Indholdsfortegnelse

Generelt:

Stokastisk variabel:

En stokastisk variabel X er en funktion defineret på Ω, der antager værdier på den reelle akse.

Tæthedsfunktion/sandsynlighedsfunktion for stokastisk variabel:

(frekvensfunktion / hyppighed)
Et godt plot af *f(x)* er histogram for kontinuerte variable og barchart for diskrete variable

Sandsynlighedsfunktion (Diskret variabel):

$f(x) = P(X = x)$

$\sum f(x) = 1$

$f(x) > 0 \quad \text{for} \, x \in S$

$f(x) = 0 \quad \text{for} \, x \notin S$

S er udfaldsrummet for X

Tæthedsfunktion (Kontinuert variabel):

$f(x) \geq 0 \quad \text{for alle } x$

$\int_{-\infty}^{\infty} f(x) \, \mathrm{d}x = 1$

$P(a < X < b) = \int_{a}^{b} f(x) \, \mathrm{d}x$

Husk at: $P(a < X < b) = P(a \leq X \leq b)$

En kontinuert stokastisk variabel har punktsandsynlighed = 0 for alle udfald, dvs. $P(X = x) = 0$

Fordelingsfunktion/sumfunktion for stokastisk variabel:

Diskret variabel:

$F(x) = P(X \leq x) = \sum_{t=-\infty}^{x} f(t)$

Et godt plot er den kumulative fordeling.

Kontinuert variabel:

$F(x) = P(X \leq x) = \int_{t=-\infty}^{x} f(t) \, \mathrm{d}t$

Generelt:

Middelværdi:

μ bliver estimeret af gennemsnittet $\bar{x}$.

Diskret variabel:

$$\mu = \sum_{alle\ x} x \cdot f(x)$$

Kontinuert variabel:

$$\mu = \int_S x \cdot f(x)\ \mathrm{d}x$$

S er udfaldsrummet for X, normalt (-∞ ; ∞).

Gennemsnit:

$\bar{x}$ er et estimat af middelværdien.

Diskret variabel:

$$x\sim = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Grupperede data (i intervaller):

$$x\sim = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot f_i$$

Fraktiler:

Definition:
Den p'ende fraktil er en værdi, x_p i datasættet, hvor mindst p % af alle observationer ≤ x_p.

For at udregne den p'ende fraktil for n observationer, udregnes $k = \frac{p}{100} \cdot n$, hvor p er i %.
Dette tal fortæller noget om, hvilken observation vi skal "tælle op til".
Vigtigt: Husk at ordne datasættet, så observationerne står i numerisk rækkefølge.

Hvis k ikke er et heltal, rundes op til næste observation, som vil have værdien x_p.

Hvis k er et heltal, tages gennemsnittet af k og k + 1.

Specielle fraktiler:
Median: 50 %
Nedre kvartil hhv. øvre kvartil : 25 % og 75 %.
Decilerne er 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% og 90%.

Eksempel 1:
Finde 25 % -fraktilen (den nedre kvartil) for 13 observationer:
Givet: p = 25 n = 13.
Derfor: $k = 0{,}25 \cdot 13 = 3{,}125$.
Da k ikke er et heltal, runder vi op til 4.
Derfor er x_p = den 4. observation.

Eksempel 2
Finde 50 %-fraktilen (medianen) for 50 observationer:
Givet: p = 50 n = 50
Derfor: $k = 0{,}50 \cdot 50 = 25$.
Da k er et heltal, skal vi finde gennemsnittet mellem observation 25 og 26.
Derfor er $x_p = \frac{x_{25} + x_{26}}{2}$.

Median:
Medianen er 50%-fraktilen, dvs. den midterste observation i det ordnede datasæt.
Hvis man har ekstremt afvigende værdier, er medianen at foretrække frem for middelværdien.

For en symmetrisk fordeling er medianen = gennemsnittet $\underline{x}$

Skæve fordelinger:
Hvis fordelingen har en "hale", så den ikke længere er symmetrisk, er den skæv.

Fordelingen er højreskæv, når halen ligger til højre (positively skewed).
Medianen for højreskæve fordelinger > $\underline{x}$.

For delingen er venstreskæv, når halen ligger til venstre (negatively skewed).
Medianen for venstreskæve fordelinger < $\underline{x}$.

Generelt:

Varians:

Diskrete variable:

$$\sigma^2 = \sum_{alle\ x} (x-\mu)^2 \cdot f(x)$$

Grupperede data (i intervaller):

$$\sigma^2 = \frac{n \cdot \sum_{i=1}^{n} x_i^2 \cdot f_i - \left(\sum_{i=1}^{n} x_i \cdot f_i \right)^2}{n \cdot (n-1)}$$

Kontinuerte variable:

$$\sigma^2 = \int_S (x-\mu)^2 \cdot f(x)\ \mathrm{d}x$$

S er udfaldsrummet for X, normalt (-∞ ; ∞).

Spredning/standardafvigelse:

$$\sigma = \sqrt{\sigma^2}$$

Variationskoefficient:

Bruges hvis man skal sammenligne variationen mellem forskellige datasæt.

$$V = \frac{\sigma}{x\sim} \cdot 100$$

Diskrete stokastiske variable:

Diskrete stokastiske variable er tællevariable, der tager heltalsværdier.

Binomialfordeling:

$X \in b(x; n, p)$

n = antal forsøg
p = sandsynlighedsparameter

Bruges når:

- Verden kan deles op i to, dvs. der kan tælles antal succes og antal fiaskoer.
- Et udfald er uafhængigt af tidligere udfald, dvs. p er konstant/ sandsynligheden for succes er ens for alle forsøg.

Eksempler på brug:

- Stikprøver med tilbagelægning
- Vælgertilslutning til Venstre blandt 500 vælgere
- Antal korrekte svar i en multiple choice test med 25 spørgsmål
- Hvor mange seksere der slås på 6 slag
- Hvor mange gange man får krone ud af 10 kast med en mønt

Sandsynlighedsfunktion:

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

NB: husk for at udregne $\binom{n}{x}$, bruges kommandoen nCr(n, x) / nCr(øverst, nederst) på TI89

Fordelingsfunktion:

$F(x) = P(X \leq x)$ *skrives* $B(x; n, p)$

Disse værdier kan findes i tabellen s. 565.

Middelværdi:

$\mu = n \cdot p$

Varians:

$\sigma^2 = n \cdot p \cdot (1-p)$

Diskrete stokastiske variable:

Diskrete stokastiske variable er tællevariable, der tager heltalsværdier.

Binomialfordeling:

Approksimation af binomialfordelingen vha. Poissonfordelingen

Binomialfordelingen kan approksimeres med poissonfordelingen, hvis n er tilstrækkeligt stort og p tilstrækkeligt lille. Herved fås $\lambda = n \cdot p$.
En tommelfingerregel er, at denne approksimation kan anvendes når $n \geq 20$ og $p \leq 0{,}05$.
Hvis $n \geq 100$, er det dog tilstrækkeligt, at $n \cdot p \leq 10$.
Hvis p er tæt på 0,50, anvendes normalfordelingen til at approksimere (se nedenfor).

Approksimation af binomialfordelingen vha. normalfordelingen

Binomialfordelingen kan approksimeres med normalfordelingen, hvis n er stor og p er tæt på 0,50. Herved udregnes Z:

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

hvor Z følger en standardnormalfordeling med $\mu = 0$ og $\sigma^2 = 1$. Denne transformation er identisk med transformationen der normalt anvendes til at standardisere normalfordelinger, da $\mu = n \cdot p$ og $\sigma^2 = n \cdot p \cdot (1-p)$ for binomialfordelingen.

Det er vigtigt at ændre grænserne for Z, da binomialfordelingen er en diskret fordeling, hvor normalfordelingen er en kontinuert fordeling.
(Forskellen ligger i at den diskrete fordeling udelukkende består af punktsandsynligheder, hvorimod punktsandsynligheden i den kontinuerte = 0.)

Hvis man fx skal finde $P(X \leq 8)$, vil den X-værdi, man skal indsætte i formlen, være 8,5, da normalfordelingens interval (7,5 ; 8,5) "dækker" binomialfordelingens X = 8.
Hvis man fx skal finde $P(X < 8)$, vil X-værdien være 7,5, da, hvad der svarer til binomialfordelingens X = 8, ikke skal være inkluderet.
Hvis man fx skal finde $P(3 \leq X < 8)$, vil de to X-værdier skulle være hhv. 2,5 og 7,5.

Diskrete stokastiske variable:

Diskrete stokastiske variable er tællevariable, der tager heltalsværdier.

Hypergeometrisk fordeling:

$X \in h(x; n, a, N)$

n = antal forsøg
a = antal defekte i populationen
N = populationens størrelse

Bruges når:

- Verden kan deles op i to, dvs. der kan tælles antal succes og antal fiaskoer.
- Et udfald er afhængigt af tidligere udfald, dvs. p ikke er konstant/sandsynligheden for succes ændrer sig for hvert forsøg.

Eksempler på brug:

- Stikprøver uden tilbagelægning
- Sandsynligheden for at få 5 hjerter hvis man udtager 5 kort fra et almindeligt spil kort.
- Sandsynligheden for at få 2 blå bolde, hvis man udtager 5 bolde fra en pose med 15 blå bolde og 5 røde.
- Sandsynligheden for at finde 10 defekte fjernsyn ud af en prøve på 20, når prøven er taget fra et parti på 80 indeholdende 34 defekte.

Sandsynlighedsfunktion:

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

NB: husk nCr(øverst, nederst) på TI89

Fordelingsfunktion:

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

Middelværdi:

$\mu = n \cdot \frac{a}{N}$

Varians:

$\sigma^2 = n \ \frac{a}{N} \ (1 - \frac{a}{N}) \ (\frac{N-n}{N-1})$

Diskrete stokastiske variable:

Diskrete stokastiske variable er tællevariable, der tager heltalsværdier.

Hypergeometrisk fordeling:

Approksimation af den hypergeometriske fordeling vha. binomialfordelingen:

Den hypergeometriske fordeling kan approksimeres med binomialfordelingen, når stikprøven (n) er lille i forhold til populationen (N).
N skal være mindst 10n, før denne approksimation kan anvendes.

Den approksimerende binomialfordelings parametre er så n = n, og $p = \frac{a}{N}$.

Diskrete stokastiske variable:

Diskrete stokastiske variable er tællevariable, der tager heltalsværdier.

Poissonfordeling:

$X \in P(\lambda)$

λ = intensiteten/gennemsnittet pr. "enhed"

Bruges når:

- Verden ikke kan deles op i enten succes eller fiasko. Fx handler det ikke om en bombe slår ned eller ej, men hvor mange bomber der slår ned.
- Fordelingen karakteriseres ved λ, som er intensiteten af vores variabel.
  λ angives på formen antal pr. noget (fx pr. dag/pr. areal/pr. person etc.).
- Poissonfordelingen bruges ofte, når der ikke er nogen naturlig øvre grænse for λ.
- Poissonfordelingen anvendes, når det ikke er muligt at tælle, hvor mange fiaskoer, der ikke er. Dvs. det totale antal fiaskoer ikke kendes.

Eksempler på brug:

- Hvor mange personer der indlægges pr. dag grundet luftforurening.
- Hvor mange kunder der går ind i et supermarked på en time.
- Hvor mange computere der går ned på en dag.
- Hvor mange genstande alkohol Janne drikker på en ganske almindelig uge.
- Hvor mange fejl der er pr. meter sejlgarn

Sandsynlighedsfunktion:

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

når $\lambda > 0$

Fordelingsfunktion:

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

Se tabel s. 570

Middelværdi:

$\mu = \lambda$

Varians:

$\sigma^2 = \lambda$

Kontinuerte stokastiske variable:

Kontinuerte stokastiske variable er måledata, der derfor er reelle tal.
Sandsynlighederne regnes i intervaller.

Normalfordelingen:

$X \in N(\mu, \sigma^2)$

μ = middelværdien
σ^2 = variansen

Standardnormalfordelingen (også kaldet Z-fordelingen) er;

$Z \in N(0, 1^2)$

P(Z≤ X) kan findes i Tabel 3.

Bruges når:
Man har kontinuerte observationer, som har en klokkeformet sandsynlighedsfunktion,

Eksempler på brug:

- Fordelingen af værnepligtiges højde
- Vægten af melposer fyldt af en robot
- Tiden det tager for en ambulance at nå frem til en givet distance
- Hvor lang tid det tager at poppe en pose mikrobølgeovnspopcorn

Tæthedsfunktion:

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

Fordelingsfunktion:

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

For standardnormalfordelingen se forrest i bogen eller Tabel 3

HUSK at for standardnormalfordelingen gælder at

$P(X \leq 1, 645) = 5\%$ $P(X \geq -1, 645) = 5\%$

$P(X \leq 1, 960) = 2, 5\%$ $P(X \geq -1, 960) = 2, 5\%$

Middelværdi:

$\mu = \mu$

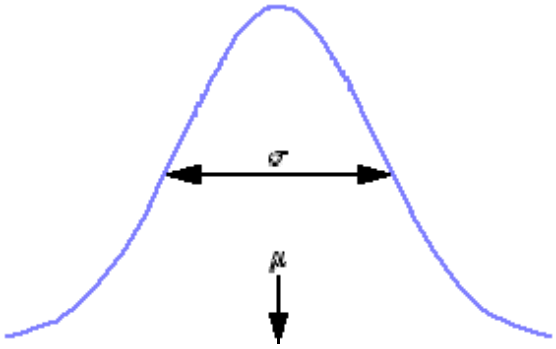
Varians:

$\sigma^2 = \sigma^2$

Kontinuerte stokastiske variable:

Kontinuerte stokastiske variable er måledata, der derfor er reelle tal.
Sandsynlighederne regnes i intervaller.

Normalfordelingen:



Standardisering:

$Y \in N(\mu, \sigma^2)$

Kan standardiseres ved at beregne:

$$Z = \frac{Y - \mu}{\sigma}$$

Der gælder da, at $Z \in N(0, 1^2)$

Og at $P(X \leq Y) = P(X \leq Z) = \Phi(Z)$

Husk i øvrigt at $P(X \geq Z) = 1 - \Phi(Z)$

$$P(Z_1 \leq X \leq Z_2) = \Phi(Z_2) - \Phi(Z_1)$$

Intervallet [μ-σ,μ+σ] rummer mere end halvdelen af udfaldene.

Kontinuerte stokastiske variable:

Kontinuerte stokastiske variable er måledata, der derfor er reelle tal.
Sandsynlighederne regnes i intervaller.

Eksponentialfordelingen:

Bruges til:
At beskrive levetider og ventetider.
At beskrive (vente)tiden mellem hændelser i poissonfordelingen.

Eksempler på brug:

- Tiden der går mellem kunders ankomst til et supermarked.
- Hvor lang tid der går mellem indlæggelser pga. luftforurening.
- Hvor lang tid der går mellem computeren går ned.
- Hvor lang tid der går mellem Jannes indtagelse af alkohol.
- Hvor mange cm der er mellem 2 fejl på et stykke sejlgarn

Tæthedsfunktion:

$$f(x) = \begin{cases} \frac{1}{\beta} \cdot e^{-\frac{x}{\beta}} & \text{for } x > 0 \text{ og } \beta > 0 \\ 0 & \text{ellers} \end{cases}$$

$$ß = \frac{1}{\lambda}$$

Fordelingsfunktion:
Sandsynligheden for at ventetiden er mindre end x:

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

Sandsynligheden for at ventetiden er over x:

$$F(X \ge x) = \int_x^\infty \frac{1}{\beta} e^{\frac{-x}{\beta}} dx = 0 - (-e^{\frac{-x}{\beta}}) = e^{\frac{-x}{\beta}}$$

Middelværdi:

$$\mu = \beta$$

Varians:

$$\sigma^2 = \beta^2$$

Sammenhæng med Poissonfordelingen:
Poisson: Diskrete hændelser pr. enhed
Eksponential: Den kontinuerte afstand mellem diskrete hændelser

Kontinuerte stokastiske variable:

Kontinuerte stokastiske variable er måledata, der derfor er reelle tal.
Sandsynlighederne regnes i intervaller.

Uniformfordelingen:

$X \in U(\alpha, \beta)$

Bruges når:
Der er en kontinuert stokastisk variabel hvor alle udfald i et interval er lige sandsynlige

Eksempel på brug:

- Trykket på fælgen under bremseklodsen, når man bremser på sin cykel.

Tæthedsfunktion:

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

For $\alpha < x < \beta$. For alle andre x er f (x) = 0.

Middelværdi:

$$\mu = \frac{\alpha + \beta}{2}$$

Varians:

$$\sigma^2 = \frac{1}{12} \cdot (\beta - \alpha)^2$$

Kontinuerte stokastiske variable:

Kontinuerte stokastiske variable er måledata, der derfor er reelle tal.
Sandsynlighederne regnes i intervaller.

Log-Normalfordelingen:

$X \in LN(\alpha, \beta)$

$\alpha = \mu$, $\beta = \sigma$.

Bruges når:
Når den stokastiske variabel er normalfordelt, hvis man tager ln af P(X).

Eksempler på brug:

- Intervallet mellem supernova-eksplosioner.
- Forstærkelsen af transistorsignaler

Tæthedsfunktion:

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

Når $x > 0$ og $\beta > 0$
Ellers er $f(x) = 0$

Middelværdi:

$$\mu = e^{\alpha + \frac{\beta^2}{2}}$$

Varians:

$$\sigma^2 = e^{2\alpha + \beta^2} \cdot \left(e^{\beta^2} - 1\right)$$

Transformering til standardnormalfordeling:

$Y \in LN(\alpha, \beta)$

Kan standardiseres ved at beregne:

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

Regneregler for stokastiske variable s. 183

Generelt
E(X) = den forventede værdi af X (på engelsk: expected value) = middelværdien = μ.
Var(X) = variansen af X = σ^2.

X er en stokastisk variabel, a og b er konstanter.

E(X)
$E(aX + b) = aE(X) + b$

$E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n)$

Var(X)

$Var(aX + b) = a^2Var(X)$

$Var(a_1X_1 + a_2X_2 + ... + a_nX_n) = (a_1)^2\,Var(X_1) + (a_2)^2\,Var(X_2) + ... + (a_n)^2\,Var(X_n)$

Vigtigt!
Når man skal bestemme variansen, er det vigtigt at finde ud af, om man får information om variansen for hvert enkelt X (uafhængige variable) eller blot om gennemsnitsvariansen for alle X (afhængige variable).
Hvis man kun kender gennemsnitsvariansen, bruges den første formel (der sættes i 2.).
Hvis man kender variansen for hvert enkelt X, svarer det til at bruge den anden formel (der sættes *ikke* i 2.). Da $a_i = 1$ for alle X_i, vil alle $(a_i)^2 = (-1)^2/1^2 = 1$. Derfor skal man blot lægge varianserne sammen.

Eksempel:

Der er blevet taget tid på, hvor lang tid det tager for en computer at læse en chip. E(X) = 2 sekunder og Var(X) = 0,05 sekunder. Desuden tager computeren præcis 60 sekunder om at starte op.
Find E(200X + 60) og Var(200X + 60)

- $E(200X + 60) = 200 \cdot E(X) + 60 = 200 \cdot 2 + 60 = 460$ sekunder
- For at finde Var(X), skal det først bestemmes om X er en afhængig eller en uafhængig variabel:
- Hvis man tester den samme chip 200 gange, er X en afhængig variabel. Derfor er $Var(200X + 60) = 200^2 \cdot Var(X) = 200^2 \cdot 0{,}05 = 2000$ sekunder
- Hvis man tester 200 forskellige chips én gang hver, er X en uafhængig variabel.
  I dette tilfælde er $Var(200X + 60) = Var(X) + Var(X) + ... + Var(X) = 200 \cdot Var(X) = 10$ sekunder.

Transformationer s. 193

Hvis man har meget skæve fordelinger, kan man formindske ekstreme værdiers indflydelse på datasættet ved at transformere datasættet. I 99% af alle tilfælde anvendes ln(x). Derved vil en venstreskæv fordeling blive klokkeformet og derfor kunne approksimeres med en normalfordeling.

For venstreskæve fordelinger: (halen er til venstre)
Her skal store værdier gøres mindre.
Her anvendes oftest:

- ln(x)
- $\frac{1}{x}$
- $\sqrt{x}$
- $\sqrt[4]{x}$

For højreskæve fordelinger (halen er til højre):
Her skal store værdier gøres større:
Der anvendes ofte:

- x^2
- x^3

Stikprøvefordelinger med kendt varians: s. 209

$\overline{X}$ er middelværdien for stikprøven.

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Uendelig population: (stikprøve lille i forhold til population)

Bruges når:

- Stikprøven er lille i forhold til populationen.
- Der kendes: μ og σ^2 (populationens middelværdi og varians)

Fordeling:

Uanset selve populationens fordeling, vil stikprøvens $\overline{X}$ altid følge en normalfordeling, når stikprøvestørrelsen gøres stor nok (se figurer s. 213).

$\overline{X}$ følger en fordeling med middelværdi μ og varians $\frac{\sigma^2}{n}$

$\overline{X}$ er kan tilnærmes med normalfordelingen

$$\overline{X} \in N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardisering:

Ifølge den centrale grænseværdisætning s. 212 vil

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

Z følger en standardnormalfordeling for $n \to \infty$

Endelig population: (stikprøve stor i forhold til population)

Bruges sjældent

Bruges når:

- Stikprøven er stor i forhold til populationen
- Der kendes: μ og σ^2 (populationens middelværdi og varians)

Fordeling:

$\overline{X}$ følger da en fordeling med middelværdi μ og varians $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

Stikprøvefordelinger med estimeret varians: s. 216

$\overline{X}$ er middelværdien for stikprøven.

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

S^2 er den estimerede varians af stikprøven.

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

t-fordelingen:

Minder meget om normalfordelingen, er også symmetrisk omkring 0. Men t-fordelingen er bredere.

Bruges til:
Det er en stikprøvefordeling for middelværdien

Fordelingen:

$$t = \frac{\overline{X} - \mu}{\left(\frac{S}{\sqrt{n}}\right)}$$

t er da en stokastisk variabel og følger en t-fordeling med parameter $\upsilon = n-1$
υ kaldes frihedsgraden

Opslag:
Tabel 4, s. 587
Ved $t_\alpha(\upsilon)$ forstås den værdi således at $P(t \geq t_\alpha) = \alpha$

HUSK at i t-fordelingen læser man større end eller lig med i tabellen, modsat normalfordelingen.

Kan approksimeres med:
Standardnormalfordelingen, hvis $\upsilon > 29$
for da er det stort set det samme.

Der gælder at:
$t(\infty)$ = standardnormalfordeling
en t-fordeling er bredere end en standardnormalfordeling

Stikprøvefordeling for variansen: s. 218

$\overline{X}$ er middelværdien for stikprøven.

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

S^2 er den estimerede varians af stikprøven.

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

χ^2 -fordelingen:

χ^2-fordelingen er ikke symmetrisk, er venstreskæv.

Bruges til:
Det er en stikprøvefordeling for variansen.

Fordelingen:

$$\chi^2 = \frac{(n-1) \cdot S^2}{\sigma^2}$$

χ^2 er da en stokastisk variabel og følger en χ^2-fordeling med parameter $\upsilon = n-1$
υ kaldes frihedsgraden

Opslag:
Tabel 5, s. 588
Ved $\chi^2{}_{\alpha}(\upsilon)$ forstås den værdi således at $P\left(\chi^2 \geq \chi_{\alpha}{}^2\right) = \alpha$

HUSK at i χ^2-fordelingen læser man større end eller lig med i tabellen, modsat normalfordelingen.

Stikprøvefordeling for sammenlignede varianser:

s. 220

$\overline{X}$ er middelværdien for stikprøven.

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

S^2 er den estimerede varians af stikprøven.

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

F -fordelingen:

F-fordelingen er ikke symmetrisk

Bruges til:

Det er en stikprøvefordeling for sammenligning af varianser.
Når:
Der haves 2 stikprøver fra to normalfordelinger med samme varians.
S_1^2 er den estimerede varians af den ene stikprøve med størrelse n_1.
S_2^2 er den estimerede varians af den ene stikprøve med størrelse n_2.

Fordelingen:

$$F = \frac{{S_1}^2}{{S_2}^2}$$

F er da en stokastisk variabel og følger en F-fordeling med parametre $\upsilon_1 = n_1\text{-}1$ og $\upsilon_2 = n_2\text{-}1$
υ kaldes frihedsgraden.

Opslag:

Tabel 6, s. 589 og frem
Ved $F_\alpha(\upsilon_1,\upsilon_2)$ forstås den værdi således at $P\left(F \geq F_\alpha\right) = \alpha$

HUSK at i F-fordelingen læser man større end eller lig med i tabellen, modsat normalfordelingen.
OBS! Der er en tabel for hver sandsynlighed se toppen. $F_{0,05}$ betyder at 5% er over værdien.

Inferens for gennemsnit s. 226

Stikprøven skal være repræsentativ for populationen. Følgende er mål for, hvor god stikprøven er:

Central estimator:
Stikprøven skal være centreret omkring den sande middelværdi (unbiased).

Efficient estimator:
Variansen på stikprøven skal være så lille som muligt.

Maksimal fejl på et estimat

Forudsætning: Observationerne skal kunne antages at være normalfordelte.

σ kendt:
Den maksimale fejl med sandsynlighed 1-α er:

$$E_{1-\alpha} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}$ findes i standardnormalfordelingen, tabel 3, eller i nederste linje i tabel 4.

σ ukendt, n > 30:

$$E_{1-\alpha} = z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

σ ukendt, n < 30:

$$E_{1-\alpha} = t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

$t_{\alpha/2}$ findes i tabel 4, med frihedsgrad v = n-1

Stikprøvestørrelse
Beregnes ud fra den maksimale fejl man vil have på sit gennemsnit, med (1-α)100% sansynlighed.

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

Konfidensinterval for gennemsnit
Intervalestimat – vi kan med 1-α sikkerhed (konfidens) antage at X ligger inden for dette interval.
σ kendt:

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

σ ukendt, n > 30:

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

σ ukendt, n < 30:

$$\bar{x} - t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

Sandsynlighed og konfidens

Sandsynligheder udtaler sig om fremtiden.
Konfidens udtaler sig om allerede målte data.

Hypotese test – generelt s. 238

Nul hypotese, H_0: $\mu = \mu_0$:
Udgangspunkt, som vi ønsker at forkaste ud fra stikprøven.

Alternativ hypotese, H_1: $\mu \neq \mu_0$:
Står i modsætning til H_0 og er den hypotese, vi ønsker at sandsynliggøre.

Ensidet alternativ:
H_1: $\mu < \mu_0$
H_1: $\mu > \mu_0$

Tosidet alternativ:
H_1: $\mu \neq \mu_0$

Fejltyper:

	Vi accepterer H	Vi forkaster H
H er sand	Korrekt beslutning	Fejl af **type 1**
H er falsk	Fejl af **type 2**	Korrekt beslutning

Type 1: Manden er uskyldig, men dømmes skyldig
Type 2: Manden er skyldig, men frikendes

P(fejl af type 1)=α , er som regel 0,05 eller 0,01 = signifikansniveauet
P(fejl af type 2)=β , kan ikke styres

Ændring af fejlen ved hypotesetest
Man kan ændre på fejlen ved hypotesetest. Dette sker ved at ændre på parametrene α eller n.
Hvis man øger n, vil man mindske sandsynligheden for fejl af både type 1 og type 2.
Hvis man øger α, mindsker man risikoen for type 1 fejl, men øger samtidig risikoen for type 2-fejl.

En **tests styrke** er defineret ved $1 - \beta$.

Inferens for Middelværdi s. 246

Hypotesetest af én middelværdi (t-test)

Bruges når:
Det er en forudsætning, at observationerne er normalfordelte.
Det skal testes, om man ud fra data kan bestemme, om den foreslåede μ_0 er sandsynlig.

Metode:

- Opstil hypoteserne H_0 (ofte $\mu = \mu_0$) og H_1, og vælg signifikansniveau α
  Husk at lighedstegnet skal stå i H_0.

- Beregn teststørrelse:

  σ kendt:
  Standardnormalfordelingen bruges:
  $$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$
  $$Z \in N(0,1^2)$$

  σ ukendt, n > 30:
  Standardnormalfordelingen bruges:
  $$Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$
  $$Z \in N(0,1^2)$$

  σ ukendt, n < 30:
  t-fordelingen bruges
  $$t = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$
  $t \in t(v)$, hvor υ = n-1

- Beregning af **kritisk værdi**:
  Slå op i t-fordelingen med υ = n – 1 frihedsgrader , z-fordelingen svarer til $v = \infty$
- Beregning af **P-værdi**:
  P-værdien er sandsynligheden for en mindst lige så ekstrem værdi som den kritiske værdi. Fås ved at slå den fundne Z-værdi op i tabellen og finde den tilsvarende sandsynlighed p = $P(Z \geq k)=1- P(Z \leq k)=$. Hvis P-værdien er mindre end α, forkastes H_0.

  Kan kun udregnes, hvis observationerne følger en normalfordeling og σ er kendt.

Inferens for Middelværdi

Hypotesetest af én middelværdi (t-test)

Metode - fortsat:

- Sammenlign teststørrelse og kritisk værdi

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 $t > t_{\alpha/2}$

- Sammenlign P-værdi og signifikansniveau.

Sammenhæng mellem hypotesetest og konfidensintervaller: s. 251

Hvis man har et (1-α) ·100% - konfidensinterval, svarer det til acceptområdet af nulhypotesen, hvis man laver en tosidet test. Dvs. at nulhypotesen accepteres, hvis den kritiske værdi ligger inden for konfidensintervallet med samme signifikansniveau.

Inferens for middelværdi

Hypotesetest for sammenligning af 2 middelværdier (t-test): s. 260

Bruges når:
Der sammenlignes gennemsnit for 2 stikprøver;

Stikprøve 1: n_1, $\overline{X}_1$, s_1^2
Stikprøve 2: n_2, $\overline{X}_2$, s_2^2

Forudsætning: x1 og x2 skal være uafhængige. Man antager at observationerne er normalfordelte.

Der er varianshomogenitet : $s_1^2 = s_2^2$. Dette kontrolleres med en F-test. Hvis der ikke er varianshomogenitet, skal man bruges "kors"-formlen i tabellen på næstsidste side.

Metode:

- Først opstilles H_0, som oftest er af formen H_0: $\mu_1 - \mu_2 = \delta$
  Som alternativ hypotese har man enten $\mu_1 - \mu_2 > \delta$, $\mu_1 - \mu_2 < \delta$ eller $\mu_1 - \mu_2 \neq \delta$

- Teststørrelsen beregnes:

  For stikprøver med kendte varianser σ_1^2 og σ_2^2

  $$Z = \frac{(X_1 - X_2) - \delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$, hvor $Z \in N(0,1^2)$

  For stikprøver $n \geq 30$, med ukendte varianser

  $$Z = \frac{(X_1 - X_2) - \delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$, hvor $Z \in N(0,1^2)$

  For stikprøver $n \leq 30$ og med ukendte varianser

  $$t = \frac{(X_1 - X_2) - \delta}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$, hvor $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ og $t \in t(\nu)$, hvor $\boldsymbol{\nu = n_1 + n_2 - 2}$

- Beregning af kritisk værdi:
  Dette gøres vha. t-fordelingstabellen, hvor man har bestemt frihedsgraden. For normalfordelingen er frihedsgraden = ∞

- Teststørrelsen og den kritiske værdi sammenlignes:

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 $t > t_{\alpha/2}$

Konfidensinterval for forskel i middelværdi

Man kan for store stikprøver beregne et (1-α) konfidensinterval for $\delta=u_1-u_2$:

$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$.

Hvis σ_1^2 og σ_2^2 kendes, bruges de i stedet for s_1^2 og s_2^2.

For små stikprøver med ukendt σ_1^2 og σ_2^2 beregnes et (1-α) – konfidensinterval for $\delta=u_1-u_2$:

$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, hvor frihedsgraden i t-fordelingen er $n_1 + n_2 - 2$.

Parret t-test s. 268

Parrede t-tests anvendes, når man har to datasæt, som er parret/uafhængige/beskriver en udvikling i populationen/stikprøven. Fx hvor mange cigaretter mødre ryger før og efter fødslen, hvor mange kilo man har tabt efter 4 uger, etc.. Det er en nødvendig forudsætning, at observationerne er normalfordelte.

Metode:

Man finder blot forskellen mellem alle sæt af værdier, D=Y-X, og behandler det som et helt almindeligt datasæt med en middelværdi, $\bar{d}$. Denne middelværdi beskriver en mulig ændring (hvis den er signifikant forskellig fra 0), og konfidensintervallet er:

$\bar{d} - t_{\alpha/2}\frac{S_D}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}$ med frihedsgrad n-1

Inferens for varians s. 281

Konfidensinterval for varians

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

Fraktilerne for χ^2 har $\upsilon = n - 1$ frihedsgrader

HUSK: konfidensintervallet er ikke nødvendigvis symmetrisk!

Hypotesetest af én varians

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

Bruges når:
Det er en forudsætning, at observationerne er normalfordelte.
Det skal testes, om man ud fra data kan bestemme, om den foreslåede σ^2 er sandsynlig.

Metode:

- Opstil hypoteser H_0 (ofte $\sigma^2 = \sigma_0^2$) og H_1, og vælg signifikansniveau (α)
  Husk at lighedstegn skal stå i H_0.
- Beregn teststørrelse: $\chi^2 = \frac{(n-1)\cdot S^2}{\sigma_0^2}$
- Beregning af kritisk værdi: slå op i χ^2-fordelingen tabel 5 med $\upsilon = n - 1$ frihedsgrader
- Sammenlign teststørrelse og kritisk værdi

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

Inferens for varians

Hypotesetest for sammenligning af to varianser, F-test s. 286

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

Bruges når:
Man har to stikprøver og vil sammenligne deres varians.
Det er antaget at data fra begge stikprøver er normalfordelt ← vigtigt!

Metode:

- Opstil hypoteser H_0 (ofte $\sigma_1^2 = \sigma_2^2$) og H_1, og vælg signifikansniveau (α)
  Husk at lighedstegn skal stå i H_0.

- Beregn teststørrelse: *se tabel nedenfor.*
  HUSK at kvadrere S'erne!

- Beregning af kritisk værdi: slå op i F-fordelingen tabel med frihedsgrader *se tabel nedenfor*

- Sammenlign teststørrelse og kritisk værdi

Alternativ hypotese	Test størrelse	Afvis nul-hypotese hvis
$\sigma_1^2 < \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F > F_\alpha(n_2 - 1, n_1 - 1)$
$\sigma_1^2 > \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F > F_\alpha(n_1 - 1, n_2 - 1)$
$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{S_M^2}{S_m^2}$	$F > F_{\alpha/2}(n_M - 1, n_m - 1)$

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

Konfidensinterval for to stikprøver hvor variansen kan antages at være ens ($=\hat{\sigma}_p$):

$$\frac{(n_1 + n_2 - 2) \cdot \hat{\sigma}_p^{\,2}}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n_1 + n_2 - 2) \cdot \hat{\sigma}_p^{\,2}}{\chi^2_{1-\alpha/2}}$$

Hvor frihedsgraden for Ki^2 er n_1+n_2-2

Inferens for andele (s. 292)

Punktestimat af andel:

$\hat{p} = \frac{x}{n}$ hvor $\hat{p} \in [0;1]$

Det kræves at n>100-200 for at få et præcist estimat.

Konfidensinterval (1-α)% for en andel:

Bruges ved en stor stikprøve.

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

Maksimal fejl på estimat: s. 296

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Hvis estimatet $\hat{p}$ bruges i stedet for parameteren p, fås et estimat af E.

Bestemmelse af stikprøvestørrelse, n: s. 296

p estimeret:

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

p ukendt (p antages at være ½, da det kræver den største stikprøve):

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

Hypotesetest for en andel: s. 298

Gælder for store stikprøver

1. H_0: $p=p_0$ H_1: $p \neq p_0$ eller $p<p_0$ eller $p>p_0$

2. Teststørrelse: $Z = \frac{X - n \cdot p_0}{\sqrt{n \cdot p_0(1-p_0)}}$

3. H_0 forkastes hvis $|Z|>z_\alpha$ ($z_{\alpha/2}$)

Hypotesetest for to andele (s. 304)

Sammenligning af andele for to forskellige binomialfordelte populationer. Kræver store stikprøver!
H_0: $p_1=p_2$

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad \textbf{hvor} \quad \hat{p} = \frac{X_1+X_2}{n_1+n_2}$$

Konfidensinterval (1-α)% for forskellen mellem to andele

p_1-p_2

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{\frac{X_1}{n_1}\left(1-\frac{X_1}{n_1}\right)}{n_1}+\frac{\frac{X_2}{n_2}\left(1-\frac{X_2}{n_2}\right)}{n_2}}$$

Hypotesetest for flere andele, antalstabeller (s. 309)

Sammenligning af andele for *flere end to* forskellige populationer, eller undersøgelse af sammenhænge med flere inddelinger. Tabellen indeholder *diskrete data*.

Fx:
Er stemmefordelingen ens 4 uger før og 2 uger før valget?
Er det de samme, som får gode karakterer i matematik og statistik?
Er der en sammenhæng mellem IQ og hårfarve?

De observerede hyppigheder er opstillet/givet i en tabel.
Start med at lægge alle værdier i de enkelte søjler og rækker sammen!

χ^2 -fordelingen er her en approksimation, derfor:
Alle de forventede hyppigheder skal være større end 5!

H_0: $p_1=p_2=\ldots=p_k$ H_1: non H_0 - dvs. altid ensidet test!
Eller:
H_0: uafhængighed mellem rækker og søjler. H_1: afhængighed – dvs. altid ensidet test!

Teststørrelse:

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

o_{ij} er den observerede værdi og e_{ij} er den forventede værdi.
e_{ij} udregnes ved at tage $\frac{\text{rækkesummen} \cdot \text{søjlesummen}}{\text{den totale sum}}$

Forkast H_0, hvis $\chi^2 > \chi_\alpha^2$(antal rækker-1)(antal søjler-1). Altid ensidet, brug altid alpha.

Goodness of fit (s. 311)

Undersøger hvor godt et datasæt følger en bestemt fordeling. Typisk er der givet to søjler, hvor den ene søjle indeholder den stokastiske variabel, og den anden søjle indeholder hyppigheder.
Man antager at datasættet har en bestemt fordeling, fx at de er normalfordelte eller poissonfordelte. Så udregnes de hyppigheder, man ville forvente, ud fra denne fordeling. (=sandsynlighederne gange det totale antal observationer)
Disse sammenlignes med de observerede værdier o_i, ved at udregne teststørrelsen χ^2.

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Hvis $|\chi^2| > \chi^2(k-1-m)$, forkastes H_0, og datasættet kan ikke antages at have denne fordeling.
k er antal kategorier / inddelinger, og m er antal estimerede parametre, som bruges i modellen.

Ex:
Hvis vi antager, at dataene følger en Poisson-fordeling med $\lambda=\lambda_0$, er m=1
Hvis vi antager, at dataene er normalfordelte, med $\mu=\mu_0$ og $\sigma^2 = \sigma_0^2$, er m=2

Hvis en forventet hyppighed er under 5, må en eller flere grupper slås sammen, så både den forventede hyppighed og den observerede hyppighed stiger.

Regressionsanalyse

Bruges når:
Det ene udfald afhænger af det andet. To kontinuerte variable.

Korrelationskoefficient s. 374

Korrelationskoefficient: $r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}) = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$ (S'erne er defineret på næste side)

Der gælder at $r \in [-1,1]$
Hvis r = 1 haves en fuldstændig lineær forbindelse med positiv hældning
Hvis r = -1 haves en fuldstændig lineær forbindelse med negativ hældning
Hvis r = 0 er den slet ikke lineær
r udtrykker graden af lineær sammenhæng

Forklaringsgraden, del af variationen der bliver dækket af modellen = r^2
Hvis $r^2 > 0,8$ er det en god model
Hvis $r^2 > 0,5$ er det en brugbar model

Regressionsanalyse

Simpel lineær model s. 338

$Y = \alpha + \beta \cdot x + \varepsilon$

Hvor
Y = afhængig kontinuert variabel
x = uafhængig kontinuert variabel
α = skæring med Y-akse
β = hældning
ε = residual (tilfældig fejl)

HUSK: at her anvendes a + bx og ikke som normalt ax + b

Mindste kvadraters metode:
Bruges når residualerne er tilfældigt spredt og ikke følger et mønster.

Går ud på at minimere den kvadratiske afstand mellem punkter og linie

Der defineres:

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

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = s_y{}^2 \cdot (n-1)$$

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

β kan estimeres med $b = \frac{S_{xy}}{S_{xx}}$

α kan estimeres med $a = \bar{y} - b \cdot x$

Excel:

- Indtast data
- Marker data
- Tryk på guiden diagram (diagram ikon)
- Vælg xy – punkt plot
- Vælg det hvor de ikke hænger sammen
- Tryk udfør
- Højreklik på en af prikkerne i diagrammet og vælg tilføj tendenslinie
- Under fanebladet type vælg lineær
- Under fanebladet indstillinger hakkes af i vis ligning og vis R-kvadreret værdi i diagram
- Tryk ok

Regressionsanalyse/inferens i regressions model

Modellens usikkerhed:
Det antages at ε er uafhængige og normalfordelte stokastiske variable med middelværdi 0 og konstant varians σ^2.

σ^2 kan estimeres med $s_e^2 = \frac{S_{yy} - (S_{xy})^2 / S_{xx}}{n-2}$

Hypotesetest om skæring med Y-aksen

Kan α fx antages at være 0?

H_0: $a = \alpha$
H_1: $a \neq \alpha$

Teststørrelsen er:

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

Kritisk værdi:
$t_{\alpha/2}(v)$
Hvor $v = n-2$

Forkast H_0 hvis $t > t_{\alpha/2}(v)$

Hypotesetest om hældningen

H_0: $b = \beta$
H_1: $b \neq \beta$
Hvis b = 0 er der ingen sammenhæng og derfor ingen model.

Teststørrelsen er:

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

Kritisk værdi:
$t_{\alpha/2}(v)$
Hvor $v = n-2$

Forkast H_0 hvis $t > t_{\alpha/2}(v)$

Regressionsanalyse/inferens i regressions model

Konfidensintervaller s. 346

Konfidensinterval for α:

$$a \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

Konfidensinterval for β:

$$b \pm t_{\alpha/2} \cdot s_e \cdot \frac{1}{\sqrt{S_{xx}}}$$

Konfidensinterval for $\alpha + \beta x_0$:
Konfidensinterval for modellen i punktet x_0.

$$(a + \beta \cdot x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Bruges til at se på hvor gennemsnittet af Y-værdier ligger. F.eks. antal kroner i gennemsnit pr. uge

Prædiktionsinterval for $\alpha + \beta x_0$:
Prædiktionsinterval for modellen i punktet x_0.

$$(a + \beta \cdot x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Bruges til at forudsige én måling. F.eks. antal kroner i næste uge.
Prædiktionsintervallet er bredere end konfidensintervallet.

Variansanalyse (Anova)

Tabellerne minder om antalstabeller, men indeholder kontinuerte målinger (Y). X inddeles i grupper.
Er der forskel i middel på grupperne?
Man skal antage at varianserne i hver gruppe er ens, og at observationerne er normalfordelte.

Ensidet variansanalyse (s. 400)

En faktor testes. (Excel: vælg Anova enkelt faktor)
Fx om
Mærkeligt nok testes forskelle i middelværdier ved at sammenligne *varianser* og deres fordeling.

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Det antages at fejlene $\varepsilon_{ij} \sim N(0, \sigma^2)$
Altså at der ikke er nogen systematik i fejlene, og at alle fejl har samme variation.

μ er gennemsnittet for alle målinger.
α_i er : i'te kategoris gennemsnit – μ. (rækkeeffekten) Det vil sige, at hvis α_i overalt er 0, er middelværdierne ens.

H_0: $\alpha_i = \alpha_j$ for alle i,j. H_1: $\alpha_i \neq \alpha_j$ mindst ét i,j

Det kan undersøges hvor meget af den totale variation (SST), der skyldes forskelle i kategoriernes middelværdier(SS(Tr)), og hvor meget der skyldes fejl(SSE).

SST=SS(Tr)+SSE

Teststørrelse F:

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

hvor N er det totale antal observationer, og k er antal kategorier eller grupper.

Forkast H_0, hvis $|F| > F_\alpha(k-1, N-k)$ (brug ALTID α, og ikke $\alpha/2$)

Sum of squares/ kvadratafvigelsessummer (hjælpestørrelser): Se også s. 404.

$$C = \frac{T_\bullet^2}{N}$$

Hvor T**.** er summen af alle observationer, og N er antal observationer,

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

Udregnes ved at kvadrere alle observationer, lægge dem sammen og trække C fra

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

Udregnes ved at lægge alle observationer sammen i i'te kategori, kvadrere denne sum, dividere den med antal observationer i kategorien, og trække C fra.
Også lig med: Den kvadrerede forskel mellem hver observation i den i'te kategori og gennemsnittet i den i'te kategori, ganget med antal observationer i den i'te kategori.

SSE = SST-SS(Tr)

Variationsanalysetabel

Variationskilde	Sum of squares	Frihedsgrader	S^2	Teststørrelse F
Behandling/Treatment	SS(Tr)	k-1	S^2_{tr}= SS(Tr)/(k-1)	S^2_{tr}/S^2_{err}
Residual/Error	SSE	N-k	S^2_{err}= SSE/(N-k)	
Total	SST	N-1		

Totale antal frihedsgrader = Frihedsgrader for behandling – Frihedsgrader for residualerne

(1-α) konfidensinterval for forskelle i middelværdi (s. 410)

$$\bar{y}_i - \bar{y}_l \pm t_{\alpha/2}\sqrt{s^2\left(\frac{1}{n_i} - \frac{1}{n_l}\right)}$$

Brug at $s^2 = S^2_{err}$= SSE/(N-k)
$\bar{y}_i$ og $\bar{y}_l$ er gennemsnit for de enkelte behandlinger.
Frihedsgraden for t er fejlens frihedsgrad, altså N-k.

Tosidet variansanalyse / randomiseret blokforsøg (s.417)

To faktorer har (muligvis) indflydelse på data. (Excel: Vælg Anova, to faktorer *uden* gentagelse).
Der er delt op i blokke, idet der er to forskellige inddelingskriterier.
Fx: hvad har indflydelse på måleresultaterne af forskellige materialers styrke: måleinstrumenterne eller materialet?

$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $\qquad \varepsilon_{ij} \sim N(0, \sigma^2)$

α_i svarer til kategoriernes indflydelse, β_j svarer til blokkenes indflydelse.
Fejlene skal igen være normalfordelte og usystematiske.

μ er gennemsnittet for alle målinger.
α_i er : i'te kategoris gennemsnit – μ.
β_j er : j'te bloks gennemsnit – μ.

Behandling:
H_0 (kategori): $\alpha_1=\alpha_2=\ldots=\alpha_a=0$ $\qquad$ H_1: non H_0

$$F_{tr} = \frac{\text{SS(Tr)/(a-1)}}{\text{SSE/(a-1)(b-1)}}$$

Forkast H_0, hvis $|F| > F_\alpha\,(a-1,(a-1)(b-1))$

Blokke:

H_0 (blok): $\beta_1 = \beta_2 = \ldots = \beta_b = 0$ H_1: non H_0

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

Forkast H_0, hvis $|F| > F_\alpha$ (b-1,(a-1)(b-1))

Det kan undersøges hvor meget af den totale variation (SST), der skyldes forskelle i kategoriernes middelværdier(SS(Tr)), forskelle i blokkenes middelværdier og hvor meget der skyldes fejl(SSE).

SST=SS(Tr)+SS(Bl) + SSE

Variationsanalysetabel s. 420

Variationskilde	Sums of squares	Frihedsgrader	S^2	**Teststørrelse F**
Behandling/Treatment	SS(Tr)	a-1	S^2_{tr}= SS(Tr)/(a-1)	$\mathbf{S^2_{tr}/S^2_{err}}$
Blokke	SS(Bl)	b-1	S^2_{bl}= SS(Bl)/(b-1)	$\mathbf{S^2_{bl}/S^2_{err}}$
Residual/Error	SSE	(a-1)(b-1)	S^2_{err}= SSE/(a-1)(b-1)	
Total	SST	N-1		

a er antal behandlinger, b er antal blokke.
Man kan behandle lige så mange faktorer man har lyst til, idet den enkelte faktors bidrag til variansen undersøges hver for sig. Den sættes altid i forhold til fejlens bidrag.
Antal faktorer = antal dimensioner

Sums of squares (s. 419)
Formler for SS(Tr), SS(Bl), SST og SSE for manuel udregning.

$$C = \frac{T_{\bullet\bullet}^{\ 2}}{ab}$$

Hvor $T_{\bullet\bullet}$ er summen af alle observationer,

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

Udregnes ved at kvadrere alle observationer, lægge dem sammen og trække C fra

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

$T_{i\bullet}$ udregnes ved at summere over de b observationer i hver treatment. $T_{i\bullet}^{\ 2}$ lægges sammen, divideres med antal blokke, og C trækkes fra.

$$SS(Bl) = \sum_{j=1}^{b} \frac{T_{\bullet j}^{\ 2}}{a} - C$$

$T_{\bullet j}$ udregnes ved at summere over de a observationer i hver blok. $T_{\bullet j}^{\ 2}$ lægges sammen, divideres med antal treatments, og C trækkes fra.

SSE = SST-SS(Tr)-SS(Bl)