HG **Econ2130** **2017**

9/3-2017

Supplement til forelesning uke 10 (6. mars).

(Det jeg ikke rakk å ta på forelesning.)

Terminologi (estimering)

- Data (konkrete tall), $x_1, x_2, \ldots$ er observasjoner av stokastiske variable, $X_1, X_2, \ldots$ Den statistiske modellen uttrykker det vi mener er rimelig å forutsette om sannsynlighetsfordelingene for $X_1, X_2, \ldots$ som gjerne avhenger av ukjente (populasjons-) størrelser, $\mu, \sigma, \lambda, \ldots$ som vi ønsker å si noe om basert på data $x_1, x_2, \ldots$
- **Eksempel.** Uid-modellen: $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte med $E(X_i) = \mu$ og $\text{var}(X_i) = \sigma^2$. **Regneeksempel** (gitt på forelesningen): X_i = diameteren 1 meter over bakken for et rent tilfeldig tre nr *i* trukket fra en skog. $E(X_i) = \mu$, som representerer gjennomsnittsdiameteren i hele skogen, antas ukjent og skal estimeres basert på $n = 3$ uavhengige observasjoner, $x_1 = 68$, $x_2 = 82$ og $x_3 = 44$.
- **Estimatet** $\bar{x} = 64.7$ er en observasjon av en stokastisk variabel, $\hat{\mu} = \bar{X}$, som kalles **estimator**. Den observerte verdien av $\hat{\mu}$ indikeres av og til ved indeksen *obs* ($\hat{\mu}_{obs} = \bar{X}_{obs} = \bar{x} = 64.7$).
- En estimator er en **observerbar** stokastisk variabel (også kalt «observator»). At den er observerbar betyr at den ikke avhenger av noen ukjente størrelser (ukjente parametre) slik at vi kan regne ut en verdi (observasjon) av den ut fra data.

  F. eks. $\bar{X}$ er observerbar mens den standardiserte $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ er ikke observerbar.

  Selv om vi ikke kan observere *Z*, har *Z* en sannsynlighets-fordeling (tilnærmet $N(0,1)$).

- Alternative estimatorer for μ betegnes ofte med aksenter $\hat{\mu}, \tilde{\mu}, \mu^*, \mu', \hat{\hat{\mu}}$, osv.
- **Definisjon.** Hvis θ (teta) er en parameter og $\hat{\theta}$ en estimator, sies $\hat{\theta}$ å være **forventningsrett** *hvis* $E(\hat{\theta}) = \theta$.

  F.eks.,vi har vist før at i *uid*-modellen gjelder:
  $E(\bar{X}) = \mu$, $\text{var}(\bar{X}) = \sigma^2/n$.
  Dermed er $\hat{\mu} = \bar{X}$ forventningsrett, $\text{var}(\hat{\mu}) = \sigma^2/n$, og $\hat{\mu}$ er tilnærmet normalfordelt for "stor" *n* ($n \geq 20$) pga sentralgrenseteoremet.

- Variansen til en forventningsrett estimator er et uttrykk for estimatorens presisjon – dvs. *desto mindre varians, desto større presisjon*.
- For to alternative forventningsrette estimatorer, velg den som har minst varians.

Estimering av populasjonsvariansen, $\text{var}(X) = \sigma^2$

Vanlig estimator for σ^2 i uid-modellen er utvalgsvariansen

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

som kan vises (bevist i boka under regel **6.3**) er forventningsrett, dvs. $E(\hat{\sigma}^2) = \sigma^2$.

I eksemplet er $n = 3$, og estimatet for σ^2 blir

$$\hat{\sigma}^2_{obs} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big|_{obs} = \frac{1}{2}\left[(68-64.7)^2 + (82-64.7)^2 + (44-64.7)^2\right] = 369.335$$

Vanlig estimator for populasjons-standardavviket $\sigma = \sqrt{\text{var}(X)}$

er, rett og slett, $\hat{\sigma} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, (utvalgs-standardavviket)

I eksemplet blir estimatet: $\hat{\sigma}_{obs} = \sqrt{\hat{\sigma}^2_{obs}} = \sqrt{369.335} = 19.2181$

Merk at estimatoren $\hat{\sigma}$ for populasjons-standardavviket **ikke er forventningsrett** !

Bevis: (vi ser bort fra muligheten at $\text{var}(\hat{\sigma}) = 0$ som $\Rightarrow$ X er en konstant.)

Siden $E(\hat{\sigma}^2) = \sigma^2$, har vi $0 < \text{var}(\hat{\sigma}) = E\left(\hat{\sigma}^2\right) - \left(E(\hat{\sigma})\right)^2 = \sigma^2 - \left(E(\hat{\sigma})\right)^2 \Rightarrow$

$\Rightarrow \left(E(\hat{\sigma})\right)^2 < \sigma^2 \Rightarrow E(\hat{\sigma}) < \sigma$

Bevis slutt.

Så, $\hat{\sigma}$ er **forventningsskjev** – den har en tendens til å underestimere σ litt, men i praksis, hvis ikke *n* er svært liten, regnes denne skjevheten for neglisjerbar.

Når n øker, vil skjevheten gå mot 0 (kan vises).

Begrepet standardfeil ("standard error" (SE) på engelsk)

La θ (teta) være en ukjent parameter og $\hat{\theta}$ en estimator

($\Rightarrow \hat{\theta}$ er en observerbar stokastisk variabel – også kalt

en **observator** (**"statistic"** på engelsk))

Estimeringsfeil	$\hat{\theta} - \theta$	(stokastisk variabel)
Kvadrert estimeringsfeil	$\left(\hat{\theta} - \theta\right)^2$	(stokastisk variabel)
Forventet kvadrert estimeringsfeil	$E\left[(\hat{\theta} - \theta)^2\right]$	(tall)

Definisjon. Standardfeil for $\hat{\theta}$ $\overset{\text{def}}{=} \sqrt{E\left[(\hat{\theta} - \theta)^2\right]} = SE(\hat{\theta})$

(Fra før: standardavviket til $\hat{\theta}$ $\overset{\text{def}}{=}$ $\sqrt{\text{var}(\hat{\theta})} = \sqrt{E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]}$)

Konklusjon: Hvis $\hat{\theta}$ er forventningsrett ($E(\hat{\theta}) = \theta$), er standardfeilen til $\hat{\theta}$ det samme som standardavviket til $\hat{\theta}$ (dvs $SE(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})} = SD(\hat{\theta})$)

Eksempel (Oppsummering)

La $x_1, x_2, \ldots, x_n$ være *n* uavhengige observasjoner av *X* med populasjonsfordeling $f(x)$, $E(X) = \mu$ og $\text{var}(X) = \sigma^2$.

Statistisk modell. $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte (uid) med felles fordeling lik den for *X* ($f(x)$), $E(X_i) = \mu$ og $\text{var}(X) = \sigma^2$, og $x_i = \left(X_i\right)_{obs}$ for $i = 1, 2, \ldots, n$

Da er $\hat{\mu} = \bar{X}$ og $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ begge forventningsrette estimatorer.

Standardfeilen til $\hat{\mu}$ er $SE(\hat{\mu}) = \sqrt{\text{var}(\hat{\mu})} = \sqrt{\text{var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, som vanligvis er ukjent og estimeres ved $SE(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}}$.

Dette gjenspeiles, f.eks., i «descriptives output») i Excel:

«Descriptives output (Excel)» for regneeksemplet

DATA (x)					
68		Mean	64.66667	<-	$\hat{\mu}_{obs} = \bar{x} = \bar{X}_{obs}$
82		Standard Error	11.09554	<-	$\mathrm{SE}(\hat{\mu})$ (estimert)
44		Median	68		
		Mode	#N/A		
		Standard Deviation	19.21805	<-	$\hat{\sigma}_{obs} = S_{obs}$
		Sample Variance	369.3333	<-	$\hat{\sigma}^2_{obs} = S^2_{obs}$
		Kurtosis	#DIV/0!		Ikke pensum
		Skewness	-0.75704		Ikke pensum
		Range	38	<-	$\max x_i - \min x_i$
		Minimum	44		
		Maximum	82		
		Sum	194		
		Count	3		

(Merk at i dette tilfellet blir «standard error» = «standard deviation» delt på roten av n=3))

Oppsummering av noen vanlig estimatorer for tre diskrete modeller i kapittel 5
(Eksempler gitt i avsnitt 6.2 og oppgaver.)

Binomisk modell. $X \sim \text{bin}(n, p)$ der $p = P(S)$ antas ukjent og n er antall binomiske forsøk.

$\Rightarrow E(X) = np$ og $\text{var}(X) = np(1-p)$

$\hat{p} = \frac{X}{n}$ (relativ frekvens av S-er i n forsøk) er forventningsrett for p.

$$(E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p)$$

med varians $\text{var}(\hat{p}) = \frac{p(1-p)}{n}$ og standardavvik (= standard feil), $\sqrt{\frac{p(1-p)}{n}}$

$$(\text{var}(\hat{p}) = \text{var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n})$$

Poisson prosess. $X \sim pois(\lambda t)$ (X er antall hendelser i løpet av t tidsenheter), der λ (forventet insisdensrate pr. tidsenhet) antas ukjent. (Modellen $\Rightarrow E(X) = \text{var}(X) = \lambda t$.)

$\hat{\lambda} = \frac{X}{t}$ (observert insidensrate pr. tidsenhet) er forventningsrett for λ .

$$(E(\hat{\lambda}) = E\left(\frac{X}{t}\right) = \frac{1}{t}E(X) = \frac{1}{t}t\lambda = \lambda \ )$$

med varians $\text{var}(\hat{\lambda}) = \text{var}\left(\frac{X}{t}\right) = \frac{1}{t^2}\text{var}(X) = \frac{1}{t^2}t\lambda = \frac{\lambda}{t}$, og standardavvik

(= standardfeil), $\sqrt{\frac{\lambda}{t}} = \frac{\sqrt{\lambda}}{\sqrt{t}}$.

Hypergeometrisk modell. $X \sim \text{hypergeom}(N, M, n)$**,** der X er antall enheter med et kjennetegn A i et rent tilfeldig utvalg på n enheter trukket fra en populasjon bestående av i alt N enheter hvorav M (ukjent) enheter har kjennetegnet A. Relativ andel av A-er i populasjonen (som vanligvis ønskes å anslås) er $p = \frac{M}{N}$.

(Modellen $\Rightarrow E(X) = n\frac{M}{N}$ og $\text{var}(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}$ (ikke vist, men oppgitt i pensum))

$\hat{p} = \frac{X}{n}$ (relativ frekvens av A-er i utvalget) er forventningsrett for $p = \frac{M}{N}$.

$$(E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}n\frac{M}{N} = \frac{M}{N} = p \ )$$

med varians, $\text{var}(\hat{p}) = \text{var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{var}(X) = \frac{1}{n^2}np(1-p)\frac{N-n}{N-1} = \frac{p(1-p)}{n}\frac{N-n}{N-1}$

og standardavvik (=standardfeil) $\sqrt{\frac{p(1-p)}{n}\frac{N-n}{N-1}}$.

Merk at hvis vi ønsker å estimere M, vil $\hat{M} = N\hat{p}$ være forventningsrett (hvorfor?). Hva blir standardfeilen for $\hat{M}$?