet}}{\sim} N(E(\hat{p}), \sqrt{\text{var}(\hat{p})}) = N(p, SE(\hat{p}))$.

[**Begrunnelse.** Dette følger av regel 5.20 som sier at hvis $\sigma^2 = \text{var}(X) = np(1-p) \geq 5$ og p ikke er veldig nær 0 eller 1, så er X tilnærmet normalfordelt, $X \overset{\text{tilnærmet}}{\sim} N(E(X), \sqrt{\text{var}(X)}) = N(np, \sqrt{np(1-p)})$. Siden $n = 900$, synes betingelsen klart å være oppfylt. Dermed kan vi bruke regel R1 (i notat til kap. 5 om normalfordelingen) som viser at

$$\hat{p} = \frac{1}{n}\cdot X \overset{\text{tilnærmet}}{\sim} N\left(\frac{1}{n}\cdot E(X),\ \frac{1}{n}\cdot SD(X)\right) = N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right) = N\left(p, SE(\hat{p})\right) \quad]$$

Av dette[2] følger at den standardiserte $\hat{p}$, $\frac{\hat{p}-p}{SE(\hat{p})}$, er tilnæmet $N(0,1)$-fordelt. Nå er standardfeilen, $SE(\hat{p})$ ukjent siden p er ukjent. Dermed, når n er stor nok (slik at $\text{var}(X) = np(1-p) \geq 5$), vil i følge modifisert regel 6.7 **(b)** denne tilnærmelsen fortsatt være akseptabel om vi erstatter den ukjente standardfeilen med estimert standard feil $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. I tråd med notasjonen slik Løvås (og Excel og STATA og andre pakker) bruker den, lar vi nå $SE(\hat{p})$ stå for den *estimerte* versjonen. Utsagnet (*) blir i denne situasjonen dermed seende ut som

$$\frac{\hat{\theta}-\theta}{SE(\hat{\theta})} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{\text{tilnærmet}}{\sim} N(0,1)$$

[2] Se det andre eksempelet etter regel R1 i notatet om normalfordeling på nettet.

som gir et tilnærmet $(1-\alpha)100\%$ KI for p: $\hat{p} \pm z_{\alpha/2} SE(\hat{p}) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Med ønsket konfidensgrad 95% trenger vi kvantilen $z_{0.025} = 1.96$ (se tabell D4 bak i Løvås), og konfidensintervallet blir utregnet som

$$\hat{p}_{obs} \pm 1.96 \cdot \sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}} = 0.15 \pm (1.96) \cdot \sqrt{\frac{(0.15)(0.85)}{900}} = 0.15 \pm 0.02 = [0.13,\ 0.17]$$

Merknad 1. *Usikkerheten* ved anslaget $\hat{p}_{obs} = 0.15$ er således i dette eksemplet beregnet til ± 0.02 (generelt $\pm z_{\alpha/2} SE(\hat{\theta})$). Vi ser dermed at begrepet "usikkerhet" ved en estimering ikke er noen absolutt størrelse. Den avhenger ikke bare av utvalgsstørrelse (n) og populasjonsvariansen til X (her $np(1-p)$ som er variansen for antall suksesser i et enkelt binomisk forsøk), men også av den subjektivt valgte konfidensgraden!

Merknad 2. (For å berolige leseren). Om du til eksamen blir bedt om å beregne et konfidensintervall som i eksemplet, trenger du naturligvis ikke, om du ikke eksplisitt blir spurt om det, å komme opp med hele begrunnelsen ovenfor. Det vil vanligvis være tilstrekkelig simpelthen å velge riktig formel i forhold til den aktuelle modellen og å kunne sette inn tallene korrekt. Du kan naturligvis ved tilleggsspørsmål risikere å bli bedt om å gjennomføre deler av argumentasjonen ovenfor for aktuelle modell-typer som omfattes av pensum.

2 Aktuelle modelltyper 1

En vanlig modelltype er *uid*-modellen (1) (engelsk *iid*)

(1) *La* $X_1, X_2, \ldots, X_n$ *være uavhengige og identisk fordelte stokastiske (uid) variable med* $E(X_i) = \mu$ og $\text{var}(X_i) = \sigma^2$, *der* μ og σ *tolkes som størrelser (som oftest ukjente) i en eller annen populasjon som data,* $x_1, x_2, \ldots, x_n$ *trekkes fra.*

Aktuelle estimatorer: $\hat{\mu} = \bar{X}$ og $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ som begge er forventningsrette (jfr avsnittet under regel 6.2 og regel 6.13).

La $\alpha/2$ -kvantilen i $N(0,1)$ - fordelingen betegnes med $z_{\alpha/2}$ (slik at $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$, der $Z \sim N(0,1)$)

La $\alpha/2$ -kvantilen i t_{n-1} - fordelingen betegnes med $t_{n-1,\alpha/2}$ (slik at $P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha$, der $T \sim t_{n-1}$).

Merk at fordelingene t_{n-1} og $N(0,1)$ ligner på hverandre: De er begge entoppet (klokkeformet) og symmetrisk rundt 0. Når *n* er "stor" (dvs. ≥ 30 omtrent), er forskjellen neglisjerbar. For små *n* er t_{n-1} karakterisert ved litt tyngre haler enn $N(0,1)$ og litt flatere kurve rundt 0 (jfr. figur 5.26 side 192 i Løvås).

Tabell 1 Konfidensintervall for μ

Situasjon	*Forutsetninger (modell)*	*n*	σ	*Standardfeil* $\sqrt{\text{var}(\hat{\mu})}$	*Estimert standardfeil*	*Pivotal* $\frac{\hat{\mu}-\mu}{SE(\hat{\mu})}$	$1-\alpha$ *KI for* μ	*Konfidens-grad*
1	(1) pluss forutsetningen $X_i \sim N(\mu, \sigma), \quad i = 1,2,\ldots,n$	Vilkårlig	Kjent	$\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	Eksakt $1-\alpha$
2	(1) pluss forutsetningen $X_i \sim N(\mu, \sigma), \quad i = 1,2,\ldots,n$	Vilkårlig	Ukjent	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$	$\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$	Eksakt $1-\alpha$
3	Bare (1) der X_i er vilkårlig fordelt	n "stor", $n \ge 30$ (til nød ≥ 20)	Ukjent	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{\text{tilnærmet}}{\sim} N(0,1)$	$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$	Tilnærmet $1-\alpha$
4	Bare (1) der X_i er vilkårlig fordelt	n liten	Ukjent				**Ikke pensum**	

Merknad 3. Uttrykket "**pivotal**" betegner en stokastisk variabel som avhenger av ukjente parametre i modellen - en variabel som derfor **ikke er observerbar** - men som har **kjent** sannsynlighetsfordeling. Pivotaler er bl.a nyttige ved konstruksjon av konfidensintervaller og tester.

Merknad 4. Siden t_{n-1}-fordelingen er tilnærmet lik $N(0,1)$ for $n \geq 30$, vil forskjellen mellom KI-ene i situasjon 2 og 3 være neglisjerbar når $n \geq 30$.

Tabell 2. Konfidensintervall for σ^2 når $X_1, X_2, \ldots, X_n$ er uavhengige og normalfordelte med $X_i \sim N(\mu, \sigma)$.

(Jeg vil antakelig ikke rekke å snakke om dette på forelesningene, så dette må leses på egenhånd[3].)

Hvis en stokastisk variabel, V, er kji-kvadratfordelt (avsnitt 5.9.1) med k frihetsgrader, skriver vi kort: $V \sim \chi^2_k$-fordelt (χ er den greske bokstaven "kji"). p-kvantilen i denne fordelingen kaller Løvås, χ_p, som er det tallet som oppfyller $P(V > \chi_p) = p$. Noen kvantiler finnes i tabell D6. Merk at kji-kvadrat fordelingen ikke er symmetrisk (jfr. figur 5.25 side 190 i Løvås) slik at vi trenger kvantiler i begge ender av fordelingen for å utlede konfidensintervallet. Se **merknad 6**.)

Modell	*n*	*Estimator*	*Pivotal*	*Nedre konfidensgrense*	*Øvre konfidensgrense*	*Konfidensgrad*
$X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte (*uid*) med $X_i \sim N(\mu, \sigma), \quad i = 1, 2, \ldots, n$	Vilkårlig	$\hat{\sigma}^2 = S^2$	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (Regel 5.22)	$\frac{(n-1)S^2}{\chi_{\alpha/2}}$	$\frac{(n-1)S^2}{\chi_{1-\alpha/2}}$	Eksakt $1-\alpha$

Merknad 5. Har vi funnet et KI for populasjonsvariansen, σ^2, kan vi lett finne et for standardavviket, σ, også. Hvis $[A, B]$ er et

[3] Les avsnitt 5.9.1 (side 190) om kji-kvadratfordelingen, med spesiell vekt på regel 5.22, og avsnitt 6.3.4 (side 230) for relevant anvendelse.

$1-\alpha$ KI for σ^2 (der A og B er positive stokastiske variable) slik at $P(A \le \sigma^2 \le B) = 1-\alpha$, så vil et $1-\alpha$ KI for σ rett og slett være gitt ved $[\sqrt{A}, \sqrt{B}]$. Dette skyldes at begivenhetene $(A \le \sigma^2 \le B)$ og $(\sqrt{A} \le \sigma \le \sqrt{B})$ er logisk ekvivalente og derfor like sannsynlige (siden funksjonen $y = \sqrt{x}$ er en voksende funksjon av *x*). [*Illustrer selv den siste setningen med et diagram over funksjonen* $y = \sqrt{x}$!].

Merknad 6. Utledning av konfidensintervallet for σ^2. (Jfr. avsnitt 6.3.4 i Løvås.) Sett $V = (n-1)S^2/\sigma^2$. I følge regel 5.22 er $V \sim \chi^2_{n-1}$-fordelt. For kvantilene[4] $\chi_{1-\alpha/2}$ og $\chi_{\alpha/2}$ har vi i følge definisjonen av kvantiler og det at kji-kvadratfordelingen er kontinuerlig, $P(V < \chi_{1-\alpha/2}) = 1 - P(V \ge \chi_{1-\alpha/2}) = 1 - P(V > \chi_{1-\alpha/2}) = 1-(1-\alpha/2) = \alpha/2$ og $P(V > \chi_{\alpha/2}) = \alpha/2$. Dermed blir (se **figur 1**): $P(\chi_{1-\alpha/2} \le V \le \chi_{\alpha/2}) = 1-\alpha$. Ved innsetting for V får vi dermed

$$1-\alpha = P\left(\chi_{1-\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi_{\alpha/2} \right) = P\left(\frac{1}{\chi_{1-\alpha/2}} \ge \frac{\sigma^2}{(n-1)S^2} \ge \frac{1}{\chi_{\alpha/2}} \right) =$$

$$= P\left(\frac{(n-1)S^2}{\chi_{1-\alpha/2}} \ge \sigma^2 \ge \frac{(n-1)S^2}{\chi_{\alpha/2}} \right) = P\left(\frac{(n-1)S^2}{\chi_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi_{1-\alpha/2}} \right)$$

I den siste likheten har vi bare ordnet om på ulikheten slik at den minste verdien kommer til venstre. Merk også at den andre likheten skyldes at når man tar den inverse av begge sider av en ulikhet mellom positive tall, snur ulikheten rundt (for eksempel $4 > 2 \Leftrightarrow 1/4 < 1/2$).

Regneeksempel 2. For de $n = 37$ kvinnehøydene (døtrene), $y_1, y_2, \ldots, y_{37}$ vi samlet inn i fjor på forelesningen 5. mars 2012, ble estimatet for populasjons-standardavviket, σ, lik $\hat{\sigma}_{obs} = S_{obs} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 5.229457$. Vi ønsker et 95% KI for σ. Som modell bruker

[4] Hvis generelt *V* er χ^2_k-fordelt (med *k* frihetsgrader), er *p*-kvantilen i Løvås definert som *et tall*, skrevet χ_p, som oppfyller $P(V > \chi_p) = p$.

vi (1) for de bakenforliggende stokastiske variablene, $Y_1, Y_2, \ldots, Y_{37}$, og antar i tillegg at de er normalfordelte, $Y_i \sim N(\mu, \sigma)$ for $i = 1, 2, \ldots, n$. (Normalfordelingsantakelsen anses vanligvis for realistisk for høydemålinger i homogene grupper.)

Konfidensgrad 0.95, gir $\alpha = 0.05$ og $\alpha/2 = 0.025$. Vi trenger altså kvantilene $\chi_{0.975}$ og $\chi_{0.025}$ i kjikvadratfordelingen med $n-1 = 36$ frihetsgrader. Tabell D6 gir kun $\chi_{0.975}$=20.57 og $\chi_{0.025} = 53.20$ for den nærmeste kji-kvadratfordelingen som er χ^2_{35}-fordelingen. χ^2_{36}-fordelingen er ikke representert i tabell D6, men vi kan bruke CHIINV-funksjonen i Excel for 36 frihetsgrader som gir $\chi_{0.975}$=21.34 og $\chi_{0.025} = 54.44$

Ut fra **merknad 5** blir 95% konfidensintervallet for σ beregnet til

$$\left[\sqrt{\frac{36S^2}{\chi_{0.025}}},\ \sqrt{\frac{36S^2}{\chi_{0.975}}}\right]_{obs} = \left[S\sqrt{\frac{36}{54.44}},\ S\sqrt{\frac{36}{21.34}}\right]_{obs} = \left[(0.81)S, (1.30)S\right]_{obs} = \left[4.24,\ 6.80\right]$$

Merk at estimatet $\hat{\sigma}_{obs} = 5.23$ ikke ligger midt i konfidensintervallet (det er altså større usikkerhet til høyre for estimatet enn til venstre - som skyldes ar kji-kvadratfordelingen er en skjev fordeling). Dette innebærer at begrepet standardfeil ikke kommer inn som noe nyttig begrep i dette tilfellet (og blir derfor ikke brukt i forbindelse med estimering av σ eller σ^2), i motsetning til konfidensintervall basert på normalfordelingen eller t-fordelingen som ovenfor.

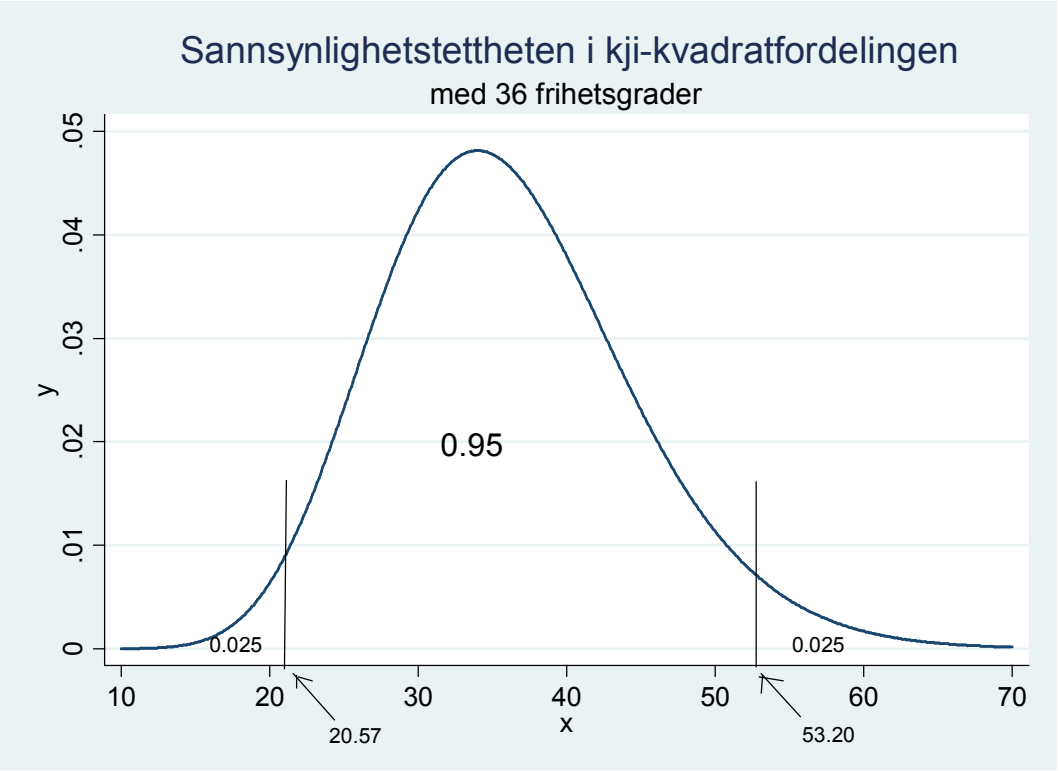
Om vi ønsker et 95% konfidensintervall for variansen, σ^2, får vi det, på grunn av **merknad 5**, ved rett og slett å kvadrere tallene i intervallet for σ: $\left[(4.24)^2, (6.80)^2\right] = \left[17.94,\ 46.22\right]$, der jeg bare tok med to desimaler i svaret for å gjøre intervallet lettere å lese i en eventuell rapport (de siste desimalene har liten tolkningsverdi uansett).

Merknad 7. Noen ganger finner vi altså ikke akkurat den fraktilen i tabellen (D6) vi er ute etter., Til eksamen, for eksempel, har vi ikke tilgang til Excel. Da er det lov å bruke "øyemålsmetoden" (i mangel av interpolasjonsmetoder som ikke er pensum). I så fall ser vi på de to nærmeste fordelingene som er representert:

Frihets-grader	$\chi_{0.975}$	$\chi_{0.025}$
35	20.57	53.20
40	24.43	59.34

På øyemål anslår vi for eksempel $\chi_{0.975}$=21.5 og $\chi_{0.025} = 54.2$ omtrent for 36 frihetsgrader, som er godt nok i en eksamensbesvarelse.

Figur 1 **χ^2-fordelingen. Graf laget med STATA**



3 Aktuelle modelltyper 2

Tabell 3 Tilnærmet konfidensintervall basert på regel 5.20 (normaltilnærming for binomisk, hypergeometrisk og poisson fordeling)

Modell	*Estimator* $\hat{\theta}$	*Standardfeil* $\sqrt{\text{var}(\hat{\theta})}$	*Estimert standardfeil* $SE(\hat{\theta})$	*Betingelse for akseptabel normal-tilnærmelse*	*Pivotal* $\frac{\hat{\theta}-\theta}{SE(\hat{\theta})}$	*Konfidensintervall (konfidensgrad tilnærmet* $1-\alpha$ *)* $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$
$X \sim \text{bin}(n,p)$	$\hat{p} = \frac{X}{n}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\text{var}(X) \geq 5$ $(np(1-p) \geq 5$	$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{\text{tilnærmet}}{\sim} N(0,1)$	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$X \sim$ hypergeom. (n, M, N) $(p = M/N)$	$\hat{p} = \frac{X}{n}$	$\sqrt{\frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}}$	$\text{var}(X) \geq 5$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}}} \overset{\text{tilnærmet}}{\sim} N(0,1)$	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}}$
$X \sim \text{pois}(t\lambda)$ [5]	$\hat{\lambda} = \frac{X}{t}$	$\sqrt{\frac{\lambda}{t}}$	$\sqrt{\frac{\hat{\lambda}}{t}}$	$\text{var}(X) \geq 5$ $(t\lambda \geq 5)$	$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/t}} \overset{\text{tilnærmet}}{\sim} N(0,1)$	$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$

[5] Husk at notasjonen $X \sim \text{pois}(m)$ er valgt slik at det som står på m's plass alltid er lik $E(X)$ (som også er lik $\text{var}(X)$ i poisson-fordelingen). Hvis det for eksempel i en oppgave fremgår at $X \sim \text{pois}(3.7)$, følger automatisk at $E(X) = \text{var}(X) = 3.7$. Av modellen i tabellen følger således at $E(X) = \text{var}(X) = t\lambda$ som impliserer at $\hat{\lambda}$ er forventningsrett siden $E(\hat{\lambda}) = E\left(\frac{X}{t}\right) = \frac{1}{t} \cdot E(X) = \frac{1}{t} \cdot t\lambda = \lambda$. Variansen (lik kvadrert standardfeil) blir $\text{var}(\hat{\lambda}) = \text{var}\left(\frac{X}{t}\right) = \frac{1}{t^2}\text{var}(X) = \frac{1}{t^2} \cdot t\lambda = \frac{\lambda}{t}$.

Merknad 7 Merk at de tre KI-ene i tabell 3 samt KI-ene i situasjon 1 og 3 i tabell 1 alle har den generelle formen angitt i regel 6.7 der SE står for standardfeil eller estimert standardfeil dersom $\mathrm{SE}(\hat{\theta})$ avhenger av ukjente parametre. Unntak fra regel 6.7 er gitt i tabell 2 og situasjon 2 under tabell 1. Argumentasjonen fra pivotal-utsagnet til konfidensintervallet er gitt i avsnitt 6.3.1. rett etter regel 6.7.

Merknad 8 I mange KI (jfr tabell 1 og 3) bruker vi altså den estimerte versjonen av standardfeilen (i tilfelle standardfeilen er ukjent) når vi utleder et KI. Det er ikke på noen måte opplagt at vi har lov til dette. Det er rimelig å tenke seg at en slik fremgangsmåte ville kunne ødelegge tilnærmelsen til N(0, 1), noe som ville gjøre konfidensgraden tvilsom. Det at vi ifølge modifisert regel 6.7 **(b)** faktisk "har lov til" å erstatte SE med en estimert versjon uten å berøre konfidensgraden vesentlig, er egentlig ganske overraskende sett i lys av en ofte betydelig usikkerhet i estimeringen av σ. For eksempel for kvinnehøydene i merknad 6 ble konfidensintervallet for σ $\left[4.24,\ 6.80\right]$ som indikerer en ikke ubetydelig usikkerhet Likevel vil etter modifisert regel 6.7 **(b)** konfidensgraden for KI-et for μ ikke bli vesentlig berørt om vi bytter ut σ med $\hat{\sigma} = S$ i standardfeilen.

Merknad 9 Det at vi har formler for usikkerhetsdelen, $\pm c \cdot SE(\hat{\theta})$, i et konfidensintervall for θ der utvalgsstørrelsen *n* inngår, gjør det mulig å bestemme nødvendig størrelse (*n*) på utvalget for å oppnå at usikkerheten ikke overstiger en gitt (akseptabel) grense. Slike beregninger kan være viktige ved planleggingen av en statistisk undersøkelse. Eksempler på slike beregninger er gitt i eksemplene 6.9, 6.11 og 6.14 i Løvås.

4 Regresjonsmodellen

KI-ene for ukjente parametre i den enkle standard regresjonmodellen med normalfordelte restledd følger samme mønsteret som situasjon 2 i tabell 1, med eneste forskjell at $n-2$ frihetsgrader benyttes i *t*-fordelingen istedenfor $n-1$ som i situasjon 2. KI-et har i alle tilfeller formen

$$\hat{\theta} \pm t_{n-2,\alpha/2} SE(\hat{\theta})$$

og konfidensgraden $1-\alpha$ gjelder eksakt for alle $n \geq 3$. Det du trenger i tillegg er derfor bare formler for $\hat{\theta}$ og $SE(\hat{\theta})$, som du finner i Løvås kap. 7, eller regresjon-II –notatet som snart legges ut på nettet. Notatet gir også eksempler på beregning av slike KI-er.