

### → Frequency table:

- ↳ h: tabel (df\$variable)
- ↳ Python: df.value\_counts()

### → Histogram:

↳ relative frequency in an interval  
the width of the interval

↳ Forms: 1) Einzogspitze := Convoluted; 2) Totzogspitze := two-peaks; 3) Symmetric := Symmetry; 4) Skjæretet := Skewed.  
↳ 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  $< 0,5$ ;  $0,5 \leq \text{weak} \leq 1$ ; very  $> 1$

### → Relative frequency:

↳  $\frac{\# \text{ of observations in an interval}}{\text{total } \# \text{ of observations}}$

### → Sum properties:

$$\begin{aligned} 1) \sum_{i=1}^n (ax_i + b) &= a \sum_{i=1}^n x_i + mb \\ 2) \sum_{i=1}^n (x_i - b)^2 &= \sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i b + nb^2 = \sum_{i=1}^n x_i^2 - 2b \sum_{i=1}^n x_i + nb^2 \end{aligned}$$

R: > view(dataset\_name) := to view the entire dataset;

> ls(dataset\_name) := to return the variables of the dataset;

> str(" ");

> table("variable\_name") := frequency table

> hist(" ", breaks = seq(0, #final, # interval)) := histogram plot;

> mean(" ");

> abline(v = x, col = 'red') := horizontal line at x;

> summary(" ");

> sd(" ");

> boxplot(" ", horizontal = TRUE, add = TRUE) := boxplot with histogram;

↳ < 0 := right-skewed;  
↳ 0 := Gausse;  
↳ > 0 := pointed.

## 2) Shuffle rules:

### → Utfallsmønster

↳ Set of all possible 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  $1 - P(A) = P(\bar{A}) = P(A)^c = P(\bar{A})$

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

Remember:  $\begin{cases} m = \text{sample size}; \\ k = \text{picked units} \\ 0! = 1 \end{cases}$

returns := # of possible outcomes

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

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

$\hookrightarrow \mu_x = \bar{X} = E[X] = \sum_{\forall x \in S} x \cdot p(x=x)$  (diskret);  $\int x \cdot f(x) dx$  (kontinuierlich)

$\hookrightarrow$  Regressionsgesetze:  $E[a] = a$ ;  $E[a \cdot X] = a \cdot E[X]$ ;  $E[a+b \cdot X] = a+b \cdot E[X]$ ;  $E[a+b \cdot X + c \cdot x^2] = a+b \cdot E[X] + c \cdot E[X^2]$

$E[Y]$  where  $Y = g(x) \Rightarrow \sum_{\forall x \in S} g(x) \cdot p(x=x)$

Variansien:

$\hookrightarrow \sigma_x^2 = \text{Var}[X] = E[(x-\mu)^2] = \sum_x (x-\mu)^2 \cdot p(x=x) = (\sum x^2 \cdot p(x=x)) - \mu^2 = E[X^2] - E[X]^2$

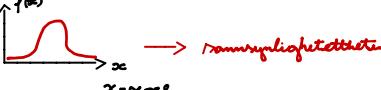
$\hookrightarrow$  Regressionsgesetze:  $\text{Var}[a] = 0$ ;  $\text{Var}[X] \geq 0$ ;  $\text{Var}[a+b \cdot X] = b^2 \cdot \text{Var}[X]$ ;  $\text{Var}[b \cdot X] = b^2 \cdot \text{Var}[X]$   
 $\text{Var}[a+b \cdot X] = b^2 \cdot \text{Var}[X]$ ;  $\text{Var}[\bar{X} - \bar{Y}] = \text{Var}[\bar{X}] + \text{Var}[\bar{Y}]$

Standardabweichung:

$\hookrightarrow \sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - \mu_x)^2}$  für Population

$\hookrightarrow s_x = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2}$  für unbekannt  $m$  und/or  $\sigma$ !

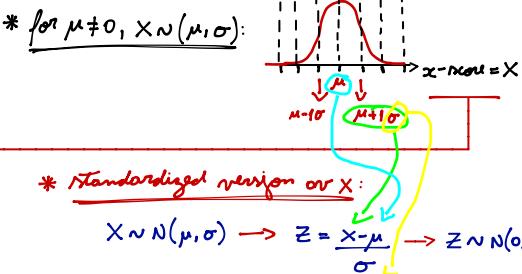
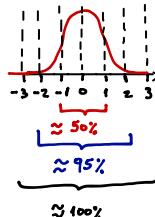
Normalverteilung:

$\hookrightarrow f(x) = g = \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$   → Densitätsfunktion

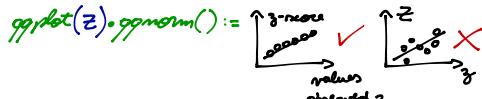
$\hookrightarrow$  Standardnormalverteilung  $N(\mu=0, \sigma=1)$ :

$\hookrightarrow$  Tabelle E3 - Kumulative Standardnormalverteilung:

| $z$                    | 0,00  | 0,01 | 0,02   | ... |
|------------------------|-------|------|--------|-----|
| $z = -1,02$            | -3,00 | ...  | ...    | ... |
| $P(Z \leq z) = 0,1539$ | -1,00 | ...  | 0,1539 | ... |
| $z = +3$               | ...   | ...  | ...    | ... |



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

ggplot(z) + pnorm(): 

\* Standardized version over  $X$ :  $X \sim N(\mu, \sigma) \rightarrow Z = \frac{X-\mu}{\sigma} \rightarrow Z \sim N(0,1)$

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

Linearkombination:

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

$\hookrightarrow$  Ex.:  $A \sim N(\mu_A, \sigma_A)$  und  $B \sim N(\mu_B, \sigma_B)$ , dann  $X = A+B$ . Wie ist  $\mu_X$  oder  $\sigma_X$ ?

$\hookrightarrow E[A+B] = E[A] + E[B] = \mu_A + \mu_B$

$\hookrightarrow \text{Var}[A+B] = E[(A+B)^2] - E[A+B]^2 = E[(\sigma_A)^2] + E[(\sigma_B)^2]$

$\hookrightarrow \text{SD}[A+B] = \sqrt{E[(\sigma_A)^2] + E[(\sigma_B)^2]}$

$$\left. \begin{array}{l} \mu_X = \mu_A + \mu_B \\ \sigma_X = \sqrt{E[(\sigma_A)^2] + E[(\sigma_B)^2]} \end{array} \right\}$$

Binomialverteilung:

$X \sim \text{bin}(\# \text{Fälle}, P(x=\text{succes}))$

$\hookrightarrow$  Ex.:  $X \sim \text{bin}(1000, 0,2)$

$\hookrightarrow E[X] = m \cdot p = 200$

$\hookrightarrow \text{Var}[X] = \sqrt{m \cdot p \cdot (1-p)} = 12,6$

$\text{bin}(m, p) \approx N(m \cdot p, \sqrt{m \cdot p \cdot (1-p)})$

\* Important rules:

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

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

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

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

\*  $\text{Cov}(X, X) = \text{Var}(X)$

\*  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

\*  $\text{Cov}(cX, Y) = c \cdot \text{Cov}(X, Y)$

\*  $\text{Cov}(X, Y+Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

### 3) Probability rules:

#### → Addition rule ('or'):

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

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

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

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

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

#### → Conditional rule:

$$\hookrightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\hookrightarrow P(\bar{A}|B) = 1 - P(A|B)$$

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

#### → Multiplication rule ('and'):

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

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

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

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

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

$$\hookrightarrow P((A \cap B)^c) = P(\bar{A} \cup \bar{B}) = P(A^c) + P(B^c) - P(A^c \cap B^c) \quad \textcircled{*}$$

#### → Bayes' rule:

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

$$\Rightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

|                         |                   | GGJJ                          | GJGJ | GJJG  | JGGJ  | GJGG | JGGG | JJGG | JJJG | JJJJ |  |
|-------------------------|-------------------|-------------------------------|------|-------|-------|------|------|------|------|------|--|
|                         |                   | GGGG                          | JGGG | JJGG  | JJJG  |      |      |      |      |      |  |
| Utfallsrom              | 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  |      |      |      |      |  |
| Scanned with CamScanner |                   |                               |      |       |       |      |      |      |      |      |  |
|                         |                   | $\frac{1}{16} + \frac{4}{16}$ |      |       |       |      |      |      |      |      |  |

#### → Cumulative distribution:

$$\hookrightarrow F(x) = P(X \leq x)$$

$$\hookrightarrow P(a < X \leq b) = F(b) - F(a)$$

$$\hookrightarrow P(X > a) = 1 - F(a)$$

$$\hookrightarrow P(X \leq b) = F(b)$$

$\Rightarrow P(x=0, y=0) = 0,09 := \text{joint probability}$

| x                       | y    |      |      |      |      | $P(X=x)$ |
|-------------------------|------|------|------|------|------|----------|
|                         | 0    | 1    | 2    | 3    | 4    |          |
| 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 | 0.01 | 0.02 | 0.01 | 0.05     |
| 7                       | 0.02 | 0.01 | 0.01 |      |      | 0.03     |
| $P(Y=y)$                |      | 0.34 | 0.35 | 0.19 | 0.09 | 0.03     |
| Scanned with CamScanner |      |      |      |      |      |          |

#### → joint distribution:

$$\hookrightarrow P(x,y) = P(x=x \text{ and } y=y) = P(x=x \cap y=y)$$

$\hookrightarrow P(x=x) \cdot P(y=y) \rightarrow x, y \text{ independent}$

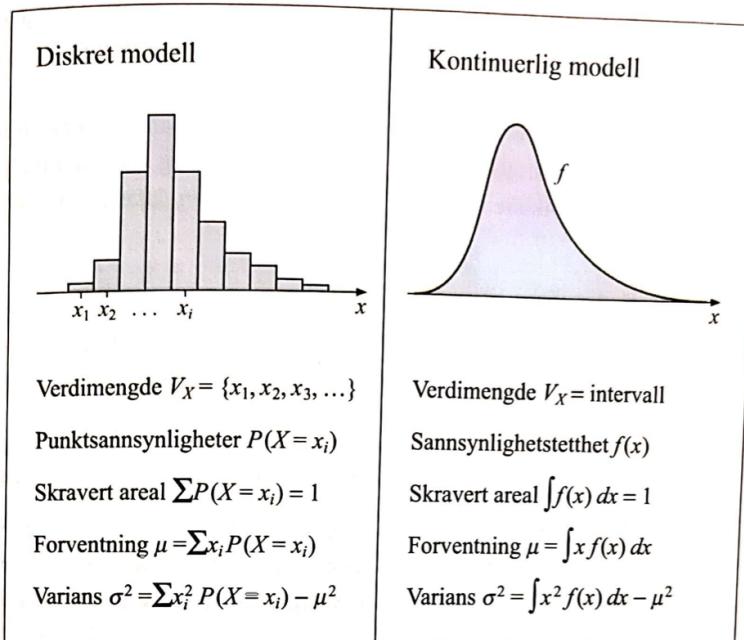
$\hookrightarrow P(x=x) \cdot P(y=y|x=x) \rightarrow x, y \text{ dependent}$

i.e.: marginal probability of Y

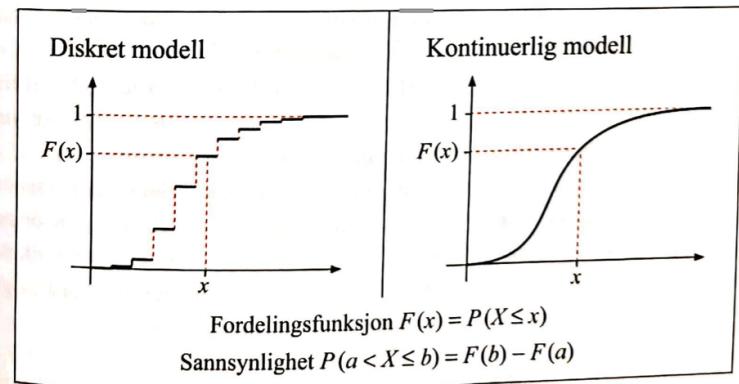
$$\hookrightarrow P(x+y=4) = 0,01 + 0,09 + 0,07 = 0,17$$

$$\hookrightarrow P(x=4|y=1) = \frac{P(x=4 \text{ and } y=1)}{P(y=1)} = \frac{0,03}{0,35} = 0,086$$

## Chaprer 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:

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

### → Varians:

$$\begin{aligned} \hookrightarrow \sigma^2 &= \text{Var}[\bar{X}] = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2 \\ \hookrightarrow E[(X-\mu)^2] &= E[X^2] - E[\mu]^2 \\ \hookrightarrow \text{Var}[bx+a] &= b^2 \text{Var}[X] \\ \hookrightarrow \sigma^2 &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu^2 \rightarrow \text{continuous} \\ \hookrightarrow \text{Var}[ax+by+c] &= a^2 \text{Var}[X] + b^2 \text{Var}[Y] + 2ab \text{Cov}[X, Y] \end{aligned}$$

only if  $X, Y$  dependent!

### → Correlation:

$$\begin{aligned} \hookrightarrow \rho(x, y) &= \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma_x \cdot \sigma_y} \\ \hookrightarrow -1 \leq \rho \leq 1 & \end{aligned}$$

### 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

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

Utnytt dette og vis at

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

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

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

### → Double expectation:

$$\hookrightarrow E[X] = \sum_{\forall y_i} P(Y=y_i) \cdot E[X|y_i] \quad (\text{side 165})$$

## Chapter 5: Probability models

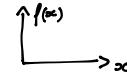
↳ Empirically := based on data; Theoretically := based on parameters

Stokastisk variable := utfallsvrom → Tall := siffror i ett utfallsvrom till en variabel. Exempel:  $X, Y, Z$ . Värdena de stokastiska variablarna:  $x, y, z$ .

Värdimängden :=  $V_X$  är mängden av alla värdena  $x$  som  $X$  kan ha.

Sannsynlighetsfördeling := förteller hur sannsynlig det är att  $X$  har de olika värdena i värdimängden.  $\rightarrow f(x) \rightarrow$

$P(X=x)$  := sannsynligheten för att den stokastiska variablen  $X$  har värde  $x$ .



\* Diskret variabel := tellevariabel;

↳ Punktsannsynlighet  $\rightarrow \sum_{x \in S} P(x=x) = 1$

↳ Tabeller med punktsannsynlighetens sannsynlighetsdiagram.

↳ Binomial fördeling:

↳  $X = \#$  nukleer på m försök

↳ Poisson-fördeling:

↳  $X = \#$  händelser i ett gitt område;  
Tidinterval, längde, volym.

↳ Geometrisk fördeling:

↳  $X = \#$  försök innan första nukleer.

\* Kontinuerlig variabel := målevariabel

↳ Sannsynlighetsstetthet  $\rightarrow \int_{x_1}^{x_2} f(x) dx = 1$

↳ Täthetskurva, tabeller med kumulativ sannsynligheten.

↳ Uniform fördeling; Normalfördeling; Exponentiellfördeling; Kvarnadratfördeling; F-fördeling:

↳  $\bar{X}$  = en mätning;

↳  $X$  = en funktion av flera mätningar.

$$V_x: 0 \ 1 \ 2 \ 3 \ 4 \ 5$$

$$P(x=\infty): \frac{1}{16} \ \frac{4}{16} \ \frac{6}{16} \ \frac{4}{16} \ \frac{1}{16} \ 1 \rightarrow \text{Punktsannsynligheten} := P(1 \leq X \leq 3) = P(X=2) + P(X=3)$$

$$P(x \leq \infty): \frac{1}{16} \ \frac{5}{16} \ \frac{9}{16} \ \frac{15}{16} \ \frac{14}{16} \ 1 \rightarrow \text{kumulativ fördeling} := P(X \leq 3) = P(X \leq 3) - P(X \leq 1)$$

→ 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;

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

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

↳  $E[X] = np$  and  $\text{Var}[X] = np(1-p)$ ;

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

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

→ Normal distribution:

↳  $X \sim \text{norm}(\mu, \sigma)$

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

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

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

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

→ Standard normal distribution:

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

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

$$P(Z > z_\alpha) = \alpha$$

→ Confidence Interval:

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

↳  $\oplus$  := larger than;  $\ominus$  := less than;

→ Central Limit theorem:

↳  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \sim \text{norm}\left(\mu, \frac{\sigma}{\sqrt{m}}\right)$  where  $X_i$  are independent and  $i=1, \dots, m$

↳ Condition:  $m \geq 30$

# CH. 6: Estimations:

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

→ Estimating of  $\mu$ :

$$\hookrightarrow \text{Unbiased estimator } \bar{X} \text{ for estimate } \hat{\mu} \rightarrow E[\hat{\mu}] = E[\bar{X}] = \mu$$

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

$$\hookrightarrow SE[\hat{\mu}] = SE[\bar{X}] = \frac{\sigma}{\sqrt{n}} \rightarrow \sigma \text{ known} \rightarrow \text{KI} := \left[ \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right] \rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$\hookrightarrow SE[\bar{X}] = \frac{s}{\sqrt{n}} \rightarrow \sigma \text{ unknown} \rightarrow \text{KI} := \left[ \bar{X} \pm t_{d.f., \alpha/2} \cdot \frac{s}{\sqrt{n}} \right] \rightarrow T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

→ Estimating of  $p$ :

$$\hookrightarrow \text{Unbiased estimator } \frac{\bar{X}}{m} \text{ for estimate } \hat{p} \rightarrow E[\hat{p}] = E\left[\frac{\bar{X}}{m}\right] = p$$

$$\hookrightarrow SE[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \rightarrow \text{KI} := \left[ \hat{p} \pm z_{\alpha/2} \cdot SE[\hat{p}] \right]$$

→ Estimating of  $\sigma^2$  and  $\sigma$ :

$$\hookrightarrow \text{Unbiased estimator } \sqrt{S^2} = S \text{ for estimate } \sigma \rightarrow E[\hat{\sigma}] = E[S] = \sigma$$

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

→ KI := page 251

Unbiased estimator Proof:

$$* E[\bar{X}] = \hat{\mu} \quad \text{Proof:} \quad E[\bar{X}] = E\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m} \sum_{i=1}^m E[X_i] = \frac{1}{m} \sum_{i=1}^m \mu = \frac{1}{m} \cdot m \cdot \mu = \hat{\mu} \quad \square$$

$$* E[S^2] = \hat{\sigma}^2 \quad \text{Proof:} \quad E[S^2] = E\left[\frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2\right] = \frac{1}{m-1} E\left[\sum_{i=1}^m (X_i - \bar{X})^2\right] \stackrel{\text{TKS3}}{=} \frac{1}{m-1} E\left[\sum_{i=1}^m X_i^2 - \bar{X}^2 \cdot m\right] = \frac{1}{m-1} \sum_{i=1}^m E[X_i^2] - m \cdot E[\bar{X}^2]$$

$$\hookrightarrow = \frac{1}{m-1} \sum_{i=1}^m (\sigma^2 + \mu^2) - m \cdot (\frac{\sigma^2 + m\mu^2}{m}) = \frac{1}{m-1} m\sigma^2 + m\mu^2 - \sigma^2 - \mu^2 m = \frac{(m-1)\sigma^2}{m-1} = \hat{\sigma}^2 \quad \square$$

Tricks 1:  $E[\bar{X}] = E\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m} \sum_{i=1}^m E[X_i] = \frac{1}{m} \sum_{i=1}^m \mu = \frac{1}{m} \cdot m \cdot \mu = \mu \quad \square$

Tricks 2:  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \leftrightarrow m \cdot \bar{X} = \sum_{i=1}^m X_i \quad \square \quad -2\bar{X} \cdot \sum_{i=1}^m X_i = -2\bar{X}m\bar{X} = -2m\bar{X}^2$

Tricks 3:  $\sum_{i=1}^m (X_i - \bar{X})^2 = \sum_{i=1}^m (X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^m X_i^2 - \sum_{i=1}^m 2X_i\bar{X} + \sum_{i=1}^m \bar{X}^2 = \sum_{i=1}^m X_i^2 - m\bar{X}^2 \quad \square$

Tricks 4:  $\text{Var}(X_i) = E(X_i^2) - E[X_i]^2 \leftrightarrow E(X_i^2) = \text{Var}(X_i) + E[X_i]^2 = \sigma^2 + \mu^2 \quad \square$

Tricks 5:  $\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{m} \sum_{i=1}^m X_i\right) = \frac{1}{m^2} \cdot \text{Var}\left(\sum_{i=1}^m X_i\right) = \frac{1}{m^2} \sum_{i=1}^m \text{Var}(X_i) = \frac{1}{m^2} \sum_{i=1}^m \sigma^2 = \frac{1}{m} \cdot m \sigma^2 = \frac{\sigma^2}{m} \quad \square$

Tricks 6:  $E[\bar{X}^2] = \text{Var}[\bar{X}] + E[\bar{X}]^2 = \frac{\sigma^2}{m} + \mu^2 \quad \square$

Tricks:

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

$$m \cdot \bar{X} = \sum_{i=1}^m X_i$$

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

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

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

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

only possible because  $X_i$  are independent!

# (Ch.6: Hypothesis testing):

|              | $H_0$ true             | $H_1$ true              |
|--------------|------------------------|-------------------------|
| Keep $H_0$   | good                   | $\beta(\text{error}^2)$ |
| Reject $H_0$ | $\alpha(\text{error})$ | good                    |

where  $\alpha$ : Significance level  $\rightarrow$  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:

$\hookrightarrow$  One group :=  $X_i \sim N(\mu, \sigma)$ ,  $i = 1, \dots, m$  OR  $X_i = \mu + \epsilon_i$ ,  $\epsilon_i \sim N(0, \sigma)$ ,  $i = 1, \dots, m$

$\hookrightarrow$  Two groups :=  $\begin{cases} \text{For } \sigma \text{ the same:} \\ \text{paired: } X_i \xrightarrow{\text{paired}} Y_i \end{cases}$   $\begin{cases} X_i \sim N(\mu_A, \sigma), i = 1, \dots, m_A \\ Y_i \sim N(\mu_B, \sigma), i = 1, \dots, m_B \end{cases}$  OR  $Y_i \sim N(\mu_D, \sigma), i = 1, \dots, m_A + m_B$  where  $\mu_D = \mu_A - \mu_B$

OR  $Y_i = \mu_D + \epsilon_i$ ,  $\epsilon_i \sim N(0, \sigma)$ ,  $i = 1, \dots, m_A + m_B$

## \* Hypothesis claims:

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

## \* Test statistic:

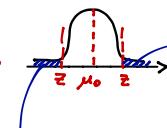
$\hookrightarrow Z\text{-test} := Z^* = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{m}}$ ;  $T\text{-test} := T^* = \frac{\bar{X} - \mu_0}{S / \sqrt{m}}$

## \* Rejection limits ( $k$ ):

$$\begin{aligned} P(Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{m}}) &= \alpha \\ &= \gamma^\alpha \text{ or } \gamma^{\frac{\alpha}{2}} \\ &\downarrow \quad \downarrow \\ \text{one-sided} & \quad \text{2-sided} \end{aligned}$$

$\hookrightarrow p\text{-test} := Z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{m}}}$  OR  $Z^* = \frac{X - mp_0}{mp_0(1-p_0)}$

## \* P-value:

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

$$P\text{-value} = P(Z \leq z_{\alpha}^* \mid H_0 \text{ is true})$$

$\hookrightarrow$  for one-sided  $H_1 := \mu_0 < 0$

$$\therefore P\text{-value} = P(|Z| \geq z_{\frac{\alpha}{2}}^* \mid H_0 \text{ is true})$$

$\hookrightarrow$  for 2-sided  $H_1 := \mu_0 \neq 0$

$$\begin{aligned} \Phi(z^*) &= P(Z \leq z^*) \rightarrow pt(\text{lower.tail} = \text{TRUE}) \\ 1 - \Phi(z^*) &= P(Z > z^*) \rightarrow pt(\text{lower.tail} = \text{FALSE}) \end{aligned}$$

## \* Rejection region:

$\hookrightarrow$  Reject  $H_0$  if:  $\begin{cases} Z^* \text{ or } T^* \geq z_{1-\alpha} \text{ or } t_{m-1, 1-\alpha} \rightarrow H_1 := \mu_0 > 0 \\ Z^* \text{ or } T^* \leq z_{\alpha} \text{ or } t_{m-1, \alpha} \rightarrow H_1 := \mu_0 < 0 \\ |Z^*| \text{ or } |T^*| \geq z_{\frac{\alpha}{2}} \text{ or } t_{m-1, \frac{\alpha}{2}} \rightarrow H_1 := \mu_0 \neq 0 \end{cases}$

$$\therefore \begin{cases} p\text{-value} > \alpha := \text{hold } H_0 \\ p\text{-value} \leq \alpha := \text{reject } H_0 \end{cases}$$

$\hookrightarrow$  Probability of obtaining a test statistic value at least as contradictory to the null hypothesis as the value that actually resulted.

$$P\text{-value} = \begin{cases} 1 - \Phi(z^*) & \rightarrow \text{calculated test statistic} \\ \Phi(z^*) & \rightarrow \text{for upper-tailed test} \rightarrow H_1 := \mu_0 > 0 \\ 2[1 - \Phi(|z^*|)] & \rightarrow \text{for lower-tailed test} \\ & \rightarrow \text{for two-tailed test} \end{cases}$$

\* Significance level  $\alpha$  := we are  $100(1-\alpha)\%$  sure that  $\mu_0$  is equal to 0.  $\rightarrow$  when  $H_0$  is true.

## 8.3: T-test for 2-groups:

### → 8.3.1: Not paired:

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  $j=1, \dots, m_2$

Hypothesis test:  $H_0: \mu_x - \mu_y = \mu_D = 0$  vs  $\mu_x > \mu_y$  or  $\mu_D > 0$

$$\mu_x < \mu_y \text{ or } \mu_D < 0$$

$$\mu_x \neq \mu_y \text{ or } \mu_D \neq 0$$

Estimations:

Unbiased estimator  $\bar{X}$  for  $\mu_x$  and  $\bar{Y}$  for  $\mu_y$ ;

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

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

$$\text{Standard error } SE[\bar{X} - \bar{Y}] = SE[\bar{D}] = S_p \cdot \sqrt{\frac{1}{m_x} + \frac{1}{m_y}}$$

$$S_p^2 = \frac{1}{m_x+m_y-2} \sum_{i=1}^{m_x+m_y} (X_i - \bar{X})^2$$

$$\text{Test statistic: } T = \frac{(\bar{X} - \bar{Y}) - \mu_D}{SE[\bar{X} - \bar{Y}]} = \frac{(\bar{X} - \bar{Y})}{S_p \sqrt{\frac{1}{m_x} + \frac{1}{m_y}}}$$

Rejection areas: → Reject  $H_0$  if:

|  |   |
|--|---|
| $T \geq t_{df, 1-\alpha}$ if $H_1: \mu_D > 0 \rightarrow \text{one-sided}$<br>$T \leq -t_{df, \alpha}$ if $H_1: \mu_D < 0 \rightarrow \text{one-sided}$<br>$ T  \geq t_{df, \frac{\alpha}{2}}$ if $H_1: \mu_D \neq 0 \rightarrow \text{two-sided}$ | <b>Remember!</b><br>* $df = m_1 + m_2 - 2$<br>* $100(1-\alpha)\%$ |
|--|---|

Confidence interval:  $[(\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}} \cdot SE[\bar{X} - \bar{Y}]]$

$p\text{-value} = P(T \geq T^* \mid H_0 \text{ is true}) =$  →  $\begin{cases} 1 - \Phi(T^*) \\ \Phi(T^*) \\ 2[1 - \Phi(T^*)] \end{cases}$

Reject  $H_0$  if:  $p\text{-value} < \alpha$

### → 8.3.2: Paired:

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

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

Hypothesis test:  $H_0: \mu_D = 0$  vs  $H_1: \mu_D \neq 0$

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

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

Test statistic:  $T = \frac{\bar{D} - \mu_D}{S/\sqrt{m}}$

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

# Variance Analysis : $\rightarrow$ for many groups

$\rightarrow$  One-way ANOVA:  $\rightarrow$  studies the effect of labeled populations on a single classification:

$\hookrightarrow k := \# \text{groups}$ ;  $m := \# \text{samples}$ ;  $Y_{ij} := \text{Observation } \# j \text{ from group } i$  (label of populations = factors);

$\hookrightarrow \text{Model} := Y_{ij} = \bar{Y}_i + \varepsilon_i$  where  $\varepsilon_i \sim N(0,1)$ ;  $i=1, \dots, k$  and  $j=1, \dots, m$  and  $\varepsilon_i$  independent

$\hookrightarrow \text{Hypothesis test: } H_0 := \mu_1 = \mu_2 = \dots = \mu_k \quad V_r \quad H_1 := \text{at least 2 of the } \mu_i \text{'s are different}$

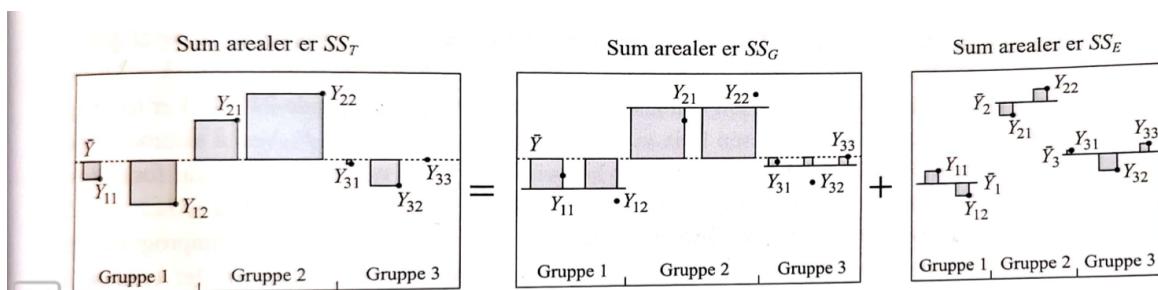
$\hookrightarrow \text{Estimations:}$

$\hookrightarrow \text{Individual sample mean} := \bar{Y}_i = \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} Y_{ij}; \text{Grand mean} := \bar{Y} = \frac{1}{k \cdot m} \cdot \sum_{i=1}^k \sum_{j=1}^m Y_{ij}$

$\hookrightarrow \text{Variances} := Y_{ij} = \bar{Y}_i + \varepsilon_i \rightarrow \varepsilon_i = Y_{ij} - \bar{Y}_i \rightarrow (Y_{ij} - \bar{Y}) = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$

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

$\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^m}}_{\text{Total sum of squares (TSS)}}$ 
 $\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^m}}_{\text{GSS}}$ 
 $\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^m}}_{\text{Error sum of squares (ESS)}}$ 
  
 $\Rightarrow m_i \cdot \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2$



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

$$\therefore \left[ \begin{array}{l} \rightarrow \text{Total Variance} := S_T^2 = \frac{TSS}{m-1} \quad \text{describer variation between groups} \\ \rightarrow \text{Variance between groups} := S_G^2 = \frac{GSS}{k-1} = MSG \\ \rightarrow \text{Variance within groups} := S_E^2 = \frac{ESS}{m-k} = MSE = \hat{\sigma}^2 \\ \therefore \hat{\sigma} = \sqrt{MSE} \end{array} \right] \quad \text{where } m = \sum_{i=1}^k m_i$$

$\hookrightarrow \text{Test statistic: } F^* = \frac{MSG}{MSE} = \frac{S_G^2}{S_E^2}$

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

\* Hold  $H_0$ :  $p\text{-value} > \alpha := \text{There is no significant difference between ...}$

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

Definition of  $\sigma$ : total variation over all groups?

proportion?

within groups?

# Variance Analysis With Contrast:

Model :=  $Y_{ij} = \mu_i + \epsilon_{ij}$  where  $\epsilon_{ij} \sim N(0, \sigma^2)$ , independent;  $i=1, \dots, k$  and  $j=1, \dots, m$

Contrast := linear combination of averages  $\rightarrow \theta = \sum_{i=1}^k c_i \mu_i$  where  $\sum_{i=1}^k c_i = 0$   
 ↳  $c$  will be chosen by what we are interested to test.

Test hypothesis := i.e.:  $\mu_1 \neq \mu_3 \Rightarrow 1\mu_1 + 0\mu_2 - 1\mu_3 = \theta = \mu_1 - \mu_3 \rightarrow H_0: \mu_1 - \mu_3 = 0 \vee H_1: \mu_1 - \mu_3 \neq 0$   
or:  $\mu_1 \neq \mu_2 \text{ and } \mu_3 \Rightarrow 1\mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = \mu_1 - \frac{\mu_2}{2} - \frac{\mu_3}{2} \rightarrow H_0: \mu_1 - \frac{\mu_2}{2} - \frac{\mu_3}{2} = 0 \vee H_1: \text{not all equals}$

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

Split variances :=  $Y_{ij} - \bar{Y} = \bar{Y}_i - \bar{Y} + Y_{ij} - \bar{Y}_i \rightarrow \sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y})^2 = \underbrace{\sum_{i=1}^k \sum_{j=1}^{m_i} (\bar{Y}_i - \bar{Y})^2}_{\text{TSS}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}_{\text{ESS}}$

$$\therefore \left\{ \begin{array}{l} \text{Total variance} := \frac{\text{TSS}}{m-1} = S_T^2 \\ \text{Treatment variance (between groups)} := MSG = \frac{\text{TSS}}{k-1} = S_{\text{Tn}}^2 \\ \text{Residual / Error variance (within groups)} := MSE = \frac{\text{ESS}}{m-k} = S_E^2 \end{array} \right. \quad \text{for } m = \sum_{i=1}^k m_i$$

Test statistic :=  $T^* = \frac{\hat{\theta} - \theta_0}{\text{SE}[\hat{\theta}]}$   $df = \frac{\alpha}{2}, m-k$

## Model assumption:

$\epsilon_{ij}$  := residual  $\rightarrow X_{ij} = \mu_i + \epsilon_{ij} \rightarrow \epsilon_{ij} = X_{ij} - \mu_i \rightarrow \text{for } \hat{\mu}_i = \bar{X}_i \rightarrow \epsilon_{ij} = X_{ij} - \bar{X}_i$

↳ Independent  $\epsilon_{ij} \sim \epsilon_j$   $\rightarrow$  correlations

↳ Normal distributed  $\epsilon_{ij} \sim N(0, \sigma^2)$   $\rightarrow$  calculate each residual  $\epsilon_{ij} = X_{ij} - \bar{X}_i$  and see the plot or median vs mean;

↳ Constant variance  $\rightarrow$  check standard 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 combinations;

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

↳ Condition = numerical variable

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

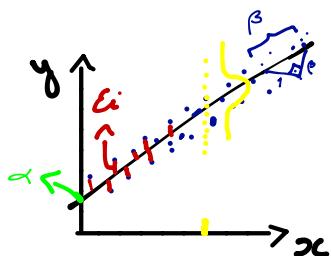
# Linear Regression:

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

→ Statistical Model:

↪  $Y_i = \alpha + \beta x_i + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$  for  $i = 1, \dots, n$ ;  $\rightarrow Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

↪  $X_i$ s are independent variable;  $Y_i$ s are dependent variable;



Where:  $\left[ \begin{array}{l} \alpha := \text{Expected value of the response variable when the explanatory variable is zero;} \\ \beta := \text{Expected change of the response variable when the explanatory variable increases 1 unit;} \end{array} \right]$

→ Estimation of parameters:

↪ Unknown parameters:  $\alpha, \beta, \sigma^2 \rightarrow$  describe the variation of the response with a fixed value of  $x$ .

↪ least squares method:

$$\hookrightarrow K = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 \rightarrow$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = n \frac{s_y}{s_x}$$

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

$$\hat{Y}_i, \hat{\sigma}^2 \text{ and } \hat{\sigma} := \left[ \begin{array}{l} E[Y_i | X_i] = E[\alpha + \beta x_i + \epsilon_i] = E[\alpha] + E[\beta x_i] + E[\epsilon_i] = \alpha + \beta x_i = \hat{Y}_i \\ \text{Var}[Y_i | X_i] = \text{Var}[\alpha + \beta x_i + \epsilon_i] = 0 + 0 + \text{Var}[\epsilon_i] = \sigma^2 \\ \text{SD}[Y_i | X_i] = \sqrt{\sigma^2} = \sigma \end{array} \right] *$$

→ Splitting Variance:

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

$\text{TSS}$                                      $\text{RSS}$

Coefficient of determination:

$$\hookrightarrow R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

↪ the proportion of the total variation in  $Y$  explained by the model.

↪ Unbiased estimator  $S^2$  for estimate  $\hat{\sigma}^2 = S^2 = \frac{ESS}{m-2} = MSE \rightarrow \hat{\sigma} = S = \sqrt{MSE}$

↪ least squares estimate of  $\sigma^2$

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

normal for  $\alpha \downarrow$

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

↳ Hypothesis testing for  $\beta$  and  $\alpha$ :

$H_0: \beta > 0 \rightarrow$  There is a positive linear comb.  
 $\beta < 0 \rightarrow$  There is a negative linear comb.  
 $\beta = 0 \rightarrow$  There is lin. comb.

↳ Statistic model:  $Y_i = \alpha + \beta X_i + \epsilon_i$ ,  $\epsilon_i \sim N(0, \sigma^2)$  for  $i = 1, \dots, n$

↳ Test hypothesis:  $H_0: \beta = 0$  vs  $H_1: \beta \neq 0 \rightarrow$  2-sided

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

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

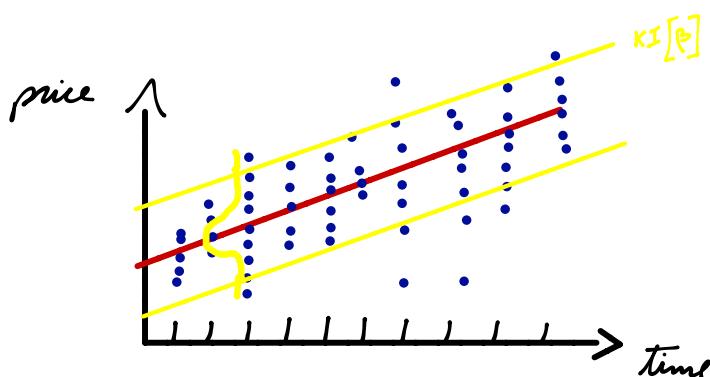
↳  $(1-\alpha)100\%$  KI for  $\beta = [\hat{\beta} \pm t_{df, \frac{\alpha}{2}} \cdot SE[\hat{\beta}]] \rightarrow$  same for  $\alpha$

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

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

→ Prediction Interval: → It is  $(1-\alpha)100\%$  likely that the true  $Y$  of random something with average  $x$  will be in that range!

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



## Chi-squared test - Analysis of categorical variables:

↳ Relationships between 2 categorical variables by crosstabs or contingency tables;

| Row<br>Category | Col. 1   | Col. 2   | Total |
|-----------------|----------|----------|-------|
| Obs. 1          | $E_{11}$ | $E_{12}$ | $R_1$ |
| Obs. 2          | $E_{21}$ | $E_{22}$ | $R_2$ |
| Total           | $K_1$    | $K_2$    | $n$   |

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

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

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

$H_0 := P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$  for all pair of  $i$  and  $j$  or  $E[X_{ij}] = n \cdot P(A_i) \cdot P(B_j) = n \hat{p}_i \hat{p}_j$  where  $E_{ij} = \frac{R_i K_j}{n}$   
 $H_1 := P(A_i \cap B_j) \neq P(A_i) \cdot P(B_j)$  for all pair of  $i$  and  $j$  or  $E[X_{ij}] \neq n \cdot P(A_i) \cdot P(B_j) \neq n \hat{p}_i \hat{p}_j$

$W = \sum_{i=1}^n \sum_{j=1}^k \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$  Reject  $H_0$  if  $W \geq \chi^2_{\alpha, df}$

\* Hypothesis :  $H_0$  := columns and rows are independent vs  $H_1$  := columns and rows are dependent

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

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

\* Reject  $H_0$  if  $W^* > \chi^2_{df, \alpha}$  where  $df = (n-1)(k-1)$  or p-value  $< \alpha$

↳  $H_0$  is true if  $\chi^2$ -statistic is near 0. Conversely,  $H_0$  is false if  $\chi^2$ -statistic is far from 0.

↳ always upper-tailed with  $\chi^2_{df, \alpha}$

## Dependency between variables:

$$\hookrightarrow \text{Covariance} := \sigma_{xy} = \text{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)] \rightarrow s_{xy} = \hat{\sigma}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

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

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

## Dependency between observations:

$\hookrightarrow$  The stronger the correlation between variables, the weaker the new information will be!

①

\* When we have dependent observations/variables:

$$\hookrightarrow \text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2] + 2 \text{cov}(X_1, X_2)$$

$$\hookrightarrow \text{Var}[X_1 - X_2] = \text{Var}[X_1] + \text{Var}[X_2] - 2 \text{cov}(X_1, X_2)$$

\* If both obs./variables have some distribution, they have the same variance.

$$\hookrightarrow \text{Then: } \text{cov}(X_1, X_2) = \text{cor}(X_1, X_2) \cdot \text{SD}[X_1] \cdot \text{SD}[X_2] = \rho \cdot \sigma^2$$

fehlaktig?

## Dependency between paired data:

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

$$\hookrightarrow \text{Var}[D_i] = \text{Var}[X_i - Y_i] = \text{Var}[X_i] + \text{Var}[Y_i] - 2 \text{cov}(X_i, Y_i) = \underbrace{\sigma_x^2 + \sigma_y^2}_{=} - 2\rho\sigma_x\sigma_y = 2\sigma^2(1-\rho)$$

$$\hookrightarrow \text{cor}(X_i, Y_i) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{if both have same variance}$$

## R code boxes:

→ Chi squared: `chisq.test(contingency-table, correct)`; `getchisq(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: = `gt(p, df)`

## \* Contingency table:

| col<br>row | Col. 1   | Col. 2   | Total |
|------------|----------|----------|-------|
| Obs. 1     | $E_{11}$ | $E_{12}$ | $R_1$ |
| Obs. 2     | $E_{21}$ | $E_{22}$ | $R_2$ |
| Total      | $K_1$    | $K_2$    | $n$   |

## \* R: $\frac{\text{cov}}{\text{cor}}$ (data[, c("D1", "D2")]) → check!

→ `scatterplot(Y~X, reg.line = True)`

→ `lm(Y~X)` := returns  $\alpha$  and  $\beta$

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

→ `anova-reg(lm(Y~X))` := returns analysis 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")` := previews table

`> hist(" ", breaks = seq(0, #final, # interval))` := histogram plot;

`> mean()`;

`> abline(v =  $\alpha$ , col = 'red')` := horizontal line at  $\alpha$ ;

`> summary()`;

`> sd()`;

`> boxplot(" ", horizontal = TRUE, add = TRUE)` := boxplot with histogram;