7. Estimering.

7.1. Innledning.

Det fins en mengde eksempler på at vi ønsker å si noe om egenskaper ved en *populasjon* på grunnlag av egenskaper ved et *tilfeldig utvalg* fra populasjonen. Tenk bare på meningsmålinger, der man vil skaffe seg kjennskap til hva "folk" mener ved å spørre et tilfeldig utvalg. Eller tenk på kvalitetskontroll, der vi vil skaffe oss kjennskap til kvaliteten av en stor produksjonsserie ved å teste et tilfeldig utvalg av produktene.

Vi ønsker altså å skaffe oss kjennskap til verdien av en eller annen ***parameter*** i populasjonen. Vanlige parametre er middelverdi, varians, eller hvor stor prosentandel av populasjonen som har en bestemt egenskap. Det naturlige er at vi først bestemmer en "beste verdi" for parameteren ut fra observasjoner av utvalget. Denne "beste verdien" skal vi kalle et ***punktestimat*** for parameteren. Men vi bør også finne et eller annen mål for ***feilmarginen*** for vårt punktestimat. Dette leder oss til begrepet ***konfidensintervall***.

Jeg skal bruke symbolet θ for en generell parameter i populasjonen. Du kan godt tenke deg at θ står for middelverdi, eller for prosentandel som har en bestemt egenskap, eller for noe annet som du er interessert i. Jeg vil bruke symbolet θ når jeg uttrykker meg generelt uten å binde meg til en bestemt parameter.

Merk at θ er en parameter i *populasjonen*. Når vi bestemmer en "beste verdi" for θ ut fra observasjoner av et utvalg, sier vi at vi finner et ***punktestimat*** av θ. Vi bruker $\hat{\theta}$ ("theta-hatt") som symbol for *punktestimatet* av θ (eller bare *estimatet* som vi vanligvis sier).

7.2. Punktestimat.

Vi skal begrense oss til å se på punktestimat for middelverdi, varians og prosentandel. Vi skal hele tiden anta at vi trekker et tilfeldig utvalg på *n* objekter fra populasjonen. Da kan vi vise at:

- Det beste estimatet for middelverdien μ i populasjonen er middelverdien
  $$\overline{x} = \frac{1}{n}\sum x_i$$
  i utvalget, der x_i står for verdiene fra utvalget.
  Eller kortere: $\hat{\mu} = \overline{x}$.

- Det beste estimat for variansen σ^2 i populasjonen er
  $$s^2 = \frac{1}{n-1}\sum\left(x_i - \overline{x}\right)^2 = \frac{1}{n-1}\left(\sum x_i^{\,2} - n\cdot\overline{x}^2\right).$$
  Eller kortere: $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum\left(x_i - \overline{x}\right)^2$.

  Dersom vi skal være pirkete, så er ikke $\sqrt{s^2}$ det aller beste estimatet for *standardavviket* σ . Men i praksis er vi ikke så pirkete, så vi sier at $\hat{\sigma} = \sqrt{s^2}$.

- Det beste estimatet for andelen *p* i populasjonen som har en bestemt egenskap, er
$$\hat{p} = \frac{x}{n}$$
der *x* er antall objekter i utvalget som har egenskapen.

Litt terminologi-pirk: Den formelen eller algoritmen som benyttes, kalles gjerne en ***estimator***. Den tallverdien som framkommer når vi anvender estimatoren, kalles et ***estimat***. I praksis bruker vi gjerne ordet *estimat* både om formelen og om tall-resultatet.

La oss se kort på hva som kreves av et "beste estimat". Vi stiller to krav:

1. Estimatet skal være ***forventningsrett***. Dette betyr at *forventningsverdien til estimatet* $\hat{\theta}$ skal være lik den sanne verdien θ. Mer jordnært (og mindre presist) kan vi si at dersom vi gjentar beregningen av estimatet uendelig mange ganger med stadig nye utvalg, skal gjennomsnittet av våre beregnede verdier av $\hat{\theta}$ være lik θ.

2. Dersom vi har valg mellom flere forventningsrette estimat, skal vi velge det estimatet som har minst varians.

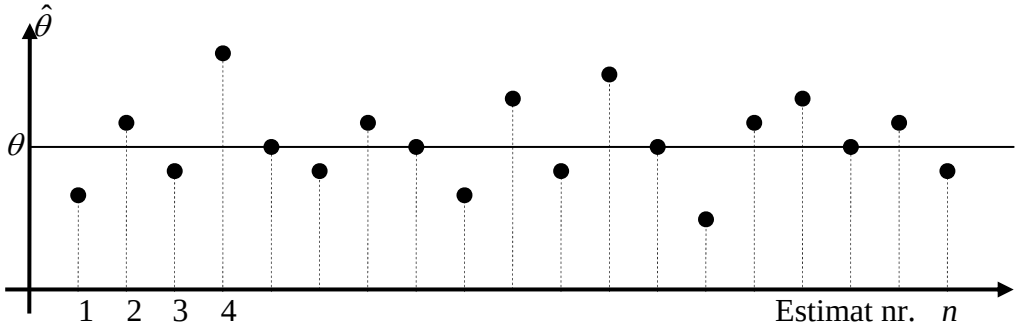
La meg illustrere situasjonen med et eksempel: Anta at du skal måle pH-verdien i et vann. Du har flere måleinstrumenter til disposisjon, men de er ikke like nøyaktige. Hvordan skal du finne det beste estimatet av den *virkelige* pH-verdien i vannet? Skal du bare bruke det mest nøyaktige instrumentet? Eller skal du bruk alle, og beregne gjennomsnittet av målingene? Ved å bruke kriteriene ovenfor, kan vi vise at det beste estimatet er et *veid gjennomsnitt* av målingene med alle instrumentene, der de mest nøyaktige instrumentene gis størst vekt. Vi kan finne vekttallene uttrykt ved variansene ("unøyaktighetene") til de enkelte instrumentene.

7.3. Konfidensintervall.

7.3.1. Prinsipper og regneregler.

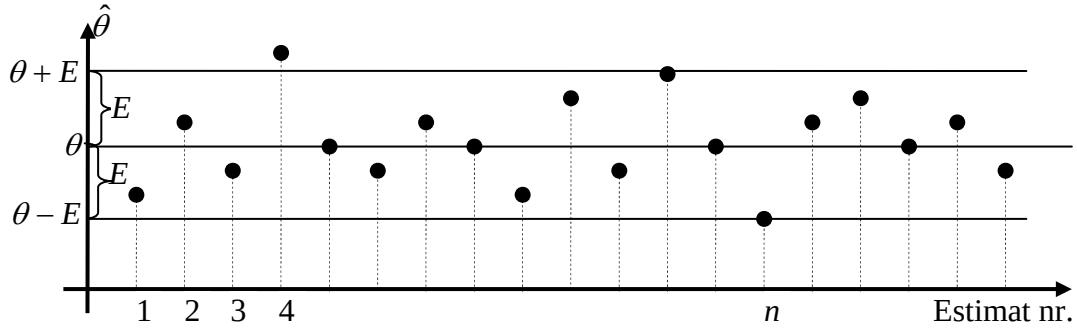
Nå blir det vanskeligere. Ikke minst fordi jeg skal prøve å få fram de generelle prinsippene, slik at beregning av konfidensintervall ikke bare skal bli manipulering med formler. Du vil ha stort utbytte av å forstå prinsippene fordi vi senere i kurset stadig kommer tilbake til samme problemstilling i ulike sammenhenger.

Utgangspunktet er at vi har en populasjon med en parameter θ. Dersom vi beregner *n* innbyrdes uavhengige estimater $\hat{\theta}_1$, $\hat{\theta}_2$, …, $\hat{\theta}_n$ av denne parameteren, vil disse estimatene spre seg rundt den rette verdien θ. Vi kan illustrere situasjonen slik:



Det er viktig å ha klart for seg at populasjons-parameteren θ har en fast verdi, mens *estimatene* $\hat{\theta}$ sprer seg rundt denne verdien.

Nå skal vi lage et intervall rundt θ som er slik at en andel α av estimatene ligger *utenfor* intervallet, og en andel $1-\alpha$ ligger *innenfor*. Videre skal vi anta at en andel $\alpha/2$ ligger utenfor på hver side. Bredden av intervallet skal vi kalle $2E$. Situasjonen er altså noe slikt:



I praksis brukes ofte $\alpha = 0.05$ (eller 5 %). Dette betyr at 95 % av estimatene ligger innenfor intervallet, og 2.5 % ligger utenfor på hver side. Det er også vanlig å bruke $\alpha = 0.10$ eller $\alpha = 0.02$.

Betraktningene ovenfor er vel og bra dersom vi har svært mange (”uendelig” mange) estimater til disposisjon. Det har vi ikke. Vanligvis har vi kun *ett* estimat $\hat{\theta}$ til disposisjon. Hva gjør vi da?

Vi snur problemet. Istedenfor å konstruere et intervall med bredde $2E$ rundt θ, lager vi et intervall med samme bredde $2E$ rundt *estimatet* $\hat{\theta}$. Et slikt intervall kaller vi et ***konfidens-intervall***. Vi kan da vise at:

La $\hat{\theta}$ være et estimat av en populasjonsparameter θ.
En andel $1-\alpha$ av alle ***konfidensintervall*** av formen

$$\left[\hat{\theta} - E,\ \hat{\theta} + E\right]$$

vil inneholde populasjonsparameteren θ.

Størrelsen $1-\alpha$ kalles ***konfidensnivået***, mens E kalles ***feilmarginen***.

Nå gjenstår det ”bare” å beregne feilmarginen E. Først må vi gjøre en viktig forutsetning:

Estimatet $\hat{\theta}$ er normalfordelt med forventningsverdi θ og standardavvik $\sigma_{\hat{\theta}}$.

Vi innfører en størrelse $z_{\alpha/2}$ som er slik at sannsynligheten for at z skal være større enn $z_{\alpha/2}$ i en standard normalfordeling er lik $\alpha/2$.

Eksempel 7.1: Finn $z_{\alpha/2}$ når $\alpha = 0.05$.

Løsning: Når $\alpha = 0.05$, er $\alpha/2 = 0.025$. Vi må da finne den z-verdien i en normalfordelings-tabell som gir et areal på $1 - 0.025 = 0.975$ til venstre for z. Vi finner av tabellen at

$$z_{\alpha/2} = z_{0.025} = \underline{\underline{1.96}}.$$

Nå kan vi vise at:

Feilmarginen E er gitt ved

$$E = z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$$

slik at konfidensintervallet er

$$\left[\hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}},\ \hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}\right].$$

Det praktiske problemet består i å påvise at $\hat{\theta}$ virkelig er normalfordelt, samt finne standard-avviket $\sigma_{\hat{\theta}}$ for estimatet. Disse problemene skal vi ta opp for hver av de situasjonene vi etter hvert kommer opp i.

7.3.2. Utledninger.

Anta at vi har mange estimater av en parameter θ. Vi trekker et *tilfeldig* estimat $\hat{\theta}$. Siden en andel $1 - \alpha$ av estimatene ligger innenfor intervallet, er sannsynligheten for at verdien av vårt estimat skal ligge innenfor intervallet

$$P\left(\theta - E < \hat{\theta} < \theta + E\right) = 1 - \alpha.$$

Vi trekker fra θ på begge sider av de to ulikhetstegnene:

*) $P\left(-E < \hat{\theta} - \theta < E\right) = 1 - \alpha.$ (denne likningen får du bruk for senere)

Multipliserer med -1 (og snur ulikhetstegnene):

$$P\left(E > \theta - \hat{\theta} > -E\right) = 1 - \alpha.$$

Snur "lese-retningen":

$$P\left(-E < \theta - \hat{\theta} < E\right) = 1 - \alpha.$$

Legger til $\hat{\theta}$ på begge sider av de to ulikhetstegnene:

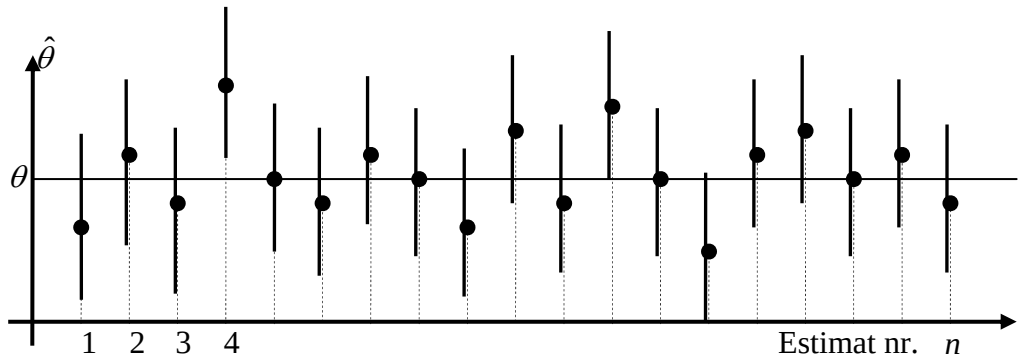
$$P\left(\hat{\theta} - E < \theta < \hat{\theta} + E\right) = 1 - \alpha.$$

Hva har vi oppnådd med disse utregningene? Vi startet med et intervall med bredde $2E$ rundt populasjonsparameteren θ. Vi endte opp med et like bredt intervall av formen

$$\left[\hat{\theta} - E,\ \hat{\theta} + E\right]$$

rundt vårt tilfeldige estimat $\hat{\theta}$.

Utregningene angir at en andel $1-\alpha$ av disse intervallene inneholder populasjons-parameteren θ. Situasjonen er illustrert nedenfor, der intervallene $\left[\hat{\theta}-E,\ \hat{\theta}+E\right]$ er tegnet inn rundt hvert av estimatene.



Nå gjenstår det å finne *E*. Vi skal da forutsette at:

> Estimatet $\hat{\theta}$ er normalfordelt med forventningsverdi θ og standardavvik $\sigma_{\hat{\theta}}$.

La $z_{\alpha/2}$ være den *z*-verdien i en standard normalfordeling som er slik at sannsynligheten for at *z* skal være større enn $z_{\alpha/2}$ er $\alpha/2$. Fordi normalfordelingen er symmetrisk kan vi da sette opp:

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1-\alpha\,.$$

Etter forutsetningen er $\hat{\theta}$ normalfordelt med forventningsverdi θ og standardavvik $\sigma_{\hat{\theta}}$. Da er

$$z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$$

standard normalfordelt, slik at

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) = 1-\alpha\,.$$

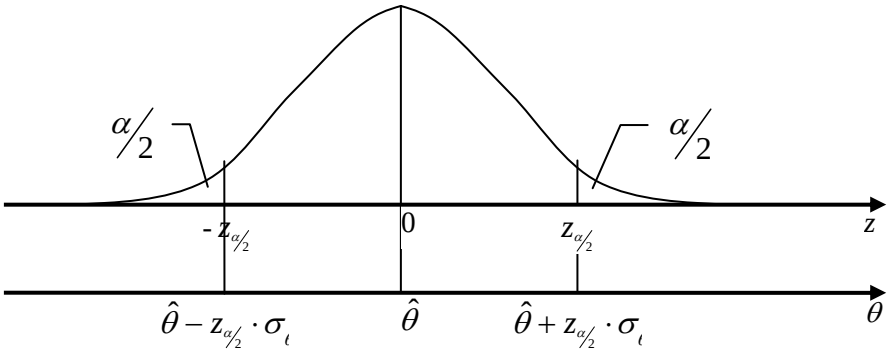
Multipliserer med $\sigma_{\hat{\theta}}$ på begge sider av ulikhetene, og får

$$P\left(-z_{\alpha/2}\cdot\sigma_{\hat{\theta}} < \hat{\theta}-\theta < z_{\alpha/2}\cdot\sigma_{\hat{\theta}}\right) = 1-\alpha\,.$$

Men hvis du sammenlikner denne likningen med *) fra forrige side, ser du at $E = z_{\alpha/2}\cdot\sigma_{\hat{\theta}}$.
Hvis vi erstatter *E* med $z_{\alpha/2}\cdot\sigma_{\hat{\theta}}$ i de videre regningene på forrige side, vil vi ende opp med

$$P\left(\hat{\theta}-z_{\alpha/2}\cdot\sigma_{\hat{\theta}} < \theta < \hat{\theta}+z_{\alpha/2}\cdot\sigma_{\hat{\theta}}\right) = 1-\alpha\,.$$

Situasjonen er illustrert på neste side.



Vi oppsummerer:

La $\hat{\theta}$ være et estimat av en populasjonsparameter θ. Dersom estimatet $\hat{\theta}$ er normalfordelt med forventningsverdi θ og standardavvik $\sigma_{\hat{\theta}}$, vil en andel $1-\alpha$ av alle ***konfidensintervall*** av formen

$$\left[\hat{\theta}-E,\ \hat{\theta}+E\right]$$

der

$$E = z_{\alpha/2}\cdot\sigma_{\hat{\theta}}$$

inneholde populasjonsparameteren θ.

7.4. Konfidensintervall for prosentandel.

Anta at en andel *p* i en populasjon har en bestemt egenskap. Vi har et tilfeldig utvalg på *n* objekter fra populasjonen. Av disse *n* objektene har *x* den bestemte egenskapen. Beste estimat for *p* er (som før nevnt) $\hat{p} = \frac{x}{n}$.

Men *x* vil være binomisk fordelt. Vi har tidligere sagt at dersom *n* er stor, vil den binomiske fordelingen nærme seg normalfordeling med forventning

$$\mathrm{E}(x) = n\cdot p$$

og varians

$$\mathrm{Var}(x) = n\cdot p\cdot(1-p).$$

At "*n* er stor" innebærer (ifølge vår bok) at $n\cdot p \geq 5$ og at $n\cdot(1-p)\geq 5$. Dessuten bør vi (om nødvendig) bruke kontinuitetskorreksjon.

Den oppmerksomme leser vil innvende at disse formlene inneholder *p*. Kan vi bruke dem når vi ikke kjenner *p*, men bare kjenner et *estimat* av *p*? Svaret er at den feilen vi gjør ved å bruke estimatet istedenfor den korrekte verdien er så liten at den ikke spiller avgjørende rolle.

Vi trekker altså *n* objekter, hvorav *x* har en bestemt egenskap. Beste estimat for andelen *p* i populasjonen som har egenskapen er

$$\hat{p} = \frac{x}{n}.$$

Da er:

$$\mathrm{Var}\left(\hat{p}\right)=\mathrm{Var}\left(\frac{x}{n}\right)=\frac{1}{n^2}\mathrm{Var}\left(X\right)=\frac{1}{n^2}\cdot n\cdot p\cdot\left(1-p\right)=\frac{p\left(1-p\right)}{n}$$

slik at standardavviket blir

$$\sigma_{\hat{p}}=\sqrt{\frac{p\left(1-p\right)}{n}}\,.$$

I praksis erstatter vi altså p med $\hat{p}$, og får at standardavviket til estimatet av p blir

$$\sigma_{\hat{p}}=\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}\,.$$

Vi summerer opp:

La p være prosentandelen som har en bestemt egenskap i en populasjon. La x være antall som har denne egenskapen når vi trekker n objekter fra populasjonen. Et konfidensintervall med konfidensnivå $1-\alpha$ for prosentandelen p er av formen

$$\left[\hat{p}-E,\ \hat{p}+E\right]$$

der

$$\hat{p}=\frac{x}{n}$$

og

$$E=z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Da har vi alt vi trenger for å gå i gang med et eksempel.

Eksempel 7.2: Ei avis spør noen tilfeldige personer om de synes at ordføreren bør gjenvelges for en ny periode. Av de $n=50$ som har noen mening, svarer $x=30$ "ja". Sett opp et 95 % konfidensintervall for andelen av velgere som vil at ordføreren bør gjenvelges.

Løsning: Vi starter med å sjekke forutsetningene: Beste estimat for p er

$$\hat{p}=\frac{x}{n}=\frac{30}{50}=0.60\,.$$

Da er $n\cdot\hat{p}=50\cdot0.60=30$ og $n\cdot\left(1-\hat{p}\right)=50\cdot\left(1-0.60\right)=20$,

slik at kravene for normalfordeling er oppfylt.
Videre er standardavviket for $\hat{p}$ gitt ved

$$\sigma_{\hat{p}}=\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}=\sqrt{\frac{0.60\cdot0.40}{50}}=\underline{0.0693}\,,$$

slik at

$$E=z_{0.05/2}\cdot\sigma_{\hat{p}}=1.96\cdot0.0693=\underline{0.136}\,.$$

Et 95 % konfidensintervall for andelen av velgere som vil at ordføreren bør gjenvelges blir da

$$\left[\hat{p}-E,\ \hat{p}+E\right]=\left[0.60-0.136,\ 0.60-0.136\right]=\underline{\underline{\left[0.464,\ 0.736\right]}}\,.$$

Vi ser av eksemplet ovenfor at feilmarginen kan bli temmelig stor. Når vi planlegger slike undersøkelser, kan det være lurt å sette en øvre grense $E_{\max}$ for tillatt feilmargin E. Da får vi:

$$E \le E_{\max} \quad \Leftrightarrow \quad z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \le E_{\max} \quad \Leftrightarrow \quad \frac{p(1-p)}{n} \le \left(\frac{E_{\max}}{z_{\alpha/2}} \right)^2$$

$$\Leftrightarrow \quad n \ge \left(\frac{z_{\alpha/2}}{E_{\max}} \right)^2 \cdot p(1-p)$$

På ny råker vi bort i problemet med at p ikke er kjent. Ettersom vi sjelden er interessert i en svært nøyaktig verdi for n, kan vi bruke en "intelligent gjetning" av p. Eller (hvis vi er forsiktig), kan vi bruke $p = 0.50$ som gir den største verdien av n. Vi har altså:

Dersom vi skal beregne et konfidensintervall med konfidensnivå $1-\alpha$ for prosentandel p i en populasjon, og kreve at feilmarginen E skal være mindre enn $E_{\max}$, må utvalget minst bestå av

$$n = \left(\frac{z_{\alpha/2}}{E_{\max}} \right)^2 \cdot p(1-p)$$

objekter. Dersom vi ikke vet noe om p, bruker vi $p = 0.50$ som gir størst verdi av n.

Eksempel 7.3: Hvor stort må utvalget minst være for at et 95 % konfidensintervall for p har $E_{\max} = 0.025$ (som gir at bredden av konfidensintervallet er 0.050)?

Løsning: Dersom vi på forhånd ikke vet noe om hvilken verdi av p vi kan forvente, blir

$$n \ge \left(\frac{z_{0.05/2}}{E_{\max}} \right)^2 \cdot p(1-p) = \left(\frac{1.96}{0.025} \right)^2 \cdot 0.50 \cdot (1-0.50) \approx \underline{\underline{1537}}.$$

7.5. Konfidensintervall for μ.

Vi skal nå gå over til å se på konfidensintervall for middelverdien μ i en populasjon. Som før skal vi plukke ut et tilfeldig utvalg på n objekter, og benytte at beste punktestimat for μ er middelverdien $\bar{x} = \frac{1}{n}\sum x_i$ for utvalget. Videre vet vi at dersom standardavviket σ i populasjonen er kjent, så er standardavviket for $\bar{x}$ gitt ved

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Dette gjør det enkelt å finne konfidensintervall for μ dersom σ er kjent og $\bar{x}$ er normalfordelt:

Dersom $\bar{x}$ er normalfordelt, er

$$\left[\bar{x} - E,\ \bar{x} + E\right]$$

der

$$E = z_{\alpha/2} \cdot \sigma_{\bar{x}} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

et konfidensintervall for μ med konfidensnivå $1 - \alpha$.

Kravet om at $\bar{x}$ er normalfordelt er oppfylt dersom:

- Populasjonen er normalfordelt, eller
- $n > 30$ (jfr. Sentralgrenseteoremet).

I praksis gjør vi ikke noe veldig galt dersom kravet om normalfordeling ikke er helt oppfylt. Vi får akseptable resultater også når $\bar{x}$ er noenlunde normalfordelt. Dette medfører at vi ofte ser at kravene ovenfor slakkes litt.

Det store praktiske problemet er at setningen ovenfor forutsetter at σ er kjent. Dette er så å si aldri oppfylt. Men før vi ser hvordan vi hanskes med problemet med ukjent σ, skal vi se på et eksempel:

Eksempel 7.4: Du vil bestemme et 90 % konfidensintervall for gjennomsnittsvekten av oppdrettsfisk i en mære. Du tar da opp 25 tilfeldig valgte fisker, og finner at gjennomsnittsvekten av disse 25 fiskene er 1.836 kg. Av tidligere erfaring vet du at standardavviket for vekten av fiskene i mæra er $\sigma = 0.16$ kg, og du tar sjansen på at denne verdien fremdeles er brukbar. La μ være gjennomsnittsvekten av *alle* fiskene i mæra (hele populasjonen). Finn et 90 % konfidensintervall for μ.

Løsning: Et 90 % konfidensintervall for μ blir da

$$\left[\bar{x} - E,\ \bar{x} + E\right]$$

der

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = z_{0.05} \cdot \frac{0.16}{\sqrt{25}} = 1.645 \cdot \frac{0.16}{5} = \underline{0.0526}.$$

Et 90 % konfidensintervall for μ blir altså

$$\left[\bar{x} - E,\ \bar{x} + E\right] = \left[1.836 - 0.0526,\ 1.836 + 0.0526\right] = \underline{\underline{\left[1.783,\ 1.889\right]}}.$$

Den store bøygen i slike problem er at standardavviket σ som regel ikke er kjent. Kan vi ikke da benytte estimatet

$$s = \sqrt{\frac{1}{n-1}\sum\left(x - \bar{x}\right)^2}$$

istedenfor σ? Innvendingen mot dette er at vi innfører en ekstra usikkerhet fordi vi opererer med et *estimat* av σ istedenfor med en sikker verdi av σ. Denne ekstra usikkerheten bør gi seg utslag i en større feilmargin E. Men hvordan skal vi finne denne nye verdien av E?

Løsningen ligger i at vi erstatter standard normalfordelingen med en ny fordeling som kalles ***t-fordelingen***. Den er svært lik standard normalfordeling, men er litt bredere. Dette fører til at feilmarginen *E* blir litt større. Ellers regner vi akkurat som før, men bruker symbolet $t_{\alpha/2}$ istedenfor $z_{\alpha/2}$ for å markere at vi bruker en *t*-fordeling.

Men det er en liten detalj til: formen på *t*-fordelingen avhenger av *n*! Vi innfører begrepet ***antall frihetsgrader*** for å angi hvilken variant av *t*-fordelingen vi skal bruke. For vårt formål er antall frihetsgrader lik $n-1$. I tabeller finner du gjerne verdier for $t_{\alpha/2}$ for ulike verdier av α der antall frihetsgrader øker fra 1 og oppover. Merk at jo større *n* blir, jo mer nærmer *t*-fordelingen seg standard normalfordelingen.

La oss se hvordan dette fungerer i praksis.

Eksempel 7.5: Vi vender igjen tilbake til fiskene i oppdrettsanlegget, der gjennomsnittsvekten av 25 tilfeldig utvalgte fisker var 1.836 kg. Men nå vet du ikke standardavviket σ for fiskene i mæra. Derimot *beregner* du standardavviket *s* for fiskene i *utvalget* ditt, og finner $s = 0.16$ kg. Finn et 90 % konfidensintervall for μ.

Løsning: Et 90 % konfidensintervall for μ blir fremdeles på formen

$$\left[\bar{x} - E,\ \bar{x} + E\right].$$

Men nå er

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = t_{0.05} \cdot \frac{0.16}{\sqrt{25}} = 1.711 \cdot \frac{0.16}{5} = \underline{0.0548}.$$

der verdien for $t_{\alpha/2}$ er funnet av tabell med $25-1=24$ frihetsgrader.

Et 90 % konfidensintervall for μ blir nå

$$\left[\bar{x} - E,\ \bar{x} + E\right] = \left[1.836 - 0.0548,\ 1.836 + 0.0548\right] = \underline{\underline{\left[1.781,\ 1.891\right]}}.$$

Du får altså et intervall som er *litt* bredere enn før, men forandringen er ikke stor. Men dersom *n* hadde vært mindre, ville bredden økt mer.

Dersom du skal finne verdier av $t_{\alpha/2}$ som ikke står i tabellen, får du interpolere på fornuftig måte. En bedre løsning er å bruke dataverktøy. Da kan du finne alle tenkelige verdier av $t_{\alpha/2}$.

7.6. Konfidensintervall for populasjonsvariansen.

Vi har allerede nevnt at beste punktestimat for populasjonsvariansen σ^2 er

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2,$$

og at

$$s = \sqrt{\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2}$$

er et brukbart estimat for populasjonens standardavvik σ.

Men du vil få ulike verdier for disse estimatene for ulike utvalg. Hvordan kan du finne et konfidensintervall for *populasjonsvariansen* (eller for *populasjonens standardavvik*)?

Her dukker det opp et nytt problem: *Estimatene av σ^2 og av σ er ikke normalfordelt!!* Dermed er hele den fine teorien i starten av dette notatet (nesten) ubrukelig.

Redningen kommer i form av en annen fordeling som kalles χ^2*-fordelingen* (uttales "kji-kvadrat-fordelingen"). Ikke la deg forvirre av den rare skrivemåten χ^2. Tenk deg at det står *x* eller noe annet velkjent.

Denne fordelingen er kun definert for positive verdier av χ^2. Den kan minne litt om normalfordelingen av utseende, men den er ikke symmetrisk. Dette betyr at vi må handtere de to "halene" av fordelingen hver for seg. Fordelingen fins også i ulike versjoner avhengig av *antall frihetsgrader*, på samme måten som *t*-fordelingen.

Så til saken. Vi kan vise at:

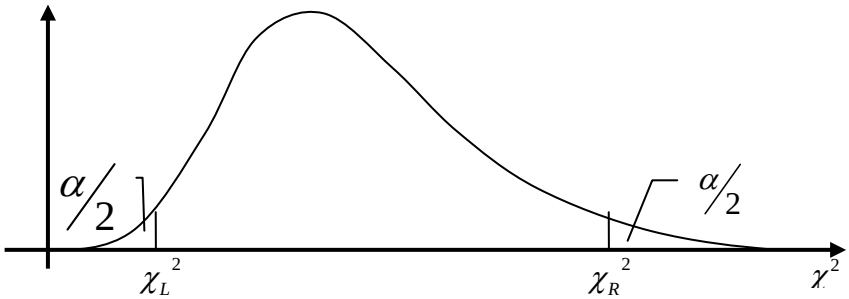
Dersom s^2 er variansen fra et utvalg av *n* objekter som er tatt tilfeldig fra en *normalfordelt populasjon* med varians σ^2, så er

$$\frac{(n-1)\cdot s^2}{\sigma^2}$$

χ^2-fordelt med $n-1$ frihetsgrader.

Merk at *populasjonen* må være normalfordelt. Dette kravet *må* være oppfylt, ellers kan du gjøre store feil.

La oss gå i gang med å konstruere et $1-\alpha$ konfidensintervall for populasjonsvariansen σ^2. Selv om detaljene i teorien fra starten av notatet ikke kan brukes, kan vi benytte resonnementet. Ved hjelp av tabell finner vi to verdier ${\chi_L}^2$ og ${\chi_R}^2$ som er slik at en andel $\alpha/2$ ligger til venstre for ${\chi_L}^2$, og en andel $\alpha/2$ ligger til høyre for ${\chi_R}^2$. Da er



$$P\left({\chi_L}^2 < \frac{(n-1)s^2}{\sigma^2} < {\chi_R}^2\right) = 1-\alpha\,.$$

Omforming:

$$P\left(\frac{\chi_L^{\,2}}{(n-1)s^2} < \frac{1}{\sigma^2} < \frac{\chi_R^{\,2}}{(n-1)s^2}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)s^2}{\chi_R^{\,2}} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^{\,2}}\right) = 1-\alpha$$

Dermed har vi at:

Konfidensintervallet for populasjonsvariansen σ^2 er

$$\left[\frac{(n-1)s^2}{\chi_R^{\,2}}, \frac{(n-1)s^2}{\chi_L^{\,2}}\right]$$

der $\chi_L^{\,2}$ og $\chi_R^{\,2}$ er hentet fra en χ^2-tabell med $n-1$ frihetsgrader.

Eksempel 7.6: Vi vender igjen tilbake til vårt oppdrettsanlegg, der vi har funnet at vekten av $n = 25$ tilfeldige fisker fra en normalfordelt populasjon har standardavvik $s = 0.16$ kg, slik at variansen blir

$$s^2 = 0.16^2 = \underline{0.0256}.$$

Finn et 90 % konfidensintervall for populasjonsvariansen σ^2.

Løsning: Av χ^2-tabell for $n-1 = 25-1 = 24$ frihetsgrader finner vi at

$$\chi_L^{\,2} = 13.848$$

mens

$$\chi_R^{\,2} = 36.415$$

når $\alpha/2 = 0.10/2 = 0.05$.

Konfidensintervallet for σ^2 blir da

$$\left[\frac{(n-1)s^2}{\chi_R^{\,2}}, \frac{(n-1)s^2}{\chi_L^{\,2}}\right] = \left[\frac{24\cdot 0.0256}{36.415}, \frac{24\cdot 0.0256}{13.848}\right] = \underline{\underline{[0.01687,\ 0.04437]}}.$$

Ved å trekke kvadratrota finner vi et tilnærmet 90 % konfidensintervall for populasjonens standardavvik σ:

$$\left[\sqrt{0.01687}, \sqrt{0.04437}\right] = \underline{\underline{[0.13,\ 0.21]}}.$$