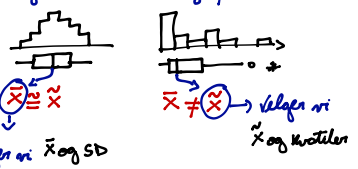
Module 1: Deskriptiv statistikk

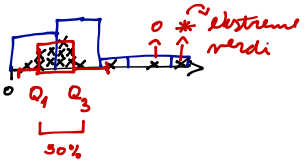
Kategoriske data // numeriske data

- frekvens → kategorier
  - frekvenstabell
  - stolpediagram
  - sektordiagram
- ← Binære or dikotome / Nominale / Ordinale

Tabeller; antall(%); sektordiagram (=stolpediagram)

Diskrete // kontinuelige data
- → Histogram, boksplott
- Symmetrisk
- Skjev-fordelt





- Kvartilene := between Q_1 and Q_3 is 50% of the data;
- SD := $\bar{X} \pm (2 \cdot SD)$ is 50% of the data.

→ Frequency table:
- R: tabel(df$variable)
- Python: df.value_counts()

→ Relative frequency:
- # of observations in an interval / total # of observations

→ Histogram:
- relative frequency in an interval / the width of the interval
- Forms: 1) Entoppet := convoluted; 2) Totoppet := two-peaks; 3) Symmetri := Symmetry; 4) Skjevhet := skewed.
  - 1) only one peak
- How do we define a histogram as symmetrical? We use skewness $\approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$ or kurtosis $\approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 - 3$
  - Symmetrical if $< 0{,}5$; $0{,}5 \leq$ weak ≤ 1; very > 1
  - < 0 := platykurtic;
  - 0 := Gauss;
  - > 0 := pointed.

→ Sum properties:

$$1)\ \sum_{i=1}^{n}(ax_i+b) = a\sum_{i=1}^{n}x_i + nb$$

$$2)\ \sum_{i=1}^{n}(x_i-b)^2 = \sum_{i=1}^{n}x_i^2 - 2x_ib + b^2 = \sum_{i=1}^{n}x_i - 2b\sum_{i=1}^{n}x_i + nb^2$$

R:
- > View(dataset_name) := to view the entire dataset;
- > ls(dataset_name) := to return the variables of the dataset;
- > str(");
- > table(" $variable_name) := frekvens tabel
- > hist(" ", breaks = seq(0, #final, # interval)) := histogram plot;
- > mean(" ");
- > abline(v = x, col='red') := horisontal line at x;
- > summary(" ");
- > sd(" ");
- > boxplot(" ", horizontal = TRUE, add = TRUE) := boxplot with histogram;

2) Shuffle rules:

→ Utfallsrom / outcomes
- Set of all possibles outcomes;

→ Hendelse / event:
- One or more sets of outcomes;

→ Probability of an event:
- $P(A) = \frac{\#\text{ favorable outcomes of } A}{\#\text{ possible outcomes of } A}$
- $0 \leq P(A) \leq 1$ and $1 - P(A) = P(A)^C = P(\bar{A})$

	replaced	not replaced
arranged	n^k	$P_{n,k} = \frac{n!}{(n-k)!}$
not arranged	$n!$	$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Remember:
- n = sample size;
- k = picked units
- $0! = 1$

returns := # of possible outcomes

Probability := $\frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}}$

Forventningsværdien: $* E[g(X)] = \sum_x g(x) \cdot p(x)$

↳ $\mu_X = \bar{X} = E[X] = \sum_{\forall x_i \in S} x \cdot P(X=x)$ (diskret); $\int_{\forall x_i \in S} x \cdot f(x)\,dx$ (kontinuerlig)

Empiriske forventningsværdien ($\bar{x}$):

↳ $\bar{x} = \frac{1}{m} \cdot \sum_{i=1}^{m} x_i$ for ukendte μ og/eller σ!

↳ **Regneregler:** $E[a] = a$; $E[aX] = a\,E[X]$; $E[a+bX] = a + b \cdot E[X]$; $E[a+bX+cX^2] = a + b \cdot E[X] + c \cdot E[X^2]$

$E[Y]$ where $Y = g(X)$ $\therefore \sum_{\forall x_i \in S} g(x) \cdot P(X=x)$

Varianser:

↳ $\sigma_X^2 = Var[X] = E[(X-\mu)^2] = \sum_x (x-\mu)^2 \cdot P(X=x) = \left(\sum x^2 \cdot P(X=x)\right) - \mu^2 = E[X^2] - E[X]^2$

↳ **Regneregler:** $Var[a] = 0$; $Var[X] \geq 0$; $Var[a+X] = Var[X]$; $Var[b \cdot X] = b^2 \cdot Var[X]$

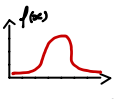
$Var[a+bX] = b^2\,Var[X]$; $Var[\bar{X} - \bar{Y}] = Var[\bar{X}] + Var[\bar{Y}]$

Standardafvigelse:

↳ $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2}$ for population

↳ $std_X = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2}$ for ukendte μ og/eller σ!

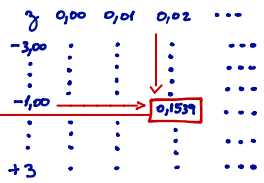
Normalfordeling:

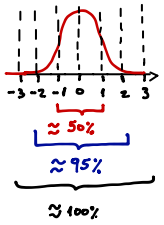
↳ $f(x) = y = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ ⟶  ⟶ sandsynlighedstætheden

↳ **Standard normalfordeling $N(\mu=0, \sigma=1)$:**

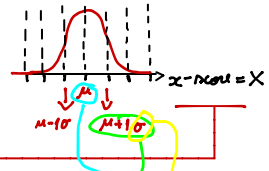
↳ Tabel E3 - kumulativ standardnormalfordeling:

↳ for $Z \sim N(0,1)$:





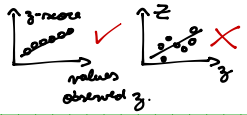
* for $\mu \neq 0$, $X \sim N(\mu, \sigma)$:



* **Standardiseret version af X:**

$X \sim N(\mu, \sigma) \longrightarrow Z = \frac{X-\mu}{\sigma} \longrightarrow Z \sim N(0,1)$

R: $pnorm(z=y) := P(Z \leq y)$ for $Z \sim N(0,1)$

$qplot(z) \cdot qpnorm() :=$ 

$* X_i \sim norm(\mu, \sigma) = \mu + \varepsilon_i$ where $\varepsilon_i \sim norm(0, \sigma)$

Linear Kombination:

↳ $\frac{1}{m}\sum_{i=1}^{m} X_i = \frac{1}{m}(X_1 + X_2 + \cdots + X_m) = \frac{1}{m} \cdot X_1 + \frac{1}{m} \cdot X_2 + \cdots + \frac{1}{m} \cdot X_m$

↳ Ex.: $A \sim N(\mu_A, \sigma_A)$ and $B \sim N(\mu_B, \sigma_B)$, then $X = A + B$. Hvad er μ_X og σ_X og $X \sim N(\mu_X, \sigma_X)$?

↳ $E[cA+B] = c\,E[A] + E[B] = c\mu_A + \mu_B$

↳ $Var[cA+B] = c^2 \cdot Var[A] + Var[B] = c^2 \cdot (\sigma_A)^2 + (\sigma_B)^2$

↳ $SD[cA+B] = \sqrt{c^2(\sigma_A)^2 + (\sigma_B)^2}$

$\mu_X = c\mu_A + \mu_B$

$\sigma_X = \sqrt{c^2(\sigma_A)^2 + (\sigma_B)^2}$ ☒

Binomisk fordeling:

$X \sim bin(\#forsøg, P(X = succes))$

↳ Ex.: $X \sim bin(1000, 0{,}2)$

↳ $E[X] = n \cdot p = 200$

↳ $Var[X] = \sqrt{n \cdot p(1-p)} = 12{,}6$

$X \sim N(200, 12{,}6)$ ☒

$bin(n, p) \approx N(np, \sqrt{np(1-p)})$

*** Important rules:**

↳ $P(X > x) = 1 - P(X \leq x)$; or $P(X \geq x+1)$

↳ $P(a < X \leq c) = P(X \leq c) - P(X \leq a)$;

↳ $P(a \leq X < c) = P(a-1 < X \leq c-1) = P(X \leq c-1) - P(X \leq a-1)$

$* Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY - XE[Y] - YE[X] + E[X]E[Y]] = E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]$

$= E[XY] - E[X]E[Y]$ ☒

* $Cov(X,X) = Var(X)$
* $Cov(X,Y) = Cov(Y,X)$
* $Cov(cX, Y) = c\,Cov(X,Y)$
* $Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)$

3) Probability rules:

→ Addition rule ('or'):

↳ $P(A \cup B) = P(A) + P(B)$ → A and B are disjoint / mutually exclusive → They can not occur at the same time;

↳ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ → not disjoint.

↳ $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$ → Included excluded rule

↳ $P(A \cup B \cup C) = 1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})$ → A, B and C are independent → $\overline{A \cup B \cup C} = \bar{A} \cap \bar{B} \cap \bar{C}$ and $P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C}) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$

↳ $= 1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})$

→ Conditional rule:

↳ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

↳ $P(\bar{A}|B) = 1 - P(A|B)$

↳ $P(A|B) = P(A)$ → A and B are independent.

→ Multiplication rule ('and'):

↳ $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$ → A, B dependent

↳ $P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$

↳ $P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) P(A|B_2) + \dots + P(B_m) \cdot P(A|B_m)$

↳ $P(A \cap B) = P(A) \cdot P(B)$ → A and B are independent;

↳ $P(A \cap B) = 0$ → A and B are disjoint.

↳ $P((A \cap B)^c) = P(A^c \cup B^c) = P(A^c) + P(B^c) - P(A^c \cap B^c)$ ⊛

→ Bayes' rule:

↳ $P(B_i|A) = \frac{P(A \cap B_i)}{P(A)}$

↳ $= \frac{P(B_i) \cdot P(A|B_i)}{P(A)}$

Utfallsrom	GGGG	GGGJ GGJG GJGG JGGG	GGJJ GJGJ GJJG JGGJ JGJG JJGG	GJJJ JGJJ JJGJ JJJG	JJJJ
Antall jenter (y)	0	1	2	3	4
$P(Y = y)$	1/16	4/16	6/16	4/16	1/16
$P(Y \leq y)$	1/16	5/16	11/16	15/16	1.0

→ $\frac{1}{16} + \frac{4}{16}$

→ Cumulative distribution:

↳ $F(x) = P(X \leq x)$

↳ $P(a < X \leq b) = F(b) - F(a)$

↳ $P(X > a) = 1 - F(a)$

↳ $P(X \leq b) = F(b)$

→ $P(x=0, y=0) = 0{,}09$:= joint probability

	y					
x	0	1	2	3	4	$P(X = x)$
0	0.09					0.09
1	0.11	0.09				0.20
2	0.07	0.12	0.07			0.26
3	0.05	0.09	0.03	0.01		0.18
4	0.01	0.03	0.05	0.02		0.11
5	0.01	0.01	0.03	0.02	0.01	0.08
6		0.01	0.01	0.02	0.01	0.05
7				0.02	0.01	0.03
$P(Y = y)$	0.34	0.35	0.19	0.09	0.03	1.00

marginal probability of X

marginal probability of Y

→ joint distribution:

↳ $P(X, Y) = P(X = x \text{ and } Y = y) = P(X = x \cap Y = y)$

↳ $P(X = x) \cdot P(Y = y)$ → X, Y independent

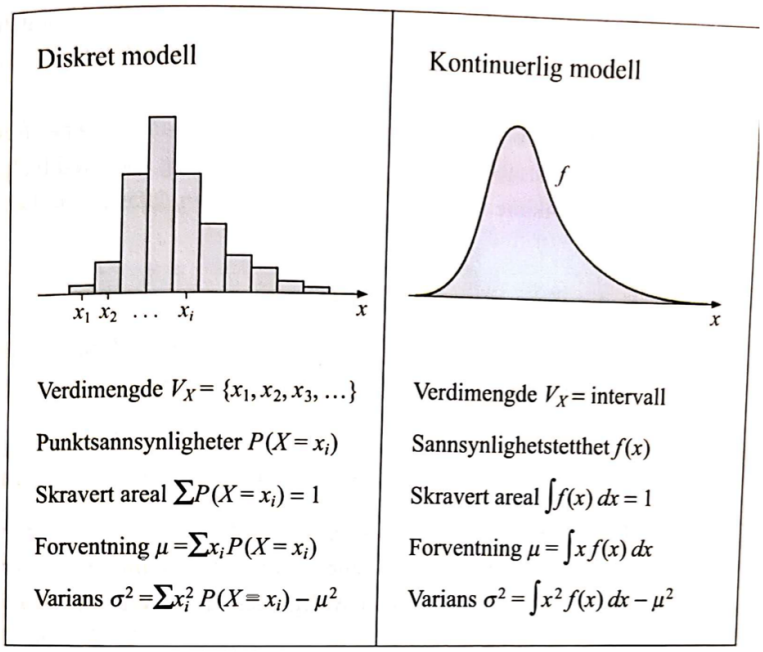
↳ $P(X = x) \cdot P(Y = y | X = x)$ → X, Y dependent

i.l.:

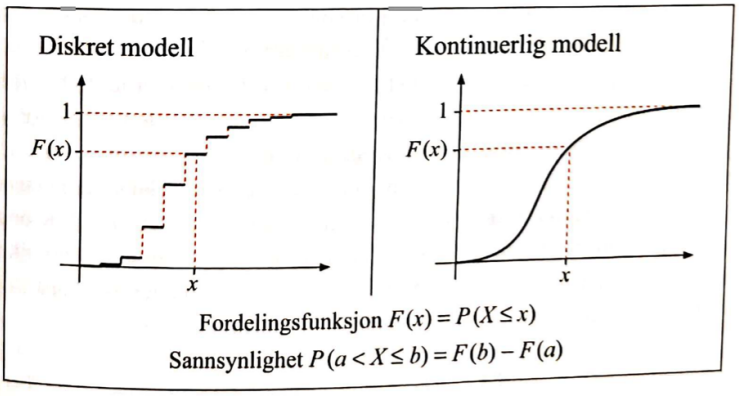
↳ $P(X + Y = 4) = 0{,}01 + 0{,}09 + 0{,}07 = 0{,}17$

↳ $P(X = 4 | Y = 1) = \frac{P(X = 4 \text{ and } Y = 1)}{P(Y = 1)} = \frac{0{,}03}{0{,}35} = 0{,}086$

Chapter 4 : Stokastiske variabler :



Figur 4.16 De viktigste forskjellene mellom diskrete og kontinuerlige modeller



Figur 4.17 De viktigste likhetene mellom diskrete og kontinuerlige modeller.

→ Expectation :

- $\mu = E[X^1] = \sum_{i=1}^{m} x_i^1 \cdot P(X = x_i)$
- $E(a+bX) = a + bE[X]$
- $E[a + bX + cX^2] = a + bE[X] + cE[X^2]$
- $E[g(x)] = \sum_{i=1}^{m} g(x_i) \cdot P(X = x_i)$
- $E[X^2] = \sum_{i=1}^{m} x_i^2 \cdot P(X = x_i)$
- $\mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$ → continuous
- $E[X \pm Y] = E[X] \pm E[Y]$
- $E[X \cdot Y] = \sum_{\forall x_i} \sum_{\forall y_i} x_i \cdot y_i \cdot P(x_i, y_i)$ → X, Y dependent
- $E[X \cdot Y] = E[X] \cdot E[Y]$ → X, Y independent

→ Variance :

- $\sigma^2 = Var[X] = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2$
- $E[(X-\mu)^2] = E[X^2] - E[\mu]^2$
- $Var[bX + a] = b^2 Var[X]$
- $\sigma^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - \mu^2$ → continuous
- $Var[aX + bY + c] = a^2 Var[X] + b^2 Var[Y] + 2ab\,Cov[X,Y]$ (only if X, Y dependent!)

→ Covariance :

- $Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X \cdot \mu_Y$

→ Correlation :

- $\rho(X,Y) = Corr[X,Y] = \frac{Cov[X,Y]}{\sigma_X \cdot \sigma_Y}$
- $-1 \le \rho \le 1$

Bevis for regel 4.19

Start med å vise at

$$E(X) = \sum_{\text{alle } x_i} x_i \cdot P(X = x_i) = \sum_{\text{alle } y_j} \sum_{\text{alle } x_i} x_i \cdot P(x_i, y_j)$$

Ta utgangspunkt i ligning (4.13) på side 157 og vis at

$$P(x_i, y_j) = P(X = x_i \mid Y = y_j) \cdot P(Y = y_j)$$

Utnytt dette og vis at

$$E(X) = \sum_{\text{alle } y_j} E(X \mid y_j) \cdot P(Y = y_j)$$

hvor den betingede forventningen til X, gitt at $Y = y_j$, er definert lik

$$E(X \mid y_j) = \sum_{\text{alle } x_i} x_i \cdot P(X = x_i \mid Y = y_j)$$

→ Double expectation :

- $E[X] = \sum_{\forall y_i} P(Y = y_i) \cdot E[X \mid y_i]$ (side 165)

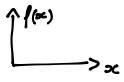
Chapter 5: Probability models

↳ Empirically := based on data ; Theoretically := based on parameters

Stokastisk variable := utfallsrom → Tall := utfall i et utfallsrom til en verdi. Eksempler: X, Y, Z. Verdiene de stokastiske variablene: x, y, z.

Verdimengden := V_X er mengden av alle verdiene x kan X kan ha.

Sannsynlighetsfordeling := fortteller hvor sannsynlig det er at X har de ulike verdiene i verdimengden. → $f(x)$ →



$P(X=x)$:= Sannsynligheten for at den stokastiske variablen X har verdien x.

* Diskret variabel := tellevariabel ;

↳ Punktsannsynlighet → $\sum_{\forall x \in S} P(X=x) = 1$

↳ Tabeller med punktsannsynligheter

↳ Sylindiagram.

↳ Binomisk fordeling:

↳ X = # suksesser på n forsøk

↳ Poisson-fordeling:

↳ X = # hendelser i et gitt areal, tidsintervall, lengde, volum.

↳ Geometrisk fordeling:

↳ X = # forsøk inntil første suksess.

* Kontinuerlig variabel := målevariabel

↳ Sannsynlighetstetthet → $\int_{\forall x \in S} f(x)\,dx = 1$

↳ Tetthetskurve, tabeller med kumulativ sannsynligheter.

↳ Uniform fordeling; Normalfordeling; Eksponensialfordeling; Kjikvadratfordeling; F-fordeling:

↳ X = en måling ;

X = en funksjon av flere målinger.

V_X:	0	1	2	3	4	T	
$P(X=x)$:	1/16	4/16	6/16	4/16	1/16	1	→ Punktsannsynligheten := $P(1 \le X \le 3) = P(X=2) + P(X=3)$
$P(X \le x)$:	1/16	5/16	11/16	15/16	16/16	1	→ Kumulativ fordeling := $P(1 \le X \le 3) = P(X \le 3) - P(X \le 1)$

→ Binomial distribution:

↳ 1) Perform a certain # of attempts (n); 2) We are interested in only 2 outcomes (A or B) in each attempt;

↳ 3) $p = P(A)$ is the same in all attempts; 4) Each attempt is independent of each other;

↳ $X \sim bin(p, n)$ where X is the # of all favorable outcome (A);

↳ $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \cdots, n$ → distribution

↳ $E[X] = np$ and $Var[X] = np(1-p)$;

↳ $X \sim bin(p, n) \sim norm(np, \sqrt{np(1-p)})$

↳ $P(X \le x) = F(x) \approx G\left(\frac{x-\mu}{\sigma}\right)$ if $\sigma^2 \ge 5$ where X is binomial

→ Normal distribution:

↳ $X \sim norm(\mu, \sigma)$

↳ $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

$$P(X \le x) = F(x) = G\left(\frac{x-\mu}{\sigma}\right)$$

$$P(X > x) = 1 - G\left(\frac{x-\mu}{\sigma}\right)$$

$$P(a \le X < b) = G\left(\frac{b-\mu}{\sigma}\right) - G\left(\frac{a-\mu}{\sigma}\right)$$

→ Standard normal distribution:

↳ $Z = \frac{X-\mu}{\sigma} \sim norm(0, 1)$

↳ $G(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-x^2}{2}}\,dx$

↳ $P(Z > z_\alpha) = \alpha$

→ Confidence interval:

↳ It is $100(1-\alpha)\%$ secure that X has a value in the interval $\mu \pm z_{\frac{\alpha}{2}} \cdot \sigma$;

↳ ⊕ := larger than ; ⊖ := less than;

→ Central limit theorem:

↳ $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim norm(\mu, \frac{\sigma}{\sqrt{n}})$ where X_is are independent and $i = 1, \cdots, n$

↳ Condition: $n \ge 30$

CH. 6: Estimations:

* **Unbiased estimator**: = the average of many estimates will approximate to the the real parameter value as long as the sample size $n \to \infty$.

→ **Estimating of μ:**

↳ Unbiased estimator $\bar{X}$ for estimate $\hat{\mu}$ → $E[\hat{\mu}] = E[\bar{X}] = \mu$

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

↳ $SE[\hat{\mu}] = SE[\bar{X}] = \frac{\sigma}{\sqrt{n}}$ → σ known → $KI := \left[\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$ → $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

↳ $SE[\bar{x}] = \frac{s}{\sqrt{n}}$ → σ unknown → $KI := \left[\bar{x} \pm t_{df, \frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$ → $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

→ **Estimating of p:**

↳ Unbiased estimator $\frac{x}{n}$ for estimate $\hat{p}$ → $E[\hat{p}] = E[\frac{x}{n}] = p$

↳ $SE[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ → $KI := \left[\hat{p} \pm z_{\frac{\alpha}{2}} \cdot SE[\hat{p}]\right]$

→ **Estimating of σ^2 and σ:**

↳ Unbiased estimator $\sqrt{s^2} = S$ for estimate σ → $E[\hat{\sigma}] = E[S] = \sigma$

↳ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ → $S = \sqrt{S^2} = \hat{\sigma}$

↳ KI := page 251

Unbiased estimator Proofs:

* $E[\bar{X}] = \hat{\mu}$ Proof → $E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\cdot\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \hat{\mu}$ ▨

* $E[s^2] = \hat{\sigma}^2$ Proof → $E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \overset{\text{TK33}}{=} \frac{1}{n-1}E\left[\sum_{i=1}^{n} x_i^2 - \bar{x}^2 n\right] = \frac{1}{n-1}\sum_{i=1}^{n} E[X_i^2] - n\cdot E[\bar{x}^2]$

↳ $= \frac{1}{n-1}\cdot\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\cdot(\frac{\sigma^2}{n}+\mu^2) = \frac{1}{n-1}\, n\sigma^2 + n\mu^2 - \sigma^2 - \mu^2 n = \frac{(n-1)\sigma^2}{(n-1)} = \hat{\sigma}^2$ ▨

Triks 1.: $E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\cdot\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \mu$ ▨

Triks 2.: $\bar{X} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \longleftrightarrow n\cdot\bar{X} = \sum_{i=1}^{n} x_i$ ▨

Triks 3: $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ ▨

($-2\bar{x}\sum_{i=1}^{n} x_i = -2\bar{x}\, n\bar{x} = -2n\bar{x}^2$; $\sum_{i=1}^{n}\bar{x}^2 = n\cdot\bar{x}^2$)

Triks 4: $Var(X_i) = E(X_i^2) - E[X_i]^2 \longleftrightarrow E(X_i^2) = Var(X_i) + E[X_i]^2 = \sigma^2 + \mu^2$ ▨

Triks 5: $Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\cdot Var\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}$ ▨

only possible because X_i's are independent!

Triks 6: $E[\bar{X}^2] = Var[\bar{X}] + E[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2$ ▨

Triks:

$E[\bar{X}] = \mu$

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

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

$E[X_i^2] = \sigma^2 + \mu^2$

$Var[\bar{X}] = \frac{\sigma^2}{n}$

$E[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2$

CH.6: Hypothesis testing:

	H_0 true	H_1 true
Keep H_0	good	$P(\text{error 2})$ (β)
Discart H_0	$P(\text{error 1})$ (α)	good

where α: Significance level → The maximum allowed probability of 1-error.

* if $\downarrow \alpha$, then $\uparrow \alpha$. However if $\uparrow m$, then $\downarrow \alpha$ and $\downarrow \beta$.

* **Statistic Model:**

↳ One group := $X_i \sim N(\mu, \sigma)$, $i = 1, \dots, m$ or $X_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma)$, $i = 1, \dots, m$

→ $E[X_i] = E[X_i] + E[\varepsilon_i] = \bar{X} + 0 = \bar{X}$ ($E[\varepsilon_i] = 0$)

↳ Two groups := For σ the same := pairwised $X_i \xrightarrow{dep} Y_i$

$$\begin{cases} X_i \sim N(\mu_A, \sigma), \ i = 1, \dots, m_A \\ Y_j \sim N(\mu_B, \sigma), \ j = 1, \dots, m_B \end{cases} \quad \text{or} \quad Y_i \sim N(\mu_D, \sigma), \ i = 1, \dots, m_A + m_B \text{ where } \mu_D = \mu_A - \mu_B$$

$$\text{or} \quad Y_i = \mu_D + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma), \ i = 1, \dots, m_A + m_B$$

* **Hypotesis claims:**

↳ $H_0 := \mu_0 = 0$ Vs $H_1 := \begin{cases} \mu_0 > 0 \rightarrow \text{one-sided} \rightarrow \alpha \\ \mu_0 < 0 \rightarrow \text{one-sided} \rightarrow \alpha \\ \mu_0 \neq 0 \rightarrow \text{two-sided} \rightarrow \frac{\alpha}{2} \end{cases}$

* **Rejection limits (k):**

$$P\left(Z = \frac{\bar{D} - \mu_0^{\,0}}{SE[\bar{D}]} \overset{k}{=} \right) = \alpha$$

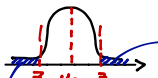
$= z_\alpha$ or $z_{\frac{\alpha}{2}}$ → one-sided, 2-sided

* **test statistic:**

↳ Z-test := $Z^* = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{m}}$; T-test := $T^* = \frac{\bar{X} - \mu_0}{S / \sqrt{m}}$

↳ p-test := $Z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ or $Z^* = \frac{X - np_0}{np_0(1-p_0)}$

* **p-value:**



p-value = $P(Z \geq Z^*_\alpha \mid H_0 \text{ is true})$ ↳ for one-sided $H_1 := \mu_0 > 0$

p-value = $P(Z \leq Z^*_\alpha \mid H_0 \text{ is true})$ ↳ for one-sided $H_1 := \mu_0 < 0$

∴ p-value = $P(|Z| \geq Z^*_{\frac{\alpha}{2}} \mid H_0 \text{ is true})$ ↳ for 2-sided $H_1 := \mu_0 \neq 0$

* **Rejection region:**

↳ Reject H_0 if:

$$\begin{cases} Z^* \text{ or } T^* \geq z_\alpha \text{ or } t_{m-1;\alpha} \rightarrow H_1 := \mu_0 > 0 \\ Z^* \text{ or } T^* \leq z_\alpha \text{ or } t_{m-1;\alpha} \rightarrow H_1 := \mu_0 < 0 \\ |Z^*| \text{ or } |T^*| \geq z_{\frac{\alpha}{2}} \text{ or } t_{m-1;\frac{\alpha}{2}} \rightarrow H_1 := \mu_0 \neq 0 \end{cases}$$

* **p-value** = $P(\text{Error type 1})$

↳ Probability of obtaining a test statistic value at least as contradictory to the null hypotesis as the value that actually resulted.

∴ p-value $> \alpha$:= hold H_0
p-value $\leq \alpha$:= reject H_0

p-value = (calculated test statistic z_0^*)

$$\begin{cases} 1 - \Phi(z_0^*) \rightarrow \text{for upper tailed test} \rightarrow H_1 := \mu_0 > 0 \\ \Phi(z_0^*) \rightarrow \text{for lower-tailed test} \\ 2\left[1 - \Phi(|z_0^*|)\right] \rightarrow \text{for two-tailed test} \end{cases}$$

* **Significance level α** := We are $100(1-\alpha)\%$ secure that μ_0 is equal to 0. → When H_0 is true.

8.3: T-test for 2-groups:

→ 8.3.1: Not pairwised:

↳ Model statistic: $X_i \sim T(\mu_x, \sigma_x)$ for $i = 1, \dots, m_1$ and $Y_j \sim T(\mu_y, \sigma_y)$ for $j = 1, \dots, m_2$

↳ Hypotesis test: $H_0 := \mu_x - \mu_y = \mu_D = 0$ Vs $\mu_x > \mu_y$ or $\mu_D > 0$

$\mu_x < \mu_y$ or $\mu_D < 0$

$\mu_x \neq \mu_y$ or $\mu_D \neq 0$

↳ Estimations:

↳ Unbiased estimator $\bar{X}$ for $\hat{\mu}_x$ and $\bar{Y}$ for $\hat{\mu}_y$;

↳ Unbiased pooled variance S_p^2 for $\hat{\sigma}^2_{x-y} = \dfrac{(m_x - 1) S_x^2 + (m_y - 1) S_y^2}{m_x + m_y - 2}$; → $\sigma^2_{x-y} = \sigma_x^2 + (-1)^2 \sigma_y^2 = \sigma_x^2 + \sigma_y^2$

where $S_\theta^2 = \dfrac{1}{m_\theta - 1} \sum_{i=1}^{m_\theta} (X_i - \bar{X})^2$

↳ Unbiased estimator S_p for $\hat{\sigma}_{x-y} = \sqrt{S_p^2}$

↳ Standard error $SE[\bar{x} - \bar{y}] = SE[\bar{D}] = S_p \cdot \sqrt{\frac{1}{m_x} + \frac{1}{m_y}}$

↳ Test statistic: $T = \dfrac{(\bar{x} - \bar{y}) - \mu_D^{\,0}}{SE[\bar{x} - \bar{y}]} = \dfrac{(\bar{x} - \bar{y})}{S_p \sqrt{\frac{1}{m_x} + \frac{1}{m_y}}}$

↳ Rejection areas: → Reject H_0 if:

- $T \geq t_{df, \alpha}$ if $H_1 := \mu_D > 0$ → one-sided
- $T \leq t_{df, \alpha}$ if $H_1 := \mu_D < 0$ → one-sided
- $|T| \geq t_{df, \frac{\alpha}{2}}$ if $H_1 := \mu_D \neq 0$ → two-sided

Remember!
* $df = m_1 + m_2 - 2$
* $100(1-\alpha)\%$

↳ Confidence Interval: $\left[(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}} \cdot SE[\bar{x} - \bar{y}]\right]$

↳ p-value $= P(T \geq T^* \mid H_0 \text{ is true}) =$ →

- $1 - \Phi(T^*)$
- $\Phi(T^*)$
- $2\left[1 - \Phi(|T^*|)\right]$

↳ Reject H_0 if: p-value $< \alpha$

→ 8.3.2: Pairwised:

↳ Paired observations $(X_1, Y_1), \dots, (X_m, Y_m)$

↳ Model: $X_i - Y_i = D_i = d_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,1)$ for $i = 1, \dots, m$ and d_i is independent;

↳ Hypotesis test: $H_0 := \mu_D = 0$ Vs $H_1 := \mu_D \neq 0$

↳ Estimations: $\hat{\mu}_D = \bar{D} = \bar{X} - \bar{Y} = \frac{1}{m} \cdot \sum_{i=1}^{m} D_i$ and $\hat{\sigma}^2 = S^2 = \frac{1}{m-1} \cdot \sum_{i=1}^{m} (D_i - \bar{D})^2$ and $\hat{\sigma} = S = \sqrt{S^2}$;

↳ Standard error: $SE[\bar{D}] = \dfrac{S}{\sqrt{m}}$

↳ test statistic := $T = \dfrac{\bar{D} - \mu_D}{S/\sqrt{m}}$

↳ $100(1-\alpha)\%$ KI := $\left[\bar{d} \pm t_{df, \frac{\alpha}{2}} \cdot SE[\bar{D}]\right]$ where $df = m - 1$

Variance Analysis: → for many groups

→ **One-Way ANOVA:** → studies the effect of labled populations on a single-classification:

↳ k := # groups; n := # samples; Y_{ij} := Observation # j from group i (label of populations := factors);

↳ Model := $Y_{ij} = \mu_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,1)$; $i = 1, \dots, k$ and $j = 1, \dots, n$ and ε_i independent

↳ Hypotesis test: $H_0 := \mu_1 = \mu_2 = \cdots = \mu_k$ Vs $H_1 :=$ at least 2 of the μ_i's are different

↳ Estimations:

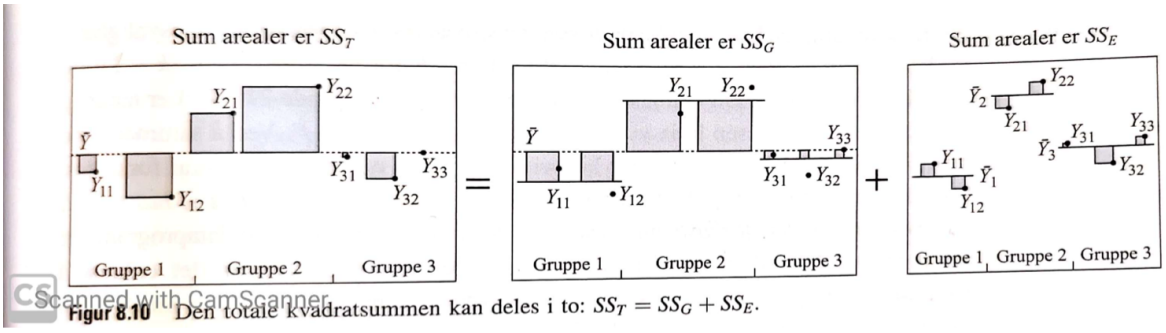
↳ Individual sample mean := $\bar{Y}_{i\cdot} = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij}$; Grand mean := $\bar{Y} = \frac{1}{k \cdot n} \cdot \sum_{i=1}^{k} \sum_{j=1}^{n} Y_{ij}$;

↳ Variances := $Y_{ij} = \bar{Y}_{i\cdot} + \varepsilon_i \rightarrow \varepsilon_i = Y_{ij} - \bar{Y}_{i\cdot} \rightarrow (Y_{ij} - \bar{Y}) = (\bar{Y}_{i\cdot} - \bar{Y}) + (Y_{ij} - \bar{Y}_{i\cdot})$

$$\therefore \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{Y}_{i\cdot} - \bar{Y})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

Total sum of squares (TSS) | GSS | Error sum of squares (ESS)

GSS → $n_i \cdot \sum_{i=1}^{k} (\bar{Y}_{i\cdot} - \bar{Y})^2$



Figur 8.10 Den totale kvadratsummen kan deles i to: $SS_T = SS_G + SS_E$.

$$\therefore \begin{cases} \rightarrow \text{Total Variance} := S_T^2 = \frac{TSS}{n-1} \\ \rightarrow \text{Variance between groups} := S_G^2 = \frac{GSS}{k-1} = MSG \\ \rightarrow \text{Variance within groups} := S_E^2 = \frac{ESS}{n-k} = MSE \end{cases} \quad \text{where } n = \sum_{i=1}^{k} n_i$$

(describe variation between groups)

↳ test statistic: $F^* = \frac{MSG}{MSE} = \frac{S_G^2}{S_E^2}$

↳ Rejection region := Reject H_0 if $\begin{cases} F^* \geq F_{df,\alpha} & \text{where } df = (k-1 \text{ and } n-k) \quad n = \sum_{i=1}^{k} n_i \\ p\text{-value} \leq \alpha \rightarrow P(F \geq F^* \mid H_0 \text{ is true}) \leq \alpha \end{cases}$

⊛ Hold H_0: p-value $> \alpha$:= There is no significant difference between ...

⊛ Proportion of explained variation → Coefficient of determination (R^2): $R^2 = \frac{SSG}{SST}$

Definition of σ: total variation over all groups?
proportion?
within groups?

Variance Analysis with Contrast:

↳ Model := $Y_{ij} = Y_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,1)$, independent ; $i = 1,\dots,k$ and $j = 1,\dots,n$

↳ Contrast := linear combination of averages $\longrightarrow \Theta = \sum_{i=1}^{k} c_i \mu_i$ where $\sum_{i=1}^{k} c_i = 0$

↳ c will be chosen by what we are interested to test.

↳ test hypotesis := i.e.: μ_1 Vs $\mu_3 := 1\cdot\mu_1 + 0\mu_2 - 1\mu_3 = \Theta = \mu_1 - \mu_3 \longrightarrow H_0 := \mu_1 - \mu_3 = 0$ Vs $H_1 := \mu_1 - \mu_3 \neq 0$

or: μ_1 Vs μ_2 and $\mu_3 := 1\mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = \mu_1 - \frac{\mu_2}{2} - \frac{\mu_3}{2} \longrightarrow H_0 := \mu_1 - \frac{\mu_2}{2} - \frac{\mu_3}{2} = 0$ Vs $H_1 :=$ not all equals

↳ Estimations := $\hat{\Theta} = \sum_{i=1}^{k} c_i \cdot \bar{Y}_i$ and $\mathrm{Var}[\hat{\theta}] = \sigma^2 \cdot \sum_{i=1}^{k} \frac{c_i}{n_i}$ and $SE[\hat{\theta}] = \sqrt{MSE \cdot \sum_{i=1}^{k} \frac{c_i^2}{n_i}}$

↳ Split variances := $Y_{ij} - \bar{Y} = \bar{Y}_i - \bar{Y} + Y_{ij} - \bar{Y}_i \longrightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$

TSS = $\sum_{i=1}^{k} n_i (Y_i - \bar{Y})^2 = TrSS$; ESS

∴
- Total variance := $\frac{TSS}{n-1} = S_T^2$
- Treatment variance (between groups) := $MSG = \frac{TrSS}{k-1} = S_{Tr}^2$
- Residual / Error variance (within groups) := $MSE = \frac{ESS}{n-k} = S_E^2$

for $n = \sum_{i=1}^{k} n_i$

↳ test statistic := $T^* = \frac{\hat{\Theta} - \Theta_0}{SE[\hat{\Theta}]}$ $\quad df = \frac{\alpha}{2}\,;\, n-k$

Model assumption:

↳ ε_{ij} := residual $\longrightarrow X_{ij} = \mu_i + \varepsilon_{ij} \longrightarrow \varepsilon_{ij} = X_{ij} - \mu_i \longrightarrow$ for $\hat{\mu}_i = \bar{X}_i \longrightarrow \varepsilon_{ij} = X_{ij} - \bar{X}_i$

↳ Independent ε_{ij} vs j $\longrightarrow$ correlations

↳ Normal distributed $\varepsilon_{ij} \sim N(0,\sigma)$; $\longrightarrow$ calculate each residual $\varepsilon_{ij} = X_{ij} - \bar{X}_i$ and see the plot or mediana Vs mean ;

↳ Constant variance. $\longrightarrow$ Check standart deviations if they are similar or plot groups X samples ;

↳ Making the model better:

↳ Use $\log(X_{ij})$ instead of X_{ij} := when not constant variance and/or not normal distributed !

Regression Analysis:

↳ 1st Plot between classes : Check : direction (+ or −) ; form (line, curve) ; power (strong, weak).

↳ correlation := How strong are the linear combination ;

↳ correlation coefficient := $r = \frac{\text{covariance between X and Y}}{S_X \cdot S_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

↳ Condition = numerical variable

↳ $r > 0$: positive ; $r < 0$: negative ; $r = 0$: no linear correlation between groups ;

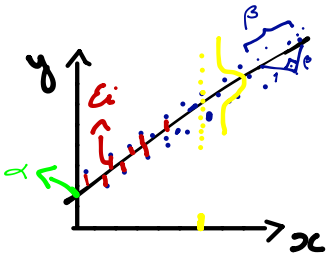
Linear Regresion:

* Explanatory variables (X_i); Response variables (Y_i);

→ Statistic Model:

↳ $Y_i = \alpha + \beta x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0,\sigma)$ for $i = 1,\dots,n$; → $Y_i \sim N(\alpha+\beta x_i, \sigma)$ → independent → ε_i := Residual →

↳ X_is are independent variable; Y_is are dependent variable;



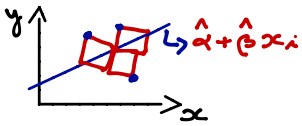
Where:
- α := Expected value of the response variable when the explanatory variable is zero;
- β := Expected changing of the response variable when the explanatory variable increases 1 unit;

→ Estimation of parameters:

↳ Unknown parameters: α, β, σ → describe the variation of the response with a fixed value of x.

↳ least squares method:

↳ $K = \sum_{i=1}^{m} [y_i - (\alpha + \beta x_i)]^2$ →



↳ $\hat{\beta} = \dfrac{\sum_{i=1}^{m}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{m}(x_i-\bar{x})^2} = \dfrac{S_{xy}}{S_x^2} = r\dfrac{S_y}{S_x}$

↳ $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

$\hat{Y}_i$, $\hat{\sigma}^2$ and $\hat{\sigma}$:=

- $E[Y_i|X_i] = E[\alpha+\beta x_i+\varepsilon_i] = E[\alpha]+E[\beta x_i]+E[\varepsilon_i]^{\to 0} = \alpha+\beta x_i = \hat{Y}_i$ (*)
- $Var[Y_i|X_i] = Var[\alpha+\beta x_i+\varepsilon_i] = 0+0+Var[\varepsilon_i] = \sigma^2$
- $SD[Y_i|X_i] = \sqrt{\sigma^2} = \sigma$

→ Splitting Variance: ($\hat{Y}_i$)

↳ $\underbrace{\sum_{i=1}^{m}(Y_i-\bar{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{m}(\hat{\alpha}+\hat{\beta}x_i-\bar{Y})^2}_{\text{Explained variation of the regression = RSS}} + \underbrace{\sum_{i=1}^{m}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}_{\text{Random/residual variation around the regression = ESS}}$ ($\hat{Y}_i$)

Coefficient of determination:

↳ $R^2 = \dfrac{RSS}{TSS} = 1 - \dfrac{ESS}{TSS}$

↳ the proportion of the total variation in Y explained by the model.

↳ Unbiased estimator S^2 for estimate $\hat{\sigma}^2 = S^2 = \dfrac{ESS}{m-2} = MSE$ → $\hat{\sigma} = S = \sqrt{MSE}$

↳ least squares estimate of σ^2

↳ $SE[\hat{\beta}] = \sqrt{\dfrac{MSE}{\sum_{i=1}^{m}(x_i-\bar{x})^2}}$ and $SE[\hat{Y}_i] = S\cdot\sqrt{\dfrac{1}{m}+\left(\dfrac{x-\bar{x}}{S/SE[\hat{\beta}]}\right)^2}$

used for α ↵

→ Testing if there is a linear combination between X, Y :

↳ Hypotese testing for β and α :

↳ statistic model : $Y_i = \alpha + \beta X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma)$ for $i = 1, \dots, m$

↳ Test hypotesis : $H_0 := \beta = 0$ (or α) vs $H_1 := \beta \neq 0$ → 2-sided

↳ test statistic : $T = \dfrac{\hat{\beta}}{SE[\hat{\beta}]}$ where $SE[\hat{\beta}] = \sqrt{\dfrac{MSE}{\sum_{i=1}^{m}(x_i - \bar{x})^2}}$ and $df = m - 2$

↳ Reject H_0 : $|T| \geq t_{df, \frac{\alpha}{2}}$ or p-value $< \alpha$

↳ $(1-\alpha)100\%$ KI for $\beta = \left[\hat{\beta} \pm t_{\frac{\alpha}{2}, df} \cdot SE[\hat{\beta}]\right]$ → same for α

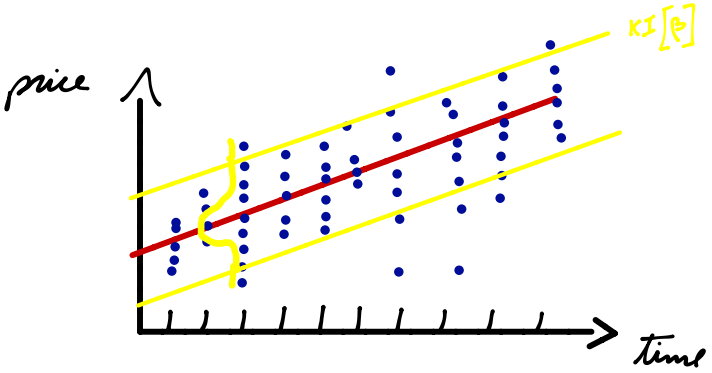
↳ $SE[E[Y|x]] = SE[\hat{Y}] = S \cdot \sqrt{\dfrac{1}{m} + \left(\dfrac{x - \bar{x}}{S/SE[\hat{\beta}]}\right)^2}$

↳ $(1-\alpha)100\%$ KI for $\hat{Y} = \left[\hat{Y} \pm t_{df, \frac{\alpha}{2}} \cdot SE[\hat{Y}]\right]$

→ Prediction Interval :

↳ $(1-\alpha)100\%$ PI for $\hat{Y} = \left[\hat{Y} \pm t_{df, \frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + \dfrac{1}{m} + \left(\dfrac{x_i - \bar{x}}{S/SE[\hat{\beta}]}\right)^2}\right]$

($\hat{Y} = \hat{\alpha} + \hat{\beta} x_i$; $\neq SE[\hat{Y}]$)



Chi squared test - Analysis of categorical variabels:

↳ Relationship between 2 categorical variabels by crosstabs or contingency tables;

Cat. / Obs.	Cat. 1	Cat. 2	Total
Obs. 1	ε_{11}	ε_{12}	R_1
Obs. 2	ε_{21}	ε_{22}	R_2
total	K_1	K_2	n

* Independent := $P(A|B) = P(A) \longrightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} \longrightarrow P(A \cap B) = P(A) \cdot P(B)$ //

* $X_{ij} \sim bin(n, p_{ij})$ where $p_{ij} = P(A_i \cap B_j)$ for X_{ij} independent

* $W \sim \chi^2_{(r-1)*(k-1)=df;\alpha}$

$H_0 := P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$ for all pair of i and j or $E[X_{ij}] = n \cdot P(A_i) \cdot P(B_j) = n \hat{p}_i \hat{p}_j$

$H_1 := P(A_i \cap B_j) \neq P(A_i) \cdot P(B_j)$ for all pair of i and j or $E[X_{ij}] \neq n \cdot P(A_i) \cdot P(B_j) \neq n \hat{p}_i \hat{p}_j$

where $E_{ij} = \frac{K_i K_j}{n}$

$$w = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(x_{ij} - \hat{\varepsilon}_{ij})^2}{\hat{\varepsilon}_{ij}}$$

Reject H_0 if $w \geq \chi^2_{\alpha;df}$

* Hypotesis: H_0 := columns and rows are independent Vs H_1 := columns and rows are dependent

* Test statistic: $w^* = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(X_{ij} - \hat{\varepsilon}_{ij})^2}{\hat{\varepsilon}_{ij}}$:= contribution of the combination (row/col) to test statistic;

↳ Where $E_{ij} = \frac{R_i \cdot K_i}{n}$:= expected number of the responses; where i := row; j := column

* Reject H_0 if $w^* > \chi^2_{df,\alpha}$ where $df = (r-1) \cdot (k-1)$

Dependency between variables:

↳ Covariance := $\sigma_{xy} = cov[X,Y] = E\left[(X-\mu_X)\cdot(Y-\mu_Y)\right] \rightarrow s_{xy} = \hat{\sigma}_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n-1}$

↳ Correlation := $\rho = corr[X,Y] = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$ for $-1 \leq \rho \leq 1 \rightarrow r = \rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{(n-1)}{(n-1)} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(x_i-\bar{x})^2 (y_i-\bar{y})^2}}$

↳ Remember: $\sigma_{xy} = \rho\,\sigma_x\,\sigma_y$ and $\sigma_{xy} = \rho\,\sigma^2$ if $\sigma_x^2 = \sigma_y^2$

Dependency between observations: $(X_1, \cdots, X_i, \cdots, X_n)$

↳ The stronger the correlation between variables, the weaker the new information will be!

①

✱ when we have dependent observations/variables:

↳ $Var[X_1+X_2] = Var[X_1] + Var[X_2] + 2\,cov(X_1,X_2)$

↳ $Var[X_1-X_2] = Var[X_1] + Var[X_2] - 2\,cov(X_1,X_2)$

✱ if both obs./variable have same distribution, they have the same variance.

↳ Then: $Cov(X_1,X_2) = Corr(X_1,X_2) \cdot SD[X_1] \cdot SD[X_2] = \rho \cdot \sigma^2$ feilaktig?

Dependency between pairwised data:

↳ $D_i = (X_i - Y_i)$ for $i = 1, \cdots, n$ and $D_1, \cdots, D_n$ independent and X_i and Y_i dependent.

↳ $Var[D_i] = Var[X_i - Y_i] = Var[X_i] + Var[Y_i] - 2\,cov(X_i, Y_i) = \underbrace{\sigma^2 + \sigma^2}_{\text{if both has same variance}} - 2\rho\sigma^2 = 2\sigma^2(1-\rho)$ //

↳ $Corr(X_i, Y_i) = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$

R code basics:

→ Chi squared: chisq.test(contingency_table, correct); qchisq(p, df);

chisq.test()$expected;

→ Variables: variable <- c(values); rep(value, times);

→ Tables: table(row_variable, column_variable); rownames(table) := get row names

→ Plots: barplot(table, beside, xlab, ylab, legend)

→ T-student := qt(p, df)

⊛ R: cov / cor (data[, c("D_1", "D_2")]) → check!

⊛ Contingency table:

cat. / obs.	Cat. 1	Cat. 2	Total
Obs. 1	ε_{11}	ε_{12}	R_1
Obs. 2	ε_{21}	ε_{22}	R_2
total	K_1	K_2	n

→ scatterplot(Y~X, reg.line = True)

→ lm(Y~X) := returns α and β

→ summary(lm(Y~X)) := returns all regression's parameters

→ anova_reg(lm(Y~X)) := returns analise of variance

→ confint(lm(Y~X), level = 0,99) := returns the confidence interval on a % level

R:
> View(dataset_name) := to view the entire dataset;
> ls(dataset_name) := to return the variables of the dataset;
> str(");
> table(" $variable_name) := freivens tabel
> hist(" ", breaks = seq(0, #final, # interval)) := histogram plot;
> mean(");
> abline(v= α, col='red') := horizontal line at x;
> summary(");
> sd(");
> boxplot(" ", horizontal = TRUE, add = TRUE) := boxplot with histogram;