Plan:

1. Oversikt over statistikk-kurset
2. Enkel data-analyse
3. Sum-notasjon
4. Kalkulator bruk
5. Regresjonslinjer og minste kvadraters metode

Oversikt over kurset:

Lærebok: Løvås Statistikk (4. utg. / 3. utg.)
Kap. 1 – 7.

Forelesn.: Tirs. 08–11 (fredag neste uke)
Veil. " 11–13
Kontortid: Tors. 09–11
Eksamen: 5. juni 3t. skriftlig eksamen
(skr. hjelpemidler + BI-kalk.)

Tema:
1. Sannsynlighetsteori
2. Data-analyse (deskriptiv statistikk)
3. Statistisk inferens (konfidensintervall, hypotesetester)

Intro: populasjon ⟷ utvalg

(alle objekt vi er interessert i)

populasjon $\xrightarrow{\text{uttrekk}}$ utvalg
- størrelsen
- tilfeldig

innhenting av data — data-analyse

kunnskap om populasjonen $\xleftarrow{\text{inferens}}$ innhenting av data

sannsynlighets-teori

Eks: populasjon: alle mulige borrehull i nordsjøen

utvalg: noen borrehull som kan bores

Eks:

	Bullet holes / sq. feet
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of plane	1.8

(2) Enkel data-analyse:

Hva kan vi si om utvalget?

Datasett, én variabel: X med n datapunkt

i	x_i
1	x_1
2	x_2
3	x_3
4	x_4
$\vdots$	$\vdots$
n	x_n

Datasett, to variabler: X, Y med n datapunkt

i	X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	
n	x_n	y_n

* Sentralmål:

i) **Gjennomsnitt:**

$$\bar{x} = \frac{x_1+x_2+\dots+x_n}{n} = \frac{1}{n}\cdot(x_1+x_2+\dots+x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Utvalgs-gjennomsnitt

Sum-notasjon

ii) **Median:** "den midterste verdien" når vi ordner $x_1, x_2, \dots, x_n$ i stigende rekkefølge.

n odde: median $= x_{((n+1)/2)}$ $\quad n=5$

n jevn: median = gj.snitt av $x_{(n/2)}$ og $x_{(n/2+1)}$

Ek: $x_1 = 5 \quad x_2 = 3 \quad x_3 = 1 \quad x_4 = 7 \quad x_5 = 2$

$x_{(1)} = 1 \quad x_{(2)} = 2 \quad x_{(3)} = 3 \quad x_{(4)} = 5 \quad x_{(5)} = 7$

median $= 3$

iii) **Modus:** den verdien som opptrer flest ganger

~~Kovarians:~~

~~Utvalgskovarians:~~ $\cancel{S_{xy} = \frac{(x_1-\bar{x})(y_1-\bar{y}) + \dots + (x_n-\bar{x})(y_n-\bar{y})}{n-1} = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}$

* Spredningsmål:

i) Standard-avvik: s_x

Utvalgsvarians: $$s_x^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$$

$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \geq 0$$

Utvalgs-standardavvik: $s_x = \sqrt{s_x^2}$

ii) Variasjonsbredde: differansen mellom største og minste verdi : $x_{(n)} - x_{(1)}$

* Kovarians:

Utvalgskovarians : $$s_{xy} = \frac{(x_1-\bar{x}) \cdot (y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{n-1}$$

$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Korrelasjonskoeffisient: $$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Eksempel:

i	x_i	y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$
1	23	42	−21	441	16	256	−336
2	34	27	−10	100	1	1	−10
3	55	20	11	121	−6	36	−66
4	60	19	16	256	−7	49	−112
5	48	22	4	16	−4	16	−16
Σ	220	130	0	934	0	358	−540

$$\bar{x} = \frac{23+34+55+60+48}{5} = \frac{220}{5} = \underline{44} \qquad \text{median}(x): \underline{48}$$

$$\bar{y} = \frac{42+27+20+19+22}{5} = \frac{130}{5} = \underline{26} \qquad \text{median}(y): \underline{22}$$

$$S_x^2 = \frac{934}{4} = 233.5 \qquad S_x = \sqrt{233.5} \simeq \underline{15.28}$$

$$S_y^2 = \frac{358}{4} = 89.5 \qquad S_y = \sqrt{89.5} \simeq \underline{9.46}$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) = \frac{-540}{4} = \underline{-135}$$

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{-135}{\sqrt{233.5}\cdot\sqrt{89.5}} \simeq \underline{-0.933}$$

③ Sum-notasjon

$$\sum_{i=1}^{n} x_i$$

Sum-notasjon

Σ: Sigma, betyr sum

x_i: uttrykket som skal summeres

$\left.\begin{matrix} n \\ i=1 \end{matrix}\right\}$: indeksverdier det skal summeres over $i = 1, 2, 3, \ldots, n$

Eks:

$$\sum_{i=1}^{4} i^2 = \underset{i=1}{1^2} + \underset{i=2}{2^2} + \underset{i=3}{3^2} + \underset{i=4}{4^2}$$

$$= 1 + 4 + 9 + 16 = \underline{\underline{30}}$$

$$\sum_{i=1}^{3} (x_i - 1)^2 = \underset{i=1}{(x_1-1)^2} + \underset{i=2}{(x_2-1)^2} + \underset{i=3}{(x_3-1)^2}$$

$$= x_1^2 - 2x_1 + 1 + x_2^2 - 2x_2 + 1 + x_3^2 - 2x_3 + 1$$

$$= \underbrace{x_1^2 + x_2^2 + x_3^2} - \underbrace{2(x_1 + x_2 + x_3)} + 3$$

$$= \sum_{i=1}^{3} (x_i^2 - 2x_i + 1) = \sum_{i=1}^{3} x_i^2 \div \sum_{i=1}^{3} 2x_i + \sum_{i=1}^{3} 1$$

$$= \sum_{i=1}^{3} x_i^2 - 2\sum_{i=1}^{3} x_i + 3 \cdot 1$$

Regneregler for summer:

i) $\sum_{i=1}^{n} (a_i \pm b_i) = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i$

ii) $\sum_{i=1}^{n} (c \cdot a_i) = c \cdot \sum_{i=1}^{n} a_i$

iii) $\sum_{i=1}^{n} c = n \cdot c$

$(a_1+b_1)+(a_2+b_2)+\dots$
$= a_1+a_2+\dots$
$+ b_1+b_2+\dots$

$ca_1+ca_2+\dots$
$= c\cdot(a_1+a_2+\dots)$

$c+c+\dots+c$
$= n\cdot c$

Husk: $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$

$x_1^2+x_2^2+\dots+x_n^2 \qquad (x_1+x_2+\dots+x_n)^2$

4 Kalkulatorbruk

i	x_i	y_i
1	23	42
2	34	27
3	55	20
4	60	19
5	48	22

Én variabel x

$\boxed{\text{C STAT}}$: tømmer statistikk-data

23 $\boxed{\Sigma+}$
34 $\boxed{\Sigma+}$
55 "
60 "
48 "

$\boxed{\bar{x},\bar{y}} \to \bar{x}=44$
$\boxed{S_x, S_y} \to S_x = 15.28$

$\boxed{\text{C STAT}}$ to var. x og y

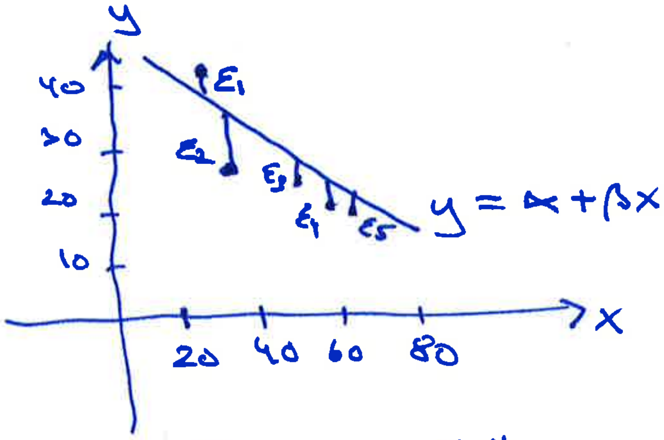
23 $\boxed{\text{INPUT}}$ 42 $\boxed{\Sigma+}$
34 " 27 "
⋮

$\boxed{\bar{x},\bar{y}} \to \bar{x}=44$
$\boxed{\bar{x},\bar{y}}$ $\boxed{\text{SWAP}} \to \bar{y}=26$
$\boxed{S_x, S_y} \to S_x = 15.28$ $\boxed{\text{SWAP}} \to S_y = 9.46$

$\boxed{\hat{x}, r}$ $\boxed{\text{SWAP}}$
$\to r = -0{,}934$

5 Minste kvadraters metode

i	x_i	y_i
1	23	42
2	34	27
3	55	20
4	60	19
5	48	22



Spredningsplott

Minste kvadraters metode:

Velge α, β slik at

$$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \varepsilon_5^2$$

er minst mulig.

Løsning:

$$\beta = r \cdot \frac{s_y}{s_x}$$

$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

Korrelasjonskoeffisienten r og stigningstallet β har samme fortegn

Noen viktige formler fra Forelesning I:

(A) Regneregler for summer:

(a_i, b_i : uttrykk med i
c : konstant)

$$\sum_{i=1}^{n}(a_i \pm b_i) = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i$$

$$\sum_{i=1}^{n} c a_i = c \cdot \sum_{i=1}^{n} a_i$$

$$\sum_{i=1}^{n} c = nc$$

(B)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(utvalgsgjennomsnitt)}$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \text{(utvalgsvarians)}$$

$$s_x = \sqrt{s_x^2} \qquad \text{(utvalgsstandardavvik)}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \qquad \text{(utvalgskovarians)}$$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \qquad \text{(korrelasjonskoeff.)}$$

Husk: $\sum_{i=1}^{n} x_i = n \cdot \bar{x}$ og $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$

(C) Regresjonslinjen:

Minste kvadraters metode

$$y = \alpha + \beta x \quad \text{der} \quad \begin{cases} \beta = r \cdot \dfrac{s_y}{s_x} \\ \alpha = \bar{y} - \beta\bar{x} \end{cases}$$