(B) Statistikk

Del A: Sannsynlighetsteori
Her var modellen og dens parametre kjent, og vi regnet basert på dette.

Eks: Binomisk → parameter p
Normalf. " μ, σ

I **Del B: Statistikk** er parametrene for modellen typisk ikke kjent. Vi skal:

i) anslå (estimere) parameterne
ii) ta stilling til påstander om parameterne

Vi skal også se på **enkel lineær regresjon**, for å studere sammenhengen mellom to variabler.

Populasjon og utvalg

Populasjon:
Alle individer/objekter vi er interessert i

Utvalg:
En delmengde av populasjonen, som vi innhenter data fra.

Viktige egenskaper:
- tilfeldig utvalg
- stort utvalg

} Når utvalget er tilfeldig og stort, kan vi trekke mer presise konklusjoner om hele populasjonen

Tilfeldig:
Alle individ/objekt i populasjonen har samme sannsynlighet for å bli trukket ut

Data-analyse: én variabel

Utvalgsstørrelse: n
Variabel: X

Datasett:

i	x_i
1	x_1
2	x_2
3	x_3
⋮	⋮
n	x_n

gjennomsnitt:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

(utvalgsgjennomsnitt)

median:

når observasjonene ordnes i stigende rekkefølge, er observasjon $(n+1)/2$ om n er oddetall, og gjennomsnittet av observasjon $n/2$, $n/2+1$ om n er partall

Standardavvik:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ (utvalgs std. avvik)

($s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ kalles utvalgs-kovarians)

Kvartiler: finnes ved å ordne observasjonene i stigende rekkefølge

øvre kvartil:
observasjon $3/4 \cdot (n+1)$

nedre kvartil:
observasjon $1/4 \cdot (n+1)$

kvartilbredde
øvre kvartil − nedre kvartil

Tolkning:
gjennomsnitt / median: sentralmål
std. avvik / kvartilbredde: spredningsmål

Beregning:

$\bar{x}$, s_x kan finnes ved hjelp av kalkulator

Eks:

7
13
9
12
14

datasett

Legges inn på kalkulator:

1. C STAT — tømmer Statistikkminne
2. 7 Σ+, 13 Σ+, 9 Σ+, 12 Σ+, 14 Σ+
3. Bruker Statistikkfn.
   - $\bar{x}$ → $\bar{x} = 11$
   - s_x → $s_x = 2.915$

median / kvartiler: finnes ved å ordne observasjonene i stigende rekkefølge (ikke mulig på kalkulator)

Grafisk framstilling:

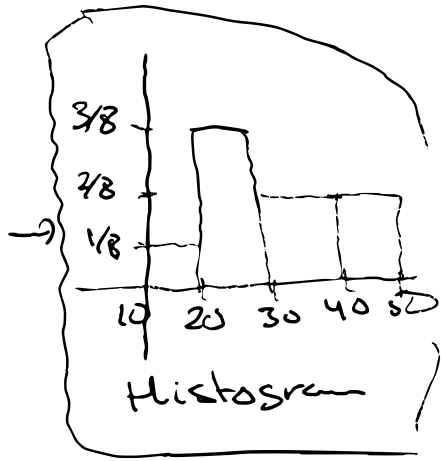
- histogram
- boksplott

} bør vite hvordan man lager disse

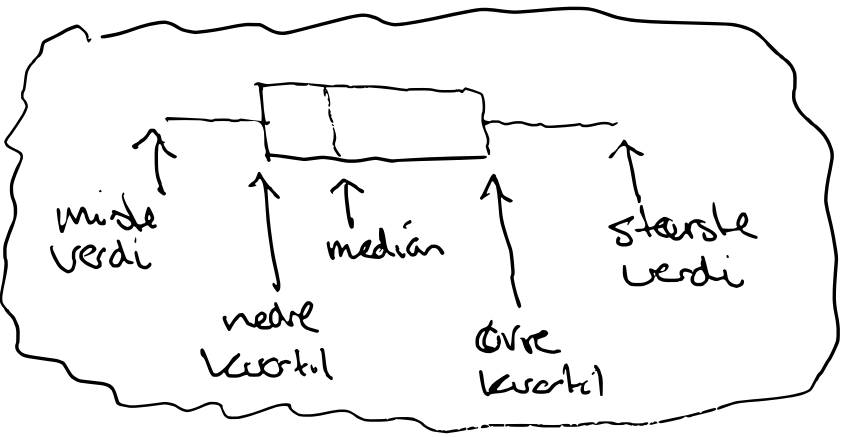
Histogram:

Frekvenstabell

	antall	frekvens
10-20	1	1/8
20-30	3	3/8
30-40	2	2/8
40-50	2	2/8
	8	



Boksplott:



Trekning fra en fordeling

Når vi skal bruke teori for å si noe populasjonen ("alle") basert på et datasett fra et utvalg (noen "få" som er trukket ut), tenker vi ofte slik:

i) Vi velger en modell for variabelen vi ser på, for eksempel at X er binomisk fordelt eller normalfordelt med ukjente parametre.

ii) Dersom utvalgsstørrelsen er n, tenker vi at $X_1, X_2, \ldots, X_n$ er stokastiske variabler som følger samme fordeling som X. Om $X \sim N(\mu, \sigma)$ er normalfordelt, sier vi at $X_1, \ldots, X_n$ er trukket fra en normalfordeling med forventing μ og std. avvik σ.

iii) Vi skriver $x_1, x_2, \ldots, x_n$ for talleverdiene i datasettet (merk: små bokstaver) og kaller dette de observerte verdiene. Vi tenker at x_i er en av de mulige verdiene X_i kan ha (med sannsynligheter som følger fra fordelingen).

iv) Dersom utvalget er tilfeldig, sier vi at $X_1, \ldots, X_n$ er uavhengige trekninger fra fordelingen til X. De er de stokastiske variablene $X_1, \ldots, X_n$ uavhengige.

Estimatorer: Stokastisk variabel som brukes til å estimere en parameter

Navn: estimator for en parameter θ kalles $\hat{\theta}$, ner noen estimatorer har egne navn

μ: Parameteren μ har

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

som estimator. Det er vanlig å skrive $\bar{X}$, ikke $\hat{\mu}$.

σ^2: Parameteren σ^2 har

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

som estimator. Det er vanlig å skrive S^2, ikke $\hat{\sigma}^2$.

P: Parameteren p har

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

som estimater når X er binomisk med parametre (n,p).

Krav til estimatorer:

i) Forventningsrett: $E(\hat{\theta}) = \theta$

ii) $\mathrm{Var}(\hat{\theta})$ minst mulig

iii) $\mathrm{Var}(\hat{\theta}) \to 0$ når $n \to \infty$

Dette er oppfyllt for $\bar{X}$, S^2, $\hat{p}$.

Noen viktige formler:

Når $X_1, \ldots, X_n$ er uavhengige trekninger fra $N(\mu,\sigma)$, så har vi:

$$E(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \sigma^2/n$$

Hvis X er binomisk fordelt med parametre (n, p), så er

$$E(\hat{p}) = p \quad , \quad Var(\hat{p}) = \frac{pq}{n}$$

med $\hat{p} = x/n$ $\quad (q = 1-p)$.

Standard feil:

Standardfeilen til en estimator er standardavviket til estimatoren

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

Punktestimat for en parameter

Når vi setter inn observerte verdier fra et utvalg i en estimator for θ, får vi et punkt-estimat for θ

μ: $\bar{x}$ $\quad$ σ^2: s^2 $\quad$ p: x/n

Konfidensintervall:

Et konfidensintervall for parameteren θ på signifikansnivå α er et intervall $[A, B]$, med stokastiske grenser A og B, slik at

$$p(A \leq \theta \leq B) = 1 - \alpha$$

Det kalles gjerne et $(1-\alpha)\cdot 100\%$ konfidensintervall for parameteren θ.

Eks: $\alpha = 0.05 \longleftrightarrow$ 95% konfidens-intervall

$$p(A \leq \theta \leq B) = 95\%$$

i) Konfidensintervall for μ, σ kjent:

Anta: $\begin{cases} X_1, \ldots, X_n \text{ uavhengige, identisk} \\ \text{fordelte med } E(X_i) = \mu \text{ og} \\ \text{Var}(X_i) = \sigma^2 \text{ med } \underline{\sigma \text{ kjent}} \end{cases}$

og

$\bar{X}$ tilnærmet normalfordelt

(X_i normalfordelt eller $n \geq 30$) for hver i

Konfidensintervall:

$$\bar{X} \pm z_{\alpha/2} \cdot \underbrace{\sigma/\sqrt{n}}_{\text{SE}(\bar{X})}, \text{ dvs } \begin{cases} A = \bar{x} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \\ B = \bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n} \end{cases}$$

$\uparrow$ punkt-estimat for μ

SE($\bar{X}$) std feil til estimatoren $\bar{X}$

"Z-intervall for μ"

ii) Konfidensintervall for μ, σ ukjent

Anta: $X_1, \ldots, X_n$ uavhengige normalfordelte med $E(X_i) = \mu$ og $\text{Var}(X_i) = \sigma^2$

Konfidensintervall:

$$\bar{x} \pm t_{\alpha/2}^{n-1} \cdot s/\sqrt{n}, \text{ dvs } \begin{cases} A = \bar{x} - t_{\alpha/2}^{n-1} \cdot s/\sqrt{n} \\ B = \bar{x} + t_{\alpha/2}^{n-1} \cdot s/\sqrt{n} \end{cases}$$

- $\bar{x}$: punkt-estimat for μ
- $s/\sqrt{n}$: estimat for std.feil til $\bar{x}$

"T-intervall for μ"

iii) Konfidensintervall for σ^2

Anta: $X_1, \ldots, X_n$ uavhengige normalfordelte med $E(X_i) = \mu$ og $\text{Var}(X_i) = \sigma^2$

Kritiske verdier:
La Y være χ^2_{n-1}-fordelt ($n-1$ df.).
Vi skriver χ^2_α for den kritiske verdien definert ved

$$P(Y > \chi^2_\alpha) = \alpha$$

for enhver α.
Merk: Vi kan ikke finne χ^2_α med kalkulator, må bruke tabell. (Tabell E6 i boken)

Konfidensintervall:

$$\left[S^2 \cdot (n-1) / \chi^2_{\alpha/2}, \; S^2(n-1) / \chi^2_{1-\alpha/2} \right]$$

Merk: χ^2_{n-1}-fordelingen er ikke symmetrisk.

iv) Konfidensintervall for p

Anta: $\begin{cases} X \text{ binomisk fordelt med} \\ \text{parametre } (n, p) \end{cases}$

og

$\begin{cases} n \text{ stor slik at } X \text{ er} \\ \text{tilnærmet normalfordelt} \end{cases}$

Konfidensintervall:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ dvs } \begin{cases} A = \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ B = \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \end{cases}$$

- $\hat{p}$: punkt-estimat for p, $\hat{p} = x/n$
- $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$: estimat for std. feilen til $\hat{p}$

Hvor stor må n være?

$$n\hat{p}\hat{q} = n\hat{p}(1-\hat{p}) \geq 5$$

Hypotesetest: Vi tar stilling til en påstand om en parameter θ.

Std oppsett:

i) Velg modell og hypoteser

H_0: nullhypotese

H_1: alternativ hypotese

Vi velger H_1 = det vi ønsker å vise, og H_0 = komplementet til H_1

ii) Velg testobservator og finner på forkastningsområde (høyresidig, venstresidig, tosidig)

iii) Velg signifikansnivå α = akseptert sannsynlighet for type I – feil.

iv) Finn kritisk verdi og bestem forkastningsområde = verdier for testobservator som sier at vi forkaster H_0

v) Samle inn data, regne ut testobservator og ta stilling til påstanden.

To mulige utfall:

- Vi forkaster H_0 ("H_1 sann")
- Vi beholder H_0 ("H_0 sann")

	H_0 sann	H_1 sann ↔ objektiv sannhet
Beholde H_0	✓	Type II feil
Forkaste H_0	Type I feil	✓

↑ beslutning

Type I feil: "uskyldig dømt"
Type II feil: "skyldig gått fri"

i) Hypotesetest for μ

Anta: $X_1, \ldots, X_n$ uavhengige normal-fordelte med $E(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$

Tre typer:

a) $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$ — høyresidig

b) $H_0: \mu \geq \mu_0$, $H_1: \mu < \mu_0$ — venstresidig

c) $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$ — tosidig

Testobservator: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ eller $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

(σ ukjent) | (σ kjent)

T-fordelt m/ $n-1$ df. når $\mu = \mu_0$ — "T-test"

std. normal-fordelt — "Z-test"

Forkastningsområde:

a) Høyresidig: $T > t_\alpha$ eller $Z > z_\alpha$

b) Venstresidig: $T < -t_\alpha$ eller $Z < -z_\alpha$

c) Tosidig: $|T| > t_{\alpha/2}$ eller $|Z| > z_{\alpha/2}$

(kritiske t-verdier her n-1 df.)

Alternativer:

- Bruk $\bar{x}$ istedet for T / Z som testobservator (skriv om forkastnings-området)

- Bruk p-verdi:

  Regn ut p-verdi og sammenlikn med signifikansnivå α istedet for å finne forkastningsområde

p-verdi: sannsynlighet for en minst like ekstrem verdi for testobservator som den vi finner

Vi forkaster H_0 om $p < \alpha$
Vi beholder H_0 om $p \geq \alpha$

Styrkefunksjon og type II -feil:

Styrkefunksjon:

$\gamma(\mu) = p(\text{forkaster } H_0 \mid \mu)$

sannsynlighet for type II -feil for hver gitt verdi (sann verdi) av μ.

ii) Hypotesetest for p = populasjonsandel

Anta: { X binomisk fordelt med parametre (n, p)
og
n stor slik at X er tilnærmet normalfordelt $(n\hat{p}(1-\hat{p}) \geq 5)$

Merk: vi må ikke forveksle p her (parameter for binomisk ford.) med p-verdi

Testobservator: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ $\left(\hat{p} = \dfrac{x}{n}\right)$

Hypoteser:

a) $H_1: p > p_0$ b) $p < p_0$ c) $p \neq p_0$

Forkastningsområde:

a) $Z > z_\alpha$ b) $Z < z_\alpha$ c) $|Z| > z_{\alpha/2}$