Afsnit 2.1-2.7

Hvad er statistik?

- Indsamling af data.
- Statistik handler ofte om at analysere en stikprøve, der er taget fra en population.
- Baseret på stikprøven, prøver vi at generalisere (eller udtale os) om populationen.

Nøgletal

- **Middelværdi** angiver tyngdepunkt eller centrering af data:
$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$
  - Eks. Har vi tallene: 12, 15, 13, 14, 16
  Middelværdien bliver: $\bar{x} = \frac{1}{5}(12+15+13+14+16) = 14$
- **Median** angiver tyngdepunkt eller centrering af data. I nogle tilfælde, f.eks. hvis man har ekstreme værdier, er medianen at fortrække frem for middelværdien:
Først skal antal n sættes i rækkefølge, hvis:
  - Ulige antal n er tallet i midten medianen.
  - Lige antal n, tages de to tal i midten ligger dem sammen og deler med 2 = madian.
- **Varians** (eller **standardafvigelsen**) siger noget om hvor maget data spreder:
  - Varians: $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - Eller varians: $\boxed{s^2 = \frac{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n \cdot (n-1)}}$
    - Eks. Har vi tallene: 12, 15, 13, 14, 16 (samme tal som i middelværdi eks. Så $\bar{x}$ er den samme).
    Variansen bliver:
$$s^2 = \frac{1}{5-1}\left((12-14)^2 + (15-14)^2 + (13-14)^2 + (14-14)^2 + (16-14)^2\right) = 2{,}5$$
  - Standardafvigelse (**spredning**): $\boxed{s = \sqrt{s^2}}$
- Vigtigt: $\bar{x}$ og s er estimerede værdier dvs. at hvis man tager en stikprøve ud af en population og beregner middelværdien og spredningen er det estimerede. μ og σ gælder for hele populationen.
- **Variationskoefficient** bruges til at sammenligne variationen mellem forskellige datasæt:
$$\boxed{V = \frac{s}{\bar{x}} \cdot 100}$$

- **Fraktiler** er punkter hvor data deles. Medianen deler data i to halvdele. Fraktiler deler data i andre dele. Ofte beregner man fraktiler:

  0, 25, 50, 75, 100 % fraktiler.

Figurer

- **Pareto diagram** siger f.eks. noget om hvor forskellige slags defekte der er i et givet system.
- **Dot diagram** er godt til at detektere fejl ved at se på outliers prikker som ligger usædvanligt.
- **Frequency distribution**: opdeling i intervaller/klasser og optælle herefter.
- **Histogram**: se side 19. god til grove data med mange tal.
- **Boxplot**: en rektangel der repræsentere midten af data og en linje repræsentere medianen. De to linjer på siderne af rektanglen repræsentere 95% og 5%.

Afsnit 4.1-4.4 og 4.6 og 4.7

En grundregel

Den klassiske sandsynlighedskoncept defineres:
Hvis der findes n lige sandsynlige udfald, hvorfra et må ske, og hændelsen s betegnes som 'succes', så er sandsynligheden for succes givet ved:

$$\boxed{\frac{s}{n}}$$

Binomialkoefficienten

Det antal forskellige måder som vi kan udvælge r objekter taget ud fra en population bestående af n forskellige objekter er:

$$\boxed{\binom{n}{r} = \frac{n!}{r!(n-r)!}}$$

Hvad er Stokastisk variable

- En funktion defineret over udfaldsrummets elementer.
- Følger en statistisk fordeling.
- Stokastisk variable angives ved store bogstaver, f.eks. X, Y, Z.
- Udfaldet fra det stokastiske variable angives ved tilsvarende små bogstaver, f.eks. x, y,z.
- Vi skelner mellem <u>diskrete</u> og <u>kontinuerte</u> stokastiske variable.

Tæthedsfunktion for diskret variabel

- For en stokastisk variabel betegnes tæthedsfunktionen ved f(x).
- For den diskrete variabel kan vi skrive tæthedsfunktionen som:

$$\boxed{f(x) = P(X = x)}$$

Fordelingsfunktion for diskret variabel

- Fordelingsfunktionen for en stokastisk variabel betegnes ved F(x).
- Fordelingsfunktionen svarer til den kumulerede tæthedsfunktion:

$$\boxed{F(x) = P(X \le x)}$$

Diskrete fordelinger

Binomial fordeling

- Vi betragter n uafhængige forsøg.
- I hvert enkelt forsøg kan udfaldet/hændelsen blive enten succes eller fiasko.
- Sandsynligheden for succes er p (og er den samme for alle n forsøg).
- Sandsynligheden for fiasko er dermed 1-p (og er den samme for alle n forsøg).
- De forskellige udfald er uafhængige.
- <u>Med tilbage lægning</u>.
- En stokastisk variabel, X, er binomial fordelt:

$$\boxed{X \approx b(x; n, p)}$$

X = antal "mærkede" i stikprøven.
p = populationsandelen = a/n, hvor a er i alt defekte.
n = stikprøvestørrelsen.
Det er lille b hvis det er præcis en mængde og store B hvis det er større eller mindre end en mængde.

- Tæthedsfunktion for en binomial fordeling:

$$\boxed{f(x) = P(X = x) = \binom{n}{x} \cdot p^x (1-p)^{n-x}}$$

$\boxed{P(X = x) = P(X \le x) - P(X \le x-1)}$, tabel 1 side 576.

- Fordelingsfunktion for binomal fordeling:

$\boxed{F(x) = P(X \le x)}$, tabel 1 side 576.

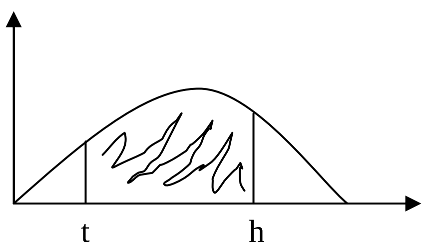
$$\boxed{P(X < x) = P(X \le x-1)}$$

$$\boxed{P(X > x) = 1 - P(X \le x-1)}$$

MOST (højst) brug:	$P(X \le udfald)$ **, direkte ved opslag tabel 1.**
MORE THAN (mere end):	$P(X > udfald) = 1 - P(X \le udfald)$
LEAST (mindst) brug:	$P(X \ge udfald) = 1 - P(X \le udfald - 1)$
LESS THAN (mindre end):	$P(X < udfald) = P(X \le udfald - 1)$

- **Middelværdi:** $\boxed{\mu = n \cdot p}$
- **Varians:** $\boxed{\sigma^2 = np \cdot (1-p)}$
- Hvis man ønsker at finde sandsynligheden for et bestemt område:

$\boxed{\text{B(h;n,p)-B(t;n,p)}}$



Den Hypergeometrisk fordeling

- En population med størrelse N.
- En stikprøve af størrelse n.
- Der er a defekte i populationen.

- Der er N-a ikke-defekte i populationen.
- x er antal defekte ud af stikprøven.
- <u>Uden tilbage lægning.</u>
- Den stokastiske variabel, X, er hypergeometrisk fordelt:

$$\boxed{X \approx h(x;n,a,N)}$$

- Tæthedsfunktion for den hypergeometriske fordeling:

$$\boxed{f(x) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}}$$

- Den hypergeometriske fordeling kan udskiftes med binomial fordelingen hvis populationen N er stor og stikprøven n er lille.

Obs! Binomial fordeling kan til forveksling bruges i tilfælde hvor n ikke er så lille i forhold til N også kunne man begå den fejl at anvende binomial distribution med n og p = a/N. se side 111.

- **Middelværdi:** $\boxed{\mu = n \cdot \frac{a}{N}}$
- **Varians:** $\boxed{\sigma^2 = n\frac{a}{N}\left(1-\frac{a}{N}\right)\left(\frac{N-n}{N-1}\right)}$

Poisson fordeling

- Poisson fordeling anvendes ofte som en fordeling (model) for tælletal, hvor der ikke er nogen naturlig øvre grænse.
- Poisson fordelingen kan ofte karakteriseres som intensitet, dvs. på formen antal/enhed.
- Parameteren λ angiver intensiteten i poisson fordelingen.
- Når n er stor og p er lille er binomial sansynligheder approksimeret til possion distribution.
- Possion fordeling anvendes til approksimation af binomiale sansynligheder, når $n \geq 20 \;\; og \;\; p \leq 0.05$, hvis $n \geq 100 \;\; er \; approksimationen \; god \; så \; længe \; np \leq 10$
- Se s. 129 for sammenligning af possion og binomial.
- Den stokastiske variabel, X, er possion fordelt:

$$\boxed{X \approx P(\lambda)}$$

- Tæthedsfunktionen:

$$\boxed{f(x) = P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}}$$

- Fordelingsfunktionen:

$\boxed{F(x) = P(X \leq x)}$, tabel 2, side 581

MOST (højst) brug:	$P(X \leq udfald)$, **direkte ved opslag tabel 2.**
MORE THEN (mere end):	$P(X > udfald) = 1 - P(X \leq udfald)$
LEAST (mindst) brug:	$P(X \geq udfald) = 1 - P(X \leq udfald - 1)$
LESS THEN (mindre end):	$P(X < udfald) = P(X \leq udfald - 1)$

- **Middelværdi:** $\boxed{\mu = \lambda}$
- **Varians:** $\boxed{\sigma^2 = \lambda}$

Middelværdi og varians for en diskret stokastisk variabel (overordnet)

- Middelværdi: $\boxed{\mu = \sum_S x \cdot f(x)}$, hvor S er udfaldsrummet for X.
- Det vides at: $\boxed{\sum_S f(x) = 1}$
- Varians: $\boxed{\sigma^2 = \sum_S (x-\mu)^2 \cdot f(x)}$, hvor S er udfaldsrummet for X.

Afsnit 5.1 og 5.2-5.6 + Afsnit 5.7,5.10,5.11 og 5.12

Tæthedsfunktionen for kontinuert variabel

- Tæthedsfunktionen betegnes f(x).
- f(x) siger noget om den relative hyppighed af udfaldet x for den stokastiske variabel X.
- For kontinuerte variable svarer tætheden ikke til sandsynligheden, dvs:

$$\boxed{f(x) \neq P(X = x)}$$

Fordelingsfunktion for kontinuert variabel

- Fordelingsfunktionen betegnes ved F(x).
- Fordelingsfunktionen svarer til den kumulerede tæthedsfunktion:

$$\boxed{F(x) = P(X \leq x)}$$

Kontinuerte fordelinger

Normal fordeling og Standardiseret normal fordeling

- Der kan ikke opstilles generelle kriterier for, hvornår en variabel er normailfordelt.
- Ofte kan man ramme rigtigt, hvis man til hvert element stiller spørgsmålet: ”hvilken værdi har elementet” og svarmuligheden er ”et tal”.

$$\boxed{X \approx N(\mu, \sigma^2)}$$

- Tæthedsfunktionen:

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

$\boxed{P(X < x)}$, aflæses i tabel 3, side 585

$$\boxed{P(X > x) = 1 - P(X < x)}$$

$$\boxed{P(a \leq X \leq b) = \int_a^b f(x)dx}$$

- **Middelværdi:** $\boxed{\mu = \mu}$
- **Varians:** $\boxed{\sigma^2 = \sigma^2}$
- En normal fordeling med middelværdien 0 og variansen 1, dvs. $\boxed{X \approx N(0,1^2)}$, kaldes en standard normal fordeling.
- En vilkårlig normal fordelt variabel $\boxed{X \approx N(\mu, \sigma^2)}$ kan standardiseres ved at beregne:

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

- Fordelingsfunktionen:

  $\boxed{F(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{z} e^{-t^2/2}\, dt}$, kan findes i tabel 3, side 585.

  $\boxed{P(X < z)}$, aflæses i tabel 3, side 585

  $\boxed{P(X > z) = 1 - P(X < z)}$

  Less (mindre end): $\boxed{P(X < a) = F\left(\frac{a-\mu}{\sigma}\right)}$

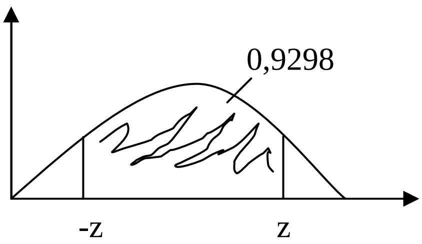
  More (større end): $\boxed{P(X > a) = 1 - F\left(\frac{a-\mu}{\sigma}\right)}$

  Between (imellem): $\boxed{P(a < X < b) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)}$

  F(Z) aflæses i tabel 3, side 585

- **Eksempel for standard normal fordeling:**

  P(-z < X < z) = 0,9298
  z = 1-0,9298/2



Log-Normal fordeling

- Log-normal fordelingen benyttes når vi har en tilfældig variable, som er på den måde at hvis man tager ln til den giver det normal distribution:

  $$\boxed{X \approx LN(\alpha, \beta^2)}$$

- Tæthedsfunktion:

  $$\boxed{f(x) = \frac{1}{\beta \cdot \sqrt{2\pi}} \cdot x^{-1} \cdot e^{-(\ln(x)-\alpha)^2/2\beta^2}}$$

- **Middelværdi:** $\boxed{\mu = e^{\alpha+\beta^2/2}}$

- **Varians:** $\boxed{\sigma^2 = e^{2\alpha/\beta^2}(e^{\beta^2} - 1)}$

- En log-normal fordelt variabel $\boxed{X \approx LN(\alpha, \beta^2)}$, kan transformeres til en standard normal fordelt variabel Z ved:

  $$\boxed{Z = \frac{\ln(X) - \alpha}{\beta}}$$

- til at finde sandsynligheden (imellem a og b):

$$\boxed{P(a < X < b) = \int_{\ln a}^{\ln b} \frac{1}{\sqrt{2\cdot\pi\beta}} e^{-(y-\alpha)^2/2\beta^2} dy = F\left(\frac{\ln b - \alpha}{\beta}\right) - F\left(\frac{\ln a - \alpha}{\beta}\right)}$$

$$\boxed{P(X < a) = F\left(\frac{\ln a - \alpha}{\beta}\right)}$$

$$\boxed{P(X > a) = 1 - F\left(\frac{\ln a - \alpha}{\beta}\right)}$$

Tabel 3 s.585

Uniform fordeling

- $\boxed{X \approx U(\alpha, \beta)}$
- Tæthedsfunktionen:

$$\boxed{f(x) = \frac{1}{\beta - \alpha}}$$

- Fordelingsfunktionen:

$$\boxed{F(x) = \int_b^a \frac{1}{\beta - \alpha} dx = \frac{a - b}{\alpha - \beta}}$$

- **Middelværdi:** $\boxed{\mu = \frac{\alpha + \beta}{2}}$
- **Varians:** $\boxed{\sigma^2 = \frac{1}{12}(\beta - \alpha)^2}$

Middelværdi og varians af en kontinuert stokastisk variabel (overordnet)

- Middelværdi: $\boxed{\mu = \int_S x \cdot f(x) dx}$, hvor S er udfaldsrummet for X.
- Varians: $\boxed{\sigma^2 = \int_S (x - \mu)^2 \cdot f(x) dx}$, hvor S er udfaldsrummet for X.

Eksponentiel fordeling

- Tæthedsfunktionen:

$$\boxed{f(x) = \frac{1}{\beta} e^{-x/\beta}}$$

- Fordelingsfunktionen:

$$\boxed{F(x) = \int_0^x \frac{1}{\beta} e^{-x/\beta} dx = 1 - e^{-x/\beta}}$$

$$\boxed{P(X < x) = F(x) = 1 - e^{-x/\beta}}$$

$$\boxed{P(X > x) = 1 - F(x) = 1 - (1 - e^{-x/\beta})}$$

- Eksponential fordeling er et special tilfælde af Gamma fordeling (α=1).
- Eksponential fordelingen anvendes f.eks. til at beskrive levetider og ventetider.
- Eksponential fordelingen anvendes f.eks. til at beskrive (vente)tiden mellem hændelser i poisson fordelingen. $\boxed{\lambda = \beta}$
- Middelværdi: $\boxed{\mu = \beta}$.

- Varians: $\boxed{\sigma^2 = \beta^2}$.

Regler for stokastisk variabel (eksempler side 186)

Vi antager at a og b er konstanter og X er en stokastisk variabel:

- E = middelværdi: $\boxed{E(aX+b) = aE(X)+b}$
- Var = Varians: $\boxed{Var(aX+b) = a^2 Var(x)}$

Følgende linear kombination gælder:

- $\boxed{E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n)}$
  $\boxed{E(X_1 - X_2) = E(X_1) - E(X_2)}$
  $\boxed{E(aX_1 - bX_2 + c) = aE(X_1) - bE(X_2) + c}$
- $\boxed{Var(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1^2Var(X_1) + a_2^2Var(X_2) + ... + a_n^2Var(X_n)}$
  $\boxed{Var(X_1 - X_2) = Var(X_1) + Var(X_2)}$, læg mærke til at der ændres her til plus.
  $\boxed{Var(aX_1 - bX_2 + c) = a^2Var(X_1) + b^2Var(X_2)}$, læg mærke til at der ændres her til plus.

Transformation

Såfremt data afviger fra at være normalt fordelt, kan man ofte med fordel transformere data, således at de transformerede data kan antages at være normal fordelt.

Afsnit 7.1-7.2, 6.1, 6.2 og 6.3

Stikprøvefordelinger

Definition af population og tilfældig stikprøve

- Tilfældig stikprøve fra en endelig population:
  Observationerne X_1, X_2,…,X_n er en tilfældig stikprøve af størrelse n fra en endelig population af størrelse N, såfremt værdierne er valgt således, at enhver delmængde af størrelse n af de N elementer fra populationen har den samme sandsynlighed for at blive valgt.
- Tilfældig stikprøve fra en uendelig population:
  Et sæt observationer X_1, X_2,…,X_n er en tilfældig stikprøve af størrelsen n fra en uendelig population f(x) såfremt:
  1. hvert X_i er en stokastisk variabel med tæthedsfunktionen f(x).
  2. De n stokastiske variable er uafhængige.

Stikprøvefordeling for middelværdien når variansen er kendt

- Uendelig population:
  Lad $\overline{X}$ være middelværdien af en stikprøve af størrelse n fra en fordeling med middelværdi μ og variansen σ^2.
  Da er $\overline{X}$ en stokastisk variabel og følger en fordeling med middelværdi μ og variansen σ^2/n.
- Endelig population:
  Lad $\overline{X}$ være middelværdien af en stikprøve af størrelse n fra en fordeling med middelværdi μ og variansen σ^2.

Da er $\overline{X}$ en stokastisk variabel og følger en fordeling med middelværdi μ og variansen $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$.

Estimation

Begreber

- Central estimator:
  En estimator $\hat{\theta}$ er central (eller ikke-biased), hvis og kun hvis, middelværdien af stikprøvefordelingen for estimatoren er lig θ.
- Efficient estimator:
  En estimator $\hat{\theta}_1$ er en mere efficient estimator af θ end estimatoren $\hat{\theta}_2$ hvis:
  1. $\hat{\theta}_1$ og $\hat{\theta}_2$ begge er centrale estimatorer af θ.
  2. variansen af stikprøvefordelingen for $\hat{\theta}_1$ er mindre end for $\hat{\theta}_2$.

Den centrale grænseværdisætning

Lad $\overline{X}$ være middelværdien af en stikprøve af størrelse n fra en fordeling med median (mean) μ og variansen σ^2, da vil:

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}}$$

Følge en $N(0,1^2)$ fordeling for $n \to \infty$.

Maksimal fejl på et estimat hvor variansen er kendt

For store værdier af n gælder:

$$\boxed{Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}}$$

Den maksimale fejl, E, på et estimat med sandsynlighed $\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$ bliver:

$\boxed{E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}$, hvor $z_{\alpha/2}$ findes i tabel 3.

To eksempler for at finde $Z_{\alpha/2}$:

$$\boxed{\begin{aligned} &\alpha = 0{,}95 \\ &\alpha/2 = \tfrac{1-0{,}95}{2} = 0{,}025 \\ &z_{\alpha/2} = z_{0{,}025} = 1{,}96 \end{aligned}}$$

$$\boxed{\begin{aligned} &\alpha = 0{,}99 \\ &\alpha/2 = \tfrac{1-0{,}99}{2} = 0{,}005 \\ &z_{\alpha/2} = z_{0{,}005} = 2{,}575 \end{aligned}}$$

Værdierne 1,96 og 2,575 blev fundet i tabel 3 s.585-586

Hvis E er kendt kan stikprøvestørrelsen n findes ved:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

$$n = \frac{1}{4} \left(\frac{z_{\alpha/2}}{E} \right)^2$$

$$n = \frac{\sigma^2}{\hat{\mu}}$$

- **Intervalestimation (konfidensinterval for middelværdi) hvor variansen er kendt**

$$-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}$$

Ved omskrivning får (1-α) **konfidensintervallet**:

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Maksimal fejl på et estimat hvor variansen <u>ikke</u> er kendt

For store værdier af n gælder:

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

Den maksimale fejl, E, på et estimat med sandsynlighed $\alpha/2 = \frac{(1-\alpha)}{2}$ bliver:

$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, hvor $t_{\alpha/2} = t(n-1)_{\alpha/2}$ findes i tabel 4 $(v = n - 1)$ og s er beregnet varians.

- **Intervalestimation (konfidensinterval for middelværdi) hvor variansen <u>ikke</u> er kendt og en stor stikprøve (n=>30)**

$\overline{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, σ blot erstattet med s. **Konfidensinterval, tabel 3.**

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

z ikke skiftet ud med t, fordi i tabel 4 går n ikke højere end 30 så derfor gøre det ingen forskel.

- **Intervalestimation (konfidensinterval for middelværdi) hvor variansen <u>ikke</u> er kendt og en lille stikprøve (n<30)**

$\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, z erstattet med t. **Konfidensinterval, tabel 4 (v=n-1).**

$$\boxed{\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}}$$

Afsnit 7.3, 7.4 og 7.5

Hypotesetest for et gennemsnit

Formulering af nul-hypotesen og alternativ hypotesen Parameter μ betragtes.

- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}}$$

Man vælger enten at <u>acceptere H_0</u> eller at <u>forkaste H_0</u>.

- Tosidet alternativ:

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{aligned} &H_1 : \mu < \mu_0 \\ &eller \\ &H_1 : \mu > \mu_0 \end{aligned}}$$

- I nulhypotesen anvendes så vidt som muligt lighedstegn.
- I alternativ hypotese placeres det udsagn som man gerne vil vise.
- Eksempelvis: en man stilles for en dommer, anklaget for noget kriminelt. Her bliver nul- og alternativ-hypotesen:

$$\boxed{\begin{aligned} H_0 &: Manden\ er\ ikke\ skyldig \\ H_1 &: Manden\ er\ skyldig \end{aligned}}$$

Trin ved hypotesetest

1. Opstil hypoteser og vælg signifikansniveau α (vælg ”risiko-niveau”).
2. Beregn teststørrelse.
3. Beregn p-værdi vha. teststørrelse. Testets p-værdi måler datas afvigelser fra H_0.
4. Sammenligne p-værdi med signifikansniveau og drag en konklusion. Alternativt kan testet udføres ved at sammenligne teststørrelse med kritisk værdi.

Beregning af teststørrelse, p-værdi og sammenligning, hvis <u>σ er kendt</u>

- Hvis nul- og alternativ-hypotese er formuleret. Og signifikansniveau α er valgt. Så kan teststørrelsen beregnes ved:

$$\boxed{Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}}$$

Der antages en normal fordeling og <u>σ er kendt.</u>

- P-værdien findes for teststørrelsen Z ved opslag i normal fordeling (tabel 3).
- Sammenligning med kritisk værdi z_α (eller $z_{\alpha/2}$ i et tosidet test).

Alternativ	Afvis

hypotese	Nul-hypotese hvis
$\mu < \mu_0$	$Z < -z_\alpha$
$\mu > \mu_0$	$Z > z_\alpha$
$\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ eller $Z > z_{\alpha/2}$

Beregning af teststørrelse, p-værdi og sammenligning, hvis <u>σ er ikke kendt (n>30)</u>

- Hvis nul- og alternativ-hypotese er formuleret. Og signifikansniveau α er valgt. Så kan teststørrelsen beregnes ved:

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Der antages en normal fordeling og <u>σ er ikke kendt.</u>

- P-værdien findes for teststørrelsen Z ved opslag i normal fordeling (tabel 3).
- Sammenligning med kritisk værdi z_α (eller $z_{\alpha/2}$ i et tosidet test).

Alternativ hypotese	Afvis Nul-hypotese hvis
$\mu < \mu_0$	$Z < -z_\alpha$
$\mu > \mu_0$	$Z > z_\alpha$
$\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ eller $Z > z_{\alpha/2}$

Beregning af teststørrelse, p-værdi og sammenligning, hvis <u>σ er ikke kendt (n<30)</u>

- Hvis nul- og alternativ-hypotese er formuleret. Og signifikansniveau α er valgt. Så kan teststørrelsen beregnes ved:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Der antages en normal fordeling og <u>σ ikke er kendt.</u>

- P-værdien findes for teststørrelsen Z ved opslag i t-fordeling (tabel 4), v=n-1.
- Sammenligning med kritisk værdi t_α (eller $t_{\alpha/2}$ i et tosidet test).

Alternativ hypotese	Afvis Nul-hypotese hvis
$\mu < \mu_0$	$t < -t_\alpha$
$\mu > \mu_0$	$t > t_\alpha$
$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ eller $t > t_{\alpha/2}$

- Hvordan kan sandsynligheden for fejl påvirkes:
  - Ved at ændre signifikansniveau α.
  - Ved at øge stikprøvestørrelse n.

Afsnit 7.6-7.9

Hypotesetest for to gennemsnit

- Sammenligner gennemsnit (middelværdier) af 2 stikprøver.
  - Stikprøve 1: $n_1,\ \overline{X}_1\ og\ s_1^2$
  - Stikprøve 2: $n_2,\ \overline{X}_2\ og\ s_2^2$

1. Formulering af hypoteser

- Parameter $\mu_1,\ \mu_2$ betragtes.
- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{array}{l} H_0 : \mu_1 - \mu_2 = \delta \\ H_1 : \mu_1 - \mu_2 \neq \delta \end{array}}$$

Man vælger enten at <u>acceptere H_0</u> eller at <u>forkaste H_0</u>.

- Tosidet alternativ:

$$\boxed{\begin{array}{l} H_0 : \mu_1 - \mu_2 = \delta \\ H_1 : \mu_1 - \mu_2 \neq \delta \end{array}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{array}{l} H_1 : \mu_1 - \mu_2 < \delta \\ eller \\ H_1 : \mu_1 - \mu_2 > \delta \end{array}}$$

- Typisk er man interesseret i at teste med $\delta = 0$.

2. beregning af teststørrelse for <u>kendte varianser $\sigma_1^2\ og\ \sigma_2^2$</u>

- Ved hypotese prøvning af 2 middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser $\sigma_1^2\ og\ \sigma_2^2$ er kendte, fås teststørrelsen:

$$\boxed{Z = \frac{(\overline{X}_1 - \overline{X}_2) - \delta}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}},\text{(tabel 3).}$$

Denne måler forskellen på to grupper og δ stort set altid nul.

3. sammenligning med kritisk værdi for <u>kendte varianser $\sigma_1^2\ og\ \sigma_2^2$</u>

- Ved hypoteseprøvning af to middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser $\sigma_1^2\ og\ \sigma_2^2$ er kendte, fås:

Alternativ hypotese	Afvis Nul-hypotese hvis
$\mu_1 - \mu_2 < \delta$	$Z < -z_\alpha$
$\mu_1 - \mu_2 > \delta$	$Z > z_\alpha$
$\mu_1 - \mu_2 \neq \delta$	$Z < -z_{\alpha/2}$ eller $Z > z_{\alpha/2}$

(tabel 3).

2. beregning af teststørrelse for <u>ikke kendte varianser σ_1^2 og σ_2^2</u>

- Ved hypotese prøvning af 2 middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser σ_1^2 *og* σ_2^2 ikke er kendte, fås teststørrelsen:

$$\boxed{Z = \frac{(\overline{X}_1 - \overline{X}_2) - \delta}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}}, \text{(tabel 3).}$$

Denne måler forskellen på to grupper og δ stort set altid nul.

3. sammenligning med kritisk værdi for <u>ikke kendte varianser σ_1^2 og σ_2^2</u>

- Ved hypoteseprøvning af to middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser σ_1^2 *og* σ_2^2 ikke er kendte, fås:

Alternativ hypotese	Afvis Nul-hypotese hvis
$\mu_1 - \mu_2 < \delta$	$Z < -z_\alpha$
$\mu_1 - \mu_2 > \delta$	$Z > z_\alpha$
$\mu_1 - \mu_2 \neq \delta$	$Z < -z_{\alpha/2}$ eller $Z > z_{\alpha/2}$

(tabel 3).

2. beregning af teststørrelse for <u>ikke kendte varianser σ_1^2 og σ_2^2, men $\sigma_1^2 = \sigma_2^2$</u>

- Ved hypotese prøvning af 2 middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser σ_1^2 *og* σ_2^2 ikke er kendte, men med $\sigma_1^2 = \sigma_2^2$, fås teststørrelsen:

$$\boxed{t = \frac{(\overline{X}_1 - \overline{X}_2) - \delta}{s_p \sqrt{1/n_1 + 1/n_2}}}$$

Denne måler forskellen på to grupper og δ stort set altid nul.

Hvor

$$\boxed{s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Og frihedsgrader:

$$\boxed{v = n_1 + n_2 - 2}$$

(tabel 4).

3. sammenligning med kritisk værdi for <u>ikke kendte varianser σ_1^2 og σ_2^2, men $\sigma_1^2 = \sigma_2^2$</u>

- Ved hypoteseprøvning af to middelværdier (μ_1 og μ_2) for data, der antages normalfordelt og varianser σ_1^2 *og* σ_2^2 ikke er kendte, men $\sigma_1^2 = \sigma_2^2$, fås:

Alternativ hypotese	Afvis Nul-hypotese hvis
$\mu_1 - \mu_2 < \delta$	$t < -t_\alpha$
$\mu_1 - \mu_2 > \delta$	$t > t_\alpha$

$\mu_1 - \mu_2 \neq \delta$	$t < -t_{\alpha/2}$ eller $t > t_{\alpha/2}$

Og frihedsgrader:

$$\boxed{v = n_1 + n_2 - 2}$$

(tabel 4).

Beregning af konfidensinterval for forskel i middelværdi for store stikprøver

- For store stikprøver beregnes et (1-α)% konfidensinterval ved:

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\text{,(tabel 3).}$$

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

Kendes σ_1^2 *og* σ_2^2 anvendes disse i stedet for s_1^2 *og* s_2^2.

Beregning af konfidensinterval for forskel i middelværdi for små stikprøver og ukendt σ_1^2 *og* σ_2^2

- For små stikprøver og ukendt σ_1^2 *og* σ_2^2, men med $\sigma_1^2 = \sigma_2^2$ beregnes et (1-α)% konfidensinterval ved:

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Og frihedsgrader:

$$\boxed{v = n_1 + n_2 - 2}$$

(tabel 4).

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

Afsnit 8.1-8.3, 6.4

Hypotesetest for en varians

χ^2 -fordeling

- Variansen for en stokastisk variabel X estimeres ved:

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

Hvor n er antallet af observationer

X_i er observationer nr. i, hvor i =1…n

$\bar{X}$ estimat af middelværdien for X

Store bogstaver => stokastiskvariabel.

- Lad S^2 være variansen af en stikprøve af størrelsen n fra en normalfordeling med varians σ^2, da er:

$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma^2},\ Ki = \chi}\text{, v = n-1 (tabel 5, s. 588).}$$

$$\boxed{P(\chi^2 \geq \chi^2_\alpha) = \alpha}$$

Konfidensinterval for en varians

Et (1-α)% konfidensinterval for en varians σ^2 fås ved:

$$\boxed{\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}$$

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

v = n-1 (tabel 5, s. 588).

Hypotesetest af en varians

1. Formulering af hypoteser

- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}}$$

Man vælger enten at <u>acceptere H_0</u> eller at <u>forkaste H_0</u>.

- Tosidet alternativ:

$$\boxed{\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{array}{l} H_1 : \sigma^2 < \sigma_0^2 \\ eller \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}}$$

- Hvor σ_0^2 er værdien der testes for.

2. teststørrelse bliver

- Lad S^2 være variansen af en stikprøve af størrelsen n fra en normalfordeling med varians σ^2, da er:

$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma_0^2},\ Ki = \chi}$$, v = n-1 (tabel 5, s. 588).

$$\boxed{P(\chi^2 \geq \chi^2_\alpha) = \alpha}$$

3. sammenligning med kritisk værdi

Alternativ hypotese	Afvis Nul-hypotese hvis
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_\alpha$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$ eller

	$\chi^2 > \chi^2_\alpha$

(tabel 5).

Hypotesetest af 2 varianser

- Sammenligner varianser af 2 stikprøver.
  - Stikprøve 1: $n_1,\ \overline{X}_1\ og\ s_1^2$
  - Stikprøve 2: $n_2,\ \overline{X}_2\ og\ s_2^2$

F-fordeling

- Lad $S_1^2\ og\ S_2^2$ være varianser af stikprøver af størrelserne $n_1\ og\ n_2$ fra en normalfordeling med varians σ^2, da er:

$\boxed{F = \frac{S_1^2}{S_2^2}}$, $v_1 = n_1-1$ og $v_2 = n_2-1$ (tabel 6a og 6b, s. 589-590).

$$\boxed{P(F \geq F_\alpha) = \alpha}$$

Hypotesetest af 2 varianser

1. Formulering af hypoteser

- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{aligned} &H_0 : \sigma_1^2 = \sigma_2^2 \\ &H_1 : \sigma_1^2 \neq \sigma_2^2 \end{aligned}}$$

Man vælger enten at <u>acceptere H_0</u> eller at <u>forkaste H_0</u>.

- Tosidet alternativ:

$$\boxed{\begin{aligned} &H_0 : \sigma_1^2 = \sigma_2^2 \\ &H_1 : \sigma_1^2 \neq \sigma_2^2 \end{aligned}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{aligned} &H_1 : \sigma_1^2 < \sigma_2^2 \\ &eller \\ &H_1 : \sigma_1^2 > \sigma_2^2 \end{aligned}}$$

2. teststørrelse bliver

$\boxed{F = \frac{S_1^2}{S_2^2}}$, $v_1 = n_1-1$ og $v_2 = n_2-1$ (tabel 6a og 6b, s. 589-590).

$$\boxed{P(F \geq F_\alpha) = \alpha}$$

3. sammenligning med kritisk værdi

Alternativ hypotese	Afvis Nul-hypotese hvis
$\sigma_1^2 < \sigma_2^2$	$F > F_\alpha(n_2-1, n_1-1)$
$\sigma_1^2 > \sigma_2^2$	$F > F_\alpha(n_1-1, n_2-1)$

$\sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha/2}(n_M - 1, n_m - 1)$

(i sidste tilfælde gælder $S_M^2 > S_m^2$)

(tabel 6a og 6b).

Afsnit 9.1-9.5

Estimation af andele

fås ved at observere antal gange x en hændelse har indtruffet uf af n forsøg:

$$\boxed{p = \frac{x}{n}}$$

Konfidensinterval for en andel

Såfremt der haves stor stikprøve, fås et (1-α)% konfidensinterval for p:

$$\boxed{\frac{x}{n} - z_{\alpha/2} \cdot \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}} < p < \frac{x}{n} + z_{\alpha/2} \cdot \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}}$$

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

Konfidensinterval for to andele

Såfremt der haves stor stikprøve, fås et (1-α)% konfidensinterval for p_1-p_2:

$$\boxed{p_1 - p_2 = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) \pm z_{\alpha/2} \cdot \sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}}}$$

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

Maksimal fejl på estimat

Den maksimale fejl, E, på et estimat med sandsynlighed $\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$ bliver:

$\boxed{E = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}}$, hvor $z_{\alpha/2}$ findes i tabel 3.

$$\boxed{p = \frac{x}{n}}$$

Bestemmelse af stikprøvestørrelse hvor p kendes

Såfremt man højst vil tillade en maksimal fejl E med (1-α)% konfidens, bestemmes den nødvendige stikprøvestørrelse ved:

$$\boxed{n = p(1-p) \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2}$$

$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$, hvor $z_{\alpha/2}$ findes i tabel 3.

Bestemmelse af stikprøvestørrelse hvor p ikke kendes

Såfremt man højst vil tillade en maksimal fejl E med (1-α)% konfidens, og p ikke kendes, bestemmes den nødvendige stikprøvestørrelse ved:

$$\boxed{n = \frac{1}{4} \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2}, \text{ p=1/2}$$

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}, \text{ hvor } z_{\alpha/2} \text{ findes i tabel 3.}$$

Hypotesetest af 1 andel

1. Formulering af hypoteser

- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{aligned} &H_0 : p = p_0 \\ &H_1 : p \neq p_0 \end{aligned}}$$

Man vælger enten at acceptere H_0 eller at forkaste H_0.

- Tosidet alternativ:

$$\boxed{\begin{aligned} &H_0 : p = p_0 \\ &H_1 : p \neq p_0 \end{aligned}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{aligned} &H_1 : p < p_0 \\ &eller \\ &H_1 : p > p_0 \end{aligned}}$$

2. teststørrelse bliver

- Såfremt stikprøvestørrelsen er tilstrækkelig stor fås teststørrelsen:

$$\boxed{Z = \frac{X - n \cdot p_0}{\sqrt{n \cdot p_0 (1 - p_0)}}}$$

3. sammenligning med kritisk værdi

Alternativ hypotese	Afvis Nul-hypotese hvis
$p < p_0$	$Z < -z_\alpha$
$p > p_0$	$Z > z_\alpha$
$p \neq p_0$	$Z < -z_\alpha$ eller $Z > z_\alpha$

(tabel 3).

Hypotesetest af 2 andel

1. Formulering af hypoteser

- Nul hypotese testes mod alternativ hypotese:

$$\boxed{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{array}}$$

Man vælger enten at <u>acceptere H_0</u> eller at <u>forkaste H_0</u>.

- Tosidet alternativ:

$$\boxed{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{array}}$$

- Ensidet alternativ, der bliver H_1 enten:

$$\boxed{\begin{array}{l} H_1 : p_1 < p_2 \\ eller \\ H_1 : p_1 > p_2 \end{array}}$$

2. teststørrelse bliver

- Såfremt stikprøvestørrelsen er tilstrækkelig stor fås teststørrelsen:

$$\boxed{Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}}$$

Hvor $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

3. sammenligning med kritisk værdi

Alternativ hypotese	Afvis Nul-hypotese hvis
$p < p_0$	$Z < -z_\alpha$
$p > p_0$	$Z > z_\alpha$
$p \neq p_0$	$Z < -z_\alpha$ eller $Z > z_\alpha$

(tabel 3).

Hypotesetest af flere andel

1. Formulering af hypoteser

- I nogle tilfælde kan man være interesseret i at vurdere om to eller flere binomialfordelinger har samme parameter p, dvs. man er interesseret i at teste nul-hypotesen:

$$\boxed{H_0 : p_1 = p_2 = ... = p_k = p}$$

Mod alternativ hypotese at disse andele ikke er ens.

	stikprøve 1	stikprøve 2	...	stikprøve k	Total
Succes	x_1	x_2	...	x_k	x
Fiasko	$n_1 - x_1$	$n_2 - x_2$	...	$n_k - x_k$	$n - x$
Total	n_1	n_2	...	n_k	n

- Under nul-hypotesen får et estimat for p:

$$\boxed{\hat{p} = \frac{x}{n}}$$

- Såfremt nul-hypotesen gælder, vil vi forvente at den j'te gruppe har e_{1j} successer og e_{2j} fiaskoer, hvor

$$\boxed{e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}}$$

$$\boxed{e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}}$$

2. teststørrelse bliver

- Teststørrelsen bliver

$$\boxed{\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

Hvor o_{ij} er observeret antal i celle (i,j) og e_{ij} er forventet antal i celle (i,j). Se tabel AAA længere nede for hvordan de beregnes.

3. sammenligning med kritisk værdi

- Vi har teststørrelsen

$$\boxed{\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

Hvor o_{ij} er observeret antal i celle (i,j) og e_{ij} er forventet antal i celle (i,j). Se tabel AAA længere nede for hvordan de beregnes

- Teststørrelsen sammenlignes med $\boxed{\chi^2_\alpha(k-1)}$
- Såfremt $\boxed{\chi^2 > \chi^2_\alpha(k-1)}$ forkastes nul-hypotesen.

Analyse af antalstabeller

1. Formulering af hypoteser

- Følgende to tabeller er eksempler på antalstabeller:
  - Opgaven kan lyde: Er stemmefordelingen ens for følgende tabel:

AAA	4 uger før	2 uger før	1 uge før	I alt
Kandidat I	79	91	93	263
Kandidat II	84	66	60	210
ved ikke	37	43	47	127
I alt	200	200	200	600

tre eksempler på hvordan o_{ij} og e_{ij} beregnes for denne tabel

(1)

$o_{11} = 79$

$$e_{11} = \frac{200 \cdot 263}{600} = 87{,}67$$

(2)

$o_{12} = 84$

$$e_{12} = \frac{200 \cdot 210}{600} = 70$$

(3)

$o_{23} = 43$

$$e_{23} = \frac{200 \cdot 127}{600} = 42{,}33$$

- Er der uafhængighed mellem inddelingskriterier:

	dårlig	middel	god
dårlig	23	60	29
middel	28	79	60
god	9	49	63

- Opstilling af nul-hypotesen:

$$\boxed{H_0 : p_{i1} = p_{i2} = p_{i3}}$$

2. teststørrelse bliver

- I en antalstabel med r rækker og c søjler, fås teststørrelsen:

$$\boxed{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

Hvor o_{ij} er observeret antal i celle (i,j) og e_{ij} er forventet antal i celle (i,j). Se tabel AAA længere oppe for hvordan de beregnes.

3. sammenligning med kritisk værdi

- Vi har teststørrelsen:

$$\boxed{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

- Teststørrelsen sammenlignes med: $\boxed{\chi^2_\alpha((r-1)(c-1))}$ tabel 5 side 588
- Såfremt $\boxed{\chi^2 > \chi^2_\alpha((r-1)(c-1))}$ det sidste led er v. forkastes nul-hypotesen.

Goodness of fit (test for fordeling)

Ofte vil man gerne teste om data (observationer) følger en specifik fordeling. Dette gøres ved at sammenligne observerede fraktiler med tilsvarende teoretiske fraktiler under forudsætning af en given fordeling. Herefter beregnes teststørrelsen ved

$$\boxed{\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}}$$

Hvor o_{ij} er observeret antal i celle (i,j) og e_{ij} er forventet antal i celle (i,j). Se tabel AAA længere oppe for hvordan de beregnes.

Teststørrelsen skal sammenlignes med kritisk værdi, der findes i $\boxed{\chi^2_\alpha(k-1-m)}$, hvor k er antal inddelinger (celler i tabellen) og m er antal estimerede parametre.

Afsnit 10.1-10.4

Sign test

Kan bruges som alternativ for:

- Hypotesetest for en middelværdi
- Parret t-test

Når ovenstående test ikke kan bruges pga. antagelse om normalfordeling.

1. Formulering af hypoteser

Sign test kan bruges til at teste hypotese om median

$$\boxed{\begin{aligned} &H_0 : \tilde{\mu} = \tilde{\mu}_D \\ &H_1 : \tilde{\mu} \neq \tilde{\mu}_D \end{aligned}}$$

Hvor $\tilde{\mu}_D$ er den værdi vi ønsker at teste.

2. teststørrelse bliver

Beregning af teststørrelse/p-værdi:

- Antal af observationer større end medianen optælles, X_+.
- Testets p-værdi kan nu findes ved at beregne sandsynligheden for (ensidet test)

$$\boxed{P(X \geq X_+)}$$

3. sammenligning med kritisk værdi

Såfremt p-værdi er mindre end signifikansniveau, forkastes H_0.

Rank-sum test

Rank-sum test (også kaldet U-test eller Wilcoxon test eller Mann-Whitney test) kan bruges som alternativ til almindelig t-test for 2 uafhængige stikprøver, i tilfælde af at normalfordelingsantagelse ikke holder.

1. Formulering af hypoteser

Rank-sum test kan altså bruges til at sammenligne medianen for 2 uafhængige stikprøver:

$$\boxed{\begin{matrix} x_1, x_2, ..., x_{n1} \\ y_1, y_2, ..., y_{n1} \end{matrix}}$$

2. teststørrelse bliver

Beregning af teststørrelse: data sorteres og rangeres (eng: ranks) i stigende rækkefælge. For hver af de to stikprøver summeres de tilhørende ranks, her benævnt W_1 og W_2, så der kan beregnes:

$$\boxed{U_1 = W_1 - \frac{n_1(n_1+1)}{2}}$$

$$\boxed{U_2 = W_2 - \frac{n_2(n_2+1)}{2}}$$

Det gælder nu, at såfremt de to stikprøver kommer fra den samme fordeling, så haves:

$$\boxed{\mu_{U_1} = \frac{n_1 \cdot n_2}{2}}$$

$$\boxed{\sigma_{U_1}^2 = \frac{n_1 \cdot n_2(n_1 + n_2 + 1)}{12}}$$

Når n_1 og n_2 er tilpas store (>8) kan vi nu anvende:

$$\boxed{Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} \approx N(0,1^2)} \text{ teststørrelse}$$

3. sammenligning med kritisk værdi

Hvis population 2 er større end population 1:
Så afvises H_0, hvis $Z < -z_\alpha$, da en lille værdi af U_1 giver en lille værdi af W_1.

Hvis population 1 er større end population 2:
Så afvises H_0, hvis $Z > z_\alpha$, da en stor værdi af U_1 giver en stor værdi af W_1.

Test for tilfældighed

- I mange undersøgelser er det vigtigt at afgøre om en stikprøve er fremkommet tilfældigt.
- Hvis vi har en sekvens med n_1 af den ene type og n_2 af en anden type (og hverken $n_1$eller n_2 er mindre end 10), f.eks.:

  K K K P K K P P K P P K P K P…
- Det totale antal skift, u, approksimeres med en normalfordeling med:

$$\boxed{\mu_u = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1} \text{ og}$$

$$\boxed{\sigma_u = \sqrt{\frac{2 \cdot n_1 \cdot n_2(2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)}}}$$

- Vi kan nu beregne p-værdien ved:

$$\boxed{Z = \frac{u - \mu_u}{\sigma_u}} \text{ idet}$$

$$\boxed{Z \approx N(0,1^2)}$$

Afsnit 11.1, 11.2, 11.6

Regressionsanalyse

- Antag at Y er en stokastisk variabel. Vi er interesseret i at modellere Y's afhængighed af en forklarende variabel x.
- Vi undersøger en lineær sammenhæng mellem Y og x, dvs. ved en regressionsmodel på formen:

$$\boxed{Y = \alpha + \beta \cdot x + \varepsilon}$$

Korrelation

- Korrelationskoefficienten r angiver den lineære sammenhæng mellem variablerne x og y.
- Korrelationskoefficienten mellem 2 variable x og y estimeres ved:

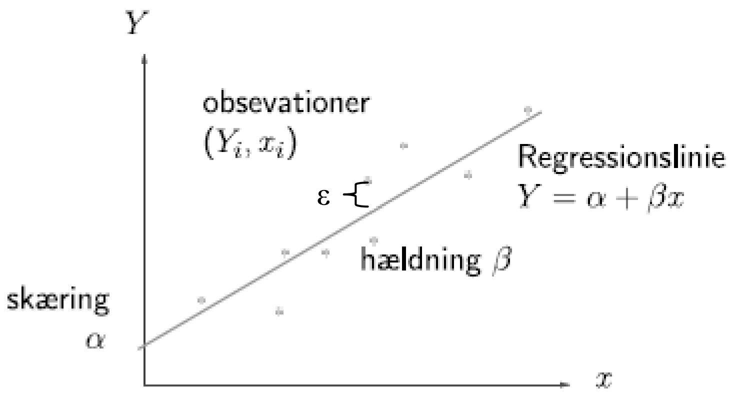
$$\boxed{r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}$$

- Det antages her, at observationerne (x_i,y_i) er sammenhørende værdier. Der gælder $r \in [-1;1]$.

Simpel lineær regressionsmodel

$$\boxed{Y = \alpha + \beta \cdot x + \varepsilon}$$

- $\alpha + \beta \cdot x$ er modellen
- ε er residual (tilfældige fejl, måle fejl eller afvigelse)
- Y afhængige variabel
- x uafhængige variabel
- α skæring med Y-aksen
- β hældning



Mindste kvadraters metode

- Antag at vi har observationerne:

x	1	2	3	4	5	6	7	8	9	10	11	12
y	16	35	45	64	86	96	106	124	134	156	164	182

- Er det en sammenhæng mellem x og y?
- Vi foreslår en model på formen $\boxed{\hat{y} = a + b \cdot x}$
- Hvordan estimeres a og b?

- a og b bestemmes ved:

$$\boxed{b = \frac{S_{xy}}{S_{xx}}}$$

$$\boxed{a = \bar{y} - b \cdot \bar{x}}$$

$$\boxed{S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ eller } \boxed{S_{xx} = \sum x_i^2 - \frac{\left(-\sum x_i\right)^2}{n}} \text{ eller } \boxed{SS_{xx} = s_x^2 \cdot (n-1)}$$

$$\boxed{S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2} \text{ eller } \boxed{S_{yy} = \sum y_i^2 - \frac{\left(-\sum y_i\right)^2}{n}} \text{ eller } \boxed{SS_{yy} = s_y^2 \cdot (n-1)}$$

$$\boxed{S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})} \text{ eller } \boxed{S_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n}}$$

- a og b er nu de værdier, der giver den regressionslinie, der minimerer den kvadratiske afstand mellem punkter og linie.
- a er et estimat for α og b er et estimat for β.

Interferens i regressionsmodel

- vi antager at de observerede data (Y_i,x_i) kan beskrives ved modellen:

$$\boxed{Y_i = \alpha + \beta \cdot x_i + \varepsilon_i}$$

Hvor det antages at ε_i er uafhængige normalfordelte stokastiske variable med middelværdi 0 og konstant varians σ^2.

- Estimatet af σ^2 bliver (varians af residualerne):

$$\boxed{s_e^2 = \frac{S_{yy} - (S_{xy})^2 / S_{xx}}{n-2}}$$

1. Formulering af hypotese om skæring med y-aksen

- Antag at vi vil teste en hypotese om skæring med y-aksen:

$$\boxed{\begin{aligned} H_0 &: a = \alpha \\ H_1 &: a \neq \alpha \end{aligned}}$$

2. teststørrelse bliver

$$\boxed{t = \frac{(a-\alpha)}{s_e} \sqrt{\frac{n \cdot S_{xx}}{S_{xx} + n \cdot (\bar{x})^2}}}$$

3. sammenligning med kritisk værdi

- Kritisk værdi findes i t-fordeling:

$$\boxed{t_{\alpha/2}(n-2)} \text{ tabel 4}$$

1. Formulering af hypotese om hældningen β

- Antag at vi vil teste en hypotese om hældningen β

$$\boxed{\begin{aligned} H_0 &: b = \beta \\ H_1 &: b \neq \beta \end{aligned}}$$

2. teststørrelse bliver

$$\boxed{t = \frac{(b-\beta)}{s_e}\sqrt{S_{xx}}}$$

3. sammenligning med kritisk værdi

- Kritisk værdi findes i t-fordeling:

$\boxed{t_{\alpha/2}(n-2)}$ tabel 4

Konfidensintervaller for α og β

- Konfidensinterval for α:

$\boxed{a \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}}$ $t_{\alpha/2}$ i tabel 4, v = n-2

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

- Konfidensinterval for β:

$\boxed{b \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}}$ $t_{\alpha/2}$ i tabel 4, v = n-2

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

*Konfidensintervaller for α+ β*x₀*

- Konfidensinterval for α+ β*x_0 svarer til et konfidensinterval for modellen i punktet x_0:

$\boxed{(a + b \cdot x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$ $t_{\alpha/2}$ i tabel 4, v = n-2

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

*Prædiktionsinterval for α+ β*x₀*

- Prædiktionsinterval for α+ β*x_0 svare til et prædiktionsinterval for modellen i punktet x_0:

$\boxed{(a + b \cdot x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$ $t_{\alpha/2}$ i tabel 4, v = n-2

$$\boxed{\alpha/2 = \frac{(1-\alpha)}{2}}$$

- Et prædiktionsinterval bliver altså større end et konfidensinterval for fastholdt α.

Korrelation og regression

- Korrelation og regression:

$\boxed{r = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} b}$ $\boxed{r^2 = \frac{S_{xx}}{S_{yy}} b^2}$, hvor

$$\boxed{S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2} \text{ eller } \boxed{S_{xx}=\sum x_i^2-\frac{\left(-\sum x_i\right)^2}{n}} \text{ eller } \boxed{SS_{xx}=s_x^2\cdot(n-1)}$$

$$\boxed{S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2} \text{ eller } \boxed{S_{yy}=\sum y_i^2-\frac{\left(-\sum y_i\right)^2}{n}} \text{ eller } \boxed{SS_{yy}=s_y^2\cdot(n-1)}$$

$$\boxed{S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})} \text{ eller } \boxed{S_{xy}=\sum x_i y_i-\frac{\sum x_i\cdot\sum y_i}{n}}$$

- Korrelationen r udtrykker graden af lineær sammenhæng.
- Korrelationen kvadreret r^2 udtrykker ”forklaringsgraden”:

S_{yy}= variation forklaret af linien + uforklaret variation:

$$\boxed{S_{yy}=\frac{S_{xy}^2}{S_{xx}}+\left(S_{yy}-\frac{S_{xy}^2}{S_{xx}}\right)}$$

Afsnit 12.1-12.3

Variansanalyse (forskel i middel)

Gruppe A	Gruppe B	Gruppe C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

- Er der forskel (i middel) på grupperne A, B og C?
- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne i hver gruppe kan antages at være normalfordelte.

En-sidet variansanalyse

- Vi betragter modellen:

$\boxed{X_{ij}=\mu+\alpha_i+\varepsilon_{ij}}$ hvor det antages $e_{ij}\approx N(0,\sigma^2)$

- μ er gennemsnit for alle målinger.
- α_i angiver niveau af ”gruppe” i.

1. Formulering af hypotese

- vi vil nu sammenligne (flere end to) middelværdier $\mu+\alpha_i$ i modellen:

$\boxed{X_{ij}=\mu+\alpha_i+\varepsilon_{ij}}$ hvor det antages $e_{ij}\approx N(0,\sigma^2)$

Dvs. hypotesen kan opstilles:

$$\boxed{\begin{array}{l}H_0:\alpha_i=\alpha_j\\ H_1:\alpha_i\neq\alpha_j\end{array}}$$

2. teststørrelse bliver

- Variansanalysetabel

Variations-kilde	Friheds-grader	Kvadratafvig. sum	Test-størrelse F
Behandlig	$k-1$	$SS(Tr)$	$\frac{SS(Tr)/(k-1)}{SSE/(N-k)}$
Residual	N-k	SSE	
Total	N-1	SST	

- Den totale varians:

$$\boxed{SST = SS(Tr) + SSE}$$

- Test størrelsen F:

$$\boxed{F = \frac{SS(Tr)/(k-1)}{SSE/(N-k)}}$$

- Måleusikkerheden (residual) varians:

$$\boxed{\sigma^2_{error} = \frac{SSE}{N-k}}$$

- Behandlingsvarians:

$$\boxed{\sigma^2_{treatment} = \frac{SS(Tr)}{k-1}}$$

Hvor k er niveauer antal slags prøver fortaget over en faktor, og N er antal observationer.

- Formler for kvadrat afvigelses sum:

$$\boxed{SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - C}$$

$$\boxed{SS(Tr) = \sum_{i=1}^{k}\frac{T_i^2}{n_i} - C}, \text{ hvor}$$

$$\boxed{C = \frac{T_{\cdot}^2}{N}}, \boxed{T_i = \sum_{j=1}^{n_i} y_{ij}}, \boxed{T. = \sum_{i=1}^{k} T_i}$$

3. sammenligning med kritisk værdi

- Teststørrelsen sammenlignes med en fraktil i F fordelingen:

$$\boxed{F \sim F_{\alpha}(k-1, N-k)}$$

Tosidet variansanalyse

	A_1	A_2	A_3
B_1	x	x	x
B_2	x	x	x
B_3	x	x	x
B_4	x	x	x

- Vi antager nu, at vi har modellen:

$\boxed{X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}}$ hvor det antages $e_{ij} \sim N(0, \sigma^2)$

Dvs. vi har to inddelingskriterier, både α og β, hvor β også kan opfattes som en blok, hvorfor designet også kaldes et randomiseret blokforsøg.

Variations-kilde	Friheds-grader	Kvadrat-afvig.sum	Test-størrelse F
Behandlig	$a-1$	$SS(Tr)$	$\frac{SS(Tr)/(a-1)}{SSE/((a-1)(b-1))}$
Blokke	$b-1$	$SS(Bl)$	$\frac{SS(Bl)/(b-1)}{SSE/((a-1)(b-1))}$
Residual	$(a-1)(b-1)$	SSE	
Total	$N-1$	SST	

- Den totale varians:

$$\boxed{SST = SS(Tr) + SS(Bl) + SSE}$$

- Test størrelsen F:

$$\boxed{F = \frac{SS(Tr)/(a-1)}{SSE/((a-1)(b-1))}} \text{ eller}$$

$$\boxed{F = \frac{SS(Bl)/(b-1)}{SSE/((a-1)(b-1))}}$$

- Måleusikkerheden (residual) varians:

$$\boxed{\sigma^2_{error} = \frac{SSE}{((a-1)(b-1))}}$$

- Behandlingens varians:

$$\boxed{\sigma^2_{treatment} = \frac{SS(Tr)}{a-1}}$$

- Blokkenes varians:

$$\boxed{\sigma^2_{blocks} = \frac{SS(Bl)}{b-1}}$$

- Formler for kvadrat afvigelses sum:

$$\boxed{SST = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - C}$$

$$\boxed{SS(Tr) = \frac{\sum_{i=1}^{a} T_{i.}^2}{b} - C}$$

$$\boxed{SS(Bl) = \frac{\sum_{j=1}^{b} T_{.j}^2}{a} - C}, \text{ hvor } \boxed{C = \frac{T_{..}^2}{ab}}$$

- Kritisk værdi for blokke:

$$\boxed{F_\alpha(b-1,(a-1)(b-1))}$$

- Kritisk værdi for behandling:

$$\boxed{F_\alpha(a-1,(a-1)(b-1))}$$

Definition på parat t-test

Hvis man måler blodtryk på 10 personer og måler højden på de samme 10 personer er det et parat t-test man ser på for at sammenligne.