Formler og regler i statistikk

ifølge lærebok Gunnar Løvås: Statistikk for universiteter og høgskoler

Kap 2. Hva er fakta om utvalget

Sentralmål

Median: midterste verdi etter sortering

Modus: hyppigst forekommende verdi

Gjennomsnitt: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Spredningsmål

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

Standardavvik: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Variasjonskoeffisisent = 100% · standardavik/gjennomsnitt

Kap 3. Sannsynlighetsregning

Utfallsrom: Mengden av alle mulige utfall i et førsøk.

Hendelse: Undermengde bestående av et eller flere mulige utfall av et forsøk.

Uniform sansynlighet: P(A) = antall gunstige utfall for hendelse A / antall mulige utfall

Relativ frekvens = antall ganger A har har inntruffet / totalt antall forsøk

Mengdelære

S: utfallsrommet

Ø: tom mengde (ingen elemeter)

Union: $A \cup B$ – alle elementer som er i A **eller** B eller i begge

Snitt: $A \cap B$ – alle elementer som er i både A **og** i B

Komplement: $\bar{A}$ innholder alle elementer som **ikke** er i A

Disjunkte mengder: A og B er disjunkte hvis der ikke har noen felles elementer

Sannsynligheter

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P\left(\overline{A \cup B}\right) = P\left(\bar{A} \cap \bar{B}\right) \qquad P\left(\bar{A} \cup \bar{B}\right) = P\left(\overline{A \cap B}\right)$$

$$A \quad og \quad B \quad disjunkte \quad \Leftrightarrow \quad P(A \cap B) = 0$$

$$A_1, A_2, ...A_n \quad er \quad parvis \quad disjunkte \quad \Leftrightarrow \quad P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$$

Betingede sannsynligheter

Def. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

Multiplikajon: $P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$

Total sannsynlighet: En og bare en av hendelsene B_1, B_2 , ..., B_n vil intreffe. Da gjelder

$$P(A) = P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + ... + P(B_n) \cdot P(A \mid B_n)$$

$$P(A) = P(B) \cdot P(A \mid B) + P(\overline{B}) \cdot P(A \mid \overline{B})$$

Bayes lov: $P(B \mid A) = \dfrac{P(B) \cdot P(A \mid B)}{P(A)}$

Annet: $P(A \mid B) + P(\overline{A} \mid B) = 1$

Uavhengige hendelser

Regel: A og B er uavhengige $\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

Multiplikasjon: Hendelsene A_1,A_2 , ..., A_n er uavhengige. Da gjelder

$$P(A_1 \cap A_2 \cap ..... \cap A_n) = P(A_1) \cdot P(A_2) \cdot ..... \cdot P(A_n)$$

$$P(A_1 \cup A_2 \cup ..... \cup A_n) = 1 - P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot ..... \cdot P(\overline{A_n})$$

Kombinatorikk

Produktregel: Forsøk i k etapper hvor hver etappe har m_i mulige utfall. Antall mulige utfall for hele forsøket blir

$m_1 \cdot m_2 \cdot \cdot \cdot \cdot m_k$.

Potensregel: Trekker k enheter av n merkede enheter med tilbakelegging. Antall mulige ordnede utvalg blir

n^k.

Permutasjoner: Trekker k enheter av n merkede enheter uten tilbakelegging. Antall mulige ordnede utvalg blir

$$P_{n,k} = n \cdot (n-1) \cdot (n-2) \cdot ..... \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

Fakultet: n forskjellige elementer kan organiseres i

$n! = n \cdot (n-1) \cdot (n-2) \cdot ..... \cdot 2 \cdot 1$ antall rekkefølger (dvs. antall ordnede utvalg av n merkede enheter)

Antall kombinasjoner: Trekker k enheter av n merkede enheter uten tilbakelegging. Totalt antall mulige ikke-ordnede kombinasjoner blir

$$C_{n,k} = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

Merk at per defefinisjon er: $0! = 1$ $\quad P_{n,0} = 1$ $\quad C_{n,0} = \binom{n}{0} = 1$

Kap 4. Stokastiske variabler

Stokastisk variabel: En stokastisk variabel X knytter et bestemt tall til et hvert utfall i utfallsrommet S

Verdimengde V_f: Mengden av alle tallverdier en stokastisk variabel X kan anta.
Diskret: V_f består av et endelig eller tellbart mengde tall
Kontinuerlig: V_f består av alle tall i et gitt intervall

Diskrete sannsynlighetsfordelinger
Alle verdier x i verdimengden opptrer med en gitt (punkt)sannsynlighet.

Punktsannsynlighet: **P(X=x)** er sannsynlighet for at den stokastiske variabelen X får verdien x

Egenskap: $\sum_{\text{alle } x \in V_f} P(X = x) = 1$

Kumulativ fordelingsfunksjon: $F(x) = P(X \leq x)$

Regler: $P(X \leq a) = F(a)$
$P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

Forventningsverdi: $\mu = E(X) = \sum_{\text{alle } x \in V_f} x \cdot P(X = x)$

Varians: $\sigma^2 = E[(X-\mu)^2] = \sum_{\text{alle } x \in V_f} (x-\mu)^2 \cdot P(X = x) \qquad \left(\sum_{\text{alle } x \in V_f} x^2 \cdot P(X = x) - \mu^2 \right)$

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

Kontinuerlige sannsynlighetsfordelinger

Sannsynligheten er gitt ved sannsynlighetstetthetfunksjon **f(x)** som oppfyller at $\int_{-\infty}^{\infty} f(x)dx = 1$

Kumulativ fordelingsfunksjon: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$

Regler: $P(X \leq a) = F(a)$
$P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

Forventningsverdi: $\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$

Varians: $\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx \qquad \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \mu^2 \right)$

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

Summasjon av stokastiske variabler (Regel 4.12 og 4.17)
Gitt stokastiske variaber $X_1, X_2, \ldots X_n$ og konstanter $a_1, a_2, \ldots a_n$ og b. Da gjelder alltid:

$E(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + b) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n) + b$

Hvis variablene $X_1, X_2, \ldots X_n$ i tillegg er statistisk uavhengige, gjelder også:

$Var(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + b) = a_1^2Var(X_1) + a_2^2Var(X_2) + \ldots + a_n^2Var(X_n)$

Kap 5. Vanlige sannsynlighetsmodeller

5.2 Binomisk modell

2 parametre: n (antall forsøk) p (sannsynlighet for suksess i hvert forsøk)

Stokastisk variabel: X (antall suksesser, 0 – n)

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

Forventning og varians: E(X) = np Var(X) = np(1-p)

Typiske eksempler: n antall terningkast X: antall 6-ere p = 1/6
Trekning fra urne <u>med</u> tilbakelegging

5.3 Hypergeometrisk modell

3 parametre: N (totalt antall enheter) M (enheter med spes. egenskap)
n (utvalg)

Stokastisk variabel: X (antall enheter i utvalget med spes. egenskap)

Punktsannsynlighet: $P(X = x) = \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

Forventning og varians: E(X) = np Var(X) = np(1-p)$\dfrac{N-n}{N-1}$ p = M/N

Typiske eksempler: - N antall enheter på lager M antall enheter med defekt
n antall stikkprøver X antall stikkprøver med defekt

- Trekning fra urne <u>uten</u> tilbakelegging (lotto)

5.5 Poissonfordelingen

Én parameter: λ forventet antall hendelser pr. enhet (ikke tid)
eller λt λ er forventet antall hendelser pr. tidsenhet, t er tidsintervall

Stokastisk variabel: X (antall hendelser i gitt enhet/tidsintervall)

Punktsannsynlighet: $P(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$ (eller $P(X = x) = \dfrac{(\lambda t)^x}{x!} e^{-\lambda t}$)

Forventning og varians: E(X) = λ Var(X) = λ (eller E(X) = λt Var(X) = λt)

Typiske eksempler: Antall anrop til en sentral pr. minutt
Antall reker i en boks rekesalat

Tilnærninger

hypergeomtrisk(N,M,n) $\approx$ binomisk(n,p) med p = M/N når N>10n

binomisk(n,p) $\approx$ poisson(λ) med λ = np når n>10 og p< 0,1

5.7 Normalfordelingen

X er normalfordelt med forventningsverdi μ og standardavvik σ: $X \sim N(\mu,\sigma)$

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

Kumulativ sannsynlighetsfordeling: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$

<u>Egenskaper:</u> Symmetri omkring μ: $F(\mu-x) = 1 - F(\mu+x)$
68 % sannsynlighet for enheter i intervallet $[\mu-\sigma,\mu+\sigma]$
95 % " - " $[\mu-2\sigma,\mu+2\sigma]$
99,7 % " - " $[\mu-3\sigma,\mu+3\sigma]$

Standard normalfordeling:
Z er standard normalfordelt med forventningsverdi $\mu = 0$ og standardavvik $\sigma = 1$, $Z \sim N(0,1)$
Kumulativ standard normafordeling $G(z) = P(Z<z)$ finnes i tabeller (D.3 side 482 i lærebok)

Regel: gitt X normalfordelt med forventningsverdi μ og standardavvik σ: $X \sim N(\mu,\sigma)$
Variabelen $Z = \frac{X-\mu}{\sigma}$ vil da være standard normalfordelt $Z \sim N(0,1)$

Beregninger: gitt X normalfordelt med forventningsverdi μ og standardavvik σ: $X \sim N(\mu,\sigma)$
$P(X<a) = F(a) = G(\frac{a-\mu}{\sigma})$
$P(a<X<b) = F(b) - F(a) = G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$
$P(X>a) = 1 - F(a) = 1 - G(\frac{a-\mu}{\sigma})$

Kvantiler (z_α) (tabell D.4 side 482 i lærebok):
Gitt en sannsynlighet α., z_α kalles da α-kvantilet og er definert som verdien z_α som gir sannsynligeten for at en tilfeldig variabel $Z \sim N(0,1)$ er større en z_α er lik α:

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

Spredningsintervall: Gitt X normalfordelt med forventningsverdi μ og st.avvik σ: $X \sim N(\mu,\sigma)$
- $100(1-\alpha)\%$ av verdiene finnes da i <u>intervallet</u> $\mu \pm z_{\alpha/2}\,\sigma$
- $100(1-\alpha)\%$ av verdiene er <u>større enn</u> $\mu - z_\alpha \cdot \sigma$
- $100(1-\alpha)\%$ av verdiene er <u>mindre enn</u> $\mu + z_\alpha \cdot \sigma$

Summasjon av normalfordelte variabler:
Gitt $X_1, X_2, \ldots X_n$ uavhengige og normalfordelte variabler $X_i \sim N(\mu_i,\sigma_i)$ $i = 1, 2, \ldots n$,
og $a_1, a_2, \ldots a_n$ er konstanter.

Variabelen $Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ er da normalfordelt $Y \sim N(\mu,\sigma)$

med $\mu = a_1 \mu_1 + a_2 \mu_2 + \ldots + a_n \mu_n$
og $\sigma^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \ldots + a_n^2 \sigma_n^2$

Spesialtilfeller (summasjon):
- Summen av n enheter av samme normalfordelte variabel $N(\mu,\sigma)$ vil være normalfordelt $N(n\mu, \sqrt{n}\,\sigma)$
- Gjennomsnittet $\overline{X}$ av n enheter av samme normalfordelte variabel $X \sim N(\mu,\sigma)$ vil være normalfordelt $\overline{X} \sim N(\mu,\sigma/\sqrt{n})$

5.8 Sentralgrenseteoremet

Sentralgrenseteoremet er et svært viktig resultat og et meget nyttig verktøy i statistikken. Det kan bevises matematisk at hvis man summerer flere variabler fra samme sannsynlighetsfordeling vil denne summen nærme seg en **normalfordeling**. Dette gjelder uansett hvilken fordeling den opprinnelige variabelen tilhørte.

Regel (5.18) La X_1, X_2, ... X_n være uavhengige variabler fra samme sannsynlighetsfordeling med forventning μ og standardavikk σ.

Da er $\overline{X} = \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)$ tilnærmet normalfordelt $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Regel (5.19) La X_1, X_2, ... X_n være uavhengige variabler fra samme sannsynlighetsfordeling med forventning μ og standardavikk σ.

Da er summen $Y = X_1 + X_2 + ... + X_n$ tilnærmet normalfordelt $Y \sim N(n\mu, \sqrt{n}\,\sigma)$

Normaltilnærming
Ved tilstrekkelig store tall kan nomalfrodelingen nyttes som en tilnærming til andre kjente fordelinger:

Regel (5.20) Hvis X er binomisk, hypergeometrisk elle poissonfordelt med forventning μ og standardavikk σ, da er X tilnærmet normalfordelt når vi forusetter at $\sigma^2 \geq 5$.

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

Binomisk: $\mu = np$ $\sigma^2 = np(1-p)$

Hypergeomertisk: $\mu = np$ $\sigma^2 = np(1-p)\frac{N-n}{N-1}$ $(p = M/N)$

Poisson: $\mu = \lambda$ $\sigma^2 = \lambda$ (alt. $\mu = \lambda t$ $\sigma^2 = \lambda t$)

<u>Tilleggsbetingelser:</u>
- Binomisk fordeling: p bør ikke være for nær 0 eller 1
- Hypergeometrisk: N mye større en n, og p = M/N bør ikke være for nær 0 eller 1
- Poissonfordeling: bør helst ha oppfylt at $\sigma^2 = \lambda t \geq 10$

Heltallskorreksjon: ved normaltilnærming til en <u>diskret</u> <u>heltallsvariabel</u> X oppnås best tilnærning ved å erstatte **x** med **x + 0,5** i kumulativ standard normalfordeling G( ).

Kap 6. Estimering og hypotesetesting

Punktestimatorer

Forventningsverdi: $\hat{\mu} = \overline{X} = \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)$ (gjennomsnitt)

Varians: $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Standardavvik: $\hat{\sigma} = S = \sqrt{S^2}$

Sannsynlighet p: $\hat{p} = \frac{X}{n}$ {ant. gunstige utfall / totalt antall forsøk}

Raten λ: $\hat{\lambda} = \frac{X}{t}$ {ant. hendelser i tidsrom t / t }

Konfidensintervaller

A. Gitt observasjoner X_1, X_2, ... X_n fra fordeling med **ukjent** μ og **kjent** σ.

100(1-α)% konfidensintervall for forventingsverdien μ er da gitt ved:

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- $z_{\alpha/2}$ er α/2-kvantilen i standard normalfordeling
- forutsetter X normalfordelt eller n > 20.

Antall observasjoner for å oppnå intervallengde L: $n \geq \left(\frac{2 \cdot z_{\alpha/2} \cdot \sigma}{L}\right)^2$

B. Gitt observasjoner X_1, X_2, ... X_n fra fordeling med **ukjent** μ og **ukjent** σ.

100(1-α)% konfidensintervall for forventingsverdien μ er da gitt ved:

$$\left[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- $t_{\alpha/2}$ er α/2-kvantilen i t-fordelingen med n-1 frihetsgrader
- forutsetter X normalfordelt eller n > 30.

Antall observasjoner for å oppnå intervallengde L: $n \geq \left(\frac{2 \cdot z_{\alpha/2} \cdot S}{L}\right)^2$ {S fra pilotstudie}

C. Binomisk modell med n forsøk og X gunstige utfall

100(1-α)% tilnærmet konfidensintervall for sannsynligheten p er gitt ved:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}\right] \qquad \left\{\hat{p} = \frac{X}{n}\right\}$$

- $z_{\alpha/2}$ er α/2-kvantilen i standard normalfordeling

- forutsetter at X er tilnærmet normalfordelt $\Leftrightarrow$ $n\hat{p}\left(1-\hat{p}\right) > 5$

Antall observasjoner for å oppnå intervallengde L: $n \geq \left(\frac{z_{\alpha/2}}{L}\right)^2$

D. Poissonfordeling med X hendelser i tidsrom t

100(1-α)% tilnærmet konfidensintervall for raten λ er gitt ved:

$$\left[\hat{\lambda} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}, \hat{\lambda} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}\right] \qquad \left\{\hat{\lambda} = \frac{X}{t}\right\}$$

- $z_{\alpha/2}$ er α/2-kvantilen i standard normalfordeling

- forutsetter at X er tilnærmet normalfordelt $\Leftrightarrow$ $\hat{\lambda}t > 10$

Hypotesetesting

p-verdi (signifikanssannsynlighet): sannsynligheten, hvis nullhypotesen H_0 er sann, for å få et resultat lik det observerte eller lengre bort fra nullhypotesen.
Generelt gjelder at hvis man finner $p < \alpha$, så kan nullhypotesen H_0 forkastes med signifikansnivå α.

A. Hypotesetest av μ når σ er **kjent** (Z-test)

Skal teste forventingsverdi til et materiale i forhold til en gitt verdi μ_0.
Gitt observasjoner $X_1, X_2, \ldots X_n$, med forutsetning om at X er normalfordelt eller n > 20, og valgt signifikansnivå α.

- Testobservator $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

- Finner z_α (α-kvantilen i standard normalfordelingen)
  eller $z_{\alpha/2}$ (α/2-kvantilen for 2-sidig test)

Test 1:	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$	Forkast H_0 hvis $Z > z_\alpha$	$p = 1 - G(Z)$
Test 2:	H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	Forkast H_0 hvis $Z < -z_\alpha$	$p = G(Z)$
Test 3:	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	Forkast H_0 hvis $\lvert Z\rvert > z_{\alpha/2}$	$p = 2G(Z)$ hvis $Z<0$ $p = 2G(-Z)$ hvis $Z>0$

B. Hypotesetest av μ når σ er u**kjent** (t-test)

Skal teste forventingsverdi til et materiale i forhold til en gitt verdi μ_0.
Gitt observasjoner $X_1, X_2, ... X_n$, med forutsetning om at X er normalfordelt eller n > 30, og valgt signifikansnivå α.

- Punktestimatorer for forventning og standardavvik $\overline{X}$ og S
- Testobservator $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$
- Finner t_α (α-kvantilen i t-fordelingen med n-1 frihetsgrader)

Test 1:	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$	Forkast H_0 hvis $T > t_\alpha$	$p < \alpha$ hvis $T > t_\alpha$
Test 2:	H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	Forkast H_0 hvis $T < -t_\alpha$	$p < \alpha$ hvis $T < -t_\alpha$
Test 3:	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	Forkast H_0 hvis $\lvert T\rvert > t_{\alpha/2}$	$p < \alpha$ hvis $\lvert T\rvert > t_{\alpha/2}$

C. Hypotesetest av sannsynlighet p

Skal teste sannsynlighet p i en binomisk modell i forhold til en gitt verdi p_0.
Gitt n forsøk med X antall ønskede resultat, med forutsetning om at X er tilnærmet normalfordelt <=> np(1-p) > 5, og valgt signifikansnivå α.

- Punktestimator for sannsynlighet $\hat{p} = \dfrac{X}{n}$ {ant. gunstige utfall / totalt antall forsøk}
- Testobservator $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$
- Finner z_α (α-kvantilen i standard normalfordelingen)
  eller $z_{\alpha/2}$ ($\alpha/2$-kvantilen for 2-sidig test)

			signifikanssannsynlighet (p-verdi)
Test 1:	H_0: $p \leq p_0$ H_1: $p > p_0$	Forkast H_0 hvis $Z > z_\alpha$	$p = 1 - G(Z)$
Test 2:	H_0: $p \geq p_0$ H_1: $p < p_0$	Forkast H_0 hvis $Z < -z_\alpha$	$p = G(Z)$
Test 3:	H_0: $p = p_0$ H_1: $p \neq p_0$	Forkast H_0 hvis $\lvert Z\rvert > z_{\alpha/2}$	$p = 2G(Z)$ hvis $Z<0$ $p = 2G(-Z)$ hvis $Z>0$

Kap 7. Analyse av sammenhenger

Korrelasjon

Gitt n observasjonspar (X_1,Y_1), (X_2,Y_2), . . . , (X_n,Y_n). Kovarians (samvariasjon) beregnes ved

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Korrelasjonskoeffisienten (r eller R) utrykker graden av lineær (rettlinjet) samsvar mellom X- og Y verdier, og beregnes ved

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}.$$

<u>For r gjelder:</u>

- verdien er alltid mellom – 1 og 1 (|r| < 1,0)
- |r| nær 0 angir ingen eller liten lineær sammenheng, |r| nær 1 angir sterk lineær sammenheng
- r < 0 angir at Y avtar med voksende X
  r > 0 angir at Y øker med voksende X

Korrelasjon utrykkes ofte ved r^2 (eller R^2) ,
(eksempler: r = -0,7 => r^2 = 0,49 r = 0,9 => r^2 = 0,81)

Enkel lineær regresjon

Minste kvadraters rette linje

Gitt n observasjonspar (X_1,Y_1), (X_2,Y_2), . . . , (X_n,Y_n). Minste kvadraters metode gir den linjen som er best tilpasset observasjonene ved

$$Y = a + bX$$

$$\text{med} \quad b = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = r \cdot \frac{S_Y}{S_X}$$

$$\text{og} \quad a = \overline{Y} - b\overline{X}$$

Kap 8. Sammenligning av grupper

Randomisering: sikre representativt utvalg for analyse ved tilfeldig utvalg og rekkefølge, anonymisering, blind og dobbelt blind undersøkelse.

Paring og blokking: Inndeling av observasjonene i par, eller i grupper som representer blokker.

Analyser av normalfordelte data:
- **T-test:** paret eller uparet sammenligning av forventningsverdi i to grupper
- **Variansanalyse (ANOVA):** sammenligning av forventningsverdi i flere enn to grupper

Analyser av <u>ikke</u> normalfordelte data:
Mann-Whitney-Wilcoxon, Paret Wilcoxon, Fortegnstest, Kruskal-Wallis og Friedman

Analyse av gruppelikhet i krysstabeller (spørreundersøkelser):
Kji-kvadrattest

T-test for sammenligning av forventet gjennomsnitt for to grupper

Gitt observasjoner $X_1, X_2, ... X_{n1}$ og $Y_1, Y_2, ... Y_{n1}$ med forutsetning om at X-er og Y-er er normalfordelte stokastiske variabler. Skal undersøke om gruppenes gjennomsnitt, μ_1 for X og μ_2 for Y, er forskjellige.

A. Uparet t-test med antatt like varianser

Antagelse: X og Y har samme varians

Interpolert varians: $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Testobservator: $$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Forkastningsgrense: t_α (α-kvantilen i t-fordelingen med $n_1 + n_2 - 2$ frihetsgrader)
eller $t_{\alpha/2}$ ($\alpha/2$-kvantilen for 2-sidig test)

B. Uparet t-test med antatt ulike varianser

Forutsetning: $n_1 > 30$ og $n_2 > 30$

Testobservator: $$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}}$$

Forkastningsgrense: t_α (α-kvantilen i t-fordelingen med $n_1 + n_2 - 2$ frihetsgrader)
eller $t_{\alpha/2}$ ($\alpha/2$-kvantilen for 2-sidig test)

C. Paret t-test

Forutsetning: Data i par (X_i, Y_i) for $i = 1,..,n$

Differanser: $D = X_i - Y_i$

Testobservator: $$T = \frac{\overline{D}}{S_D / \sqrt{n}}$$

Forkastningsgrense: t_α (α-kvantilen i t-fordelingen med n-1 frihetsgrader)
eller $t_{\alpha/2}$ ($\alpha/2$-kvantilen for 2-sidig test)

- -

<u>Test 1:</u>	H_0: $\mu_1 \le \mu_2$	H_1: $\mu_1 > \mu_2$	Forkast H_0 hvis $T > t_\alpha$	$p < \alpha$ hvis $T > t_\alpha$
<u>Test 2:</u>	H_0: $\mu_1 \ge \mu_2$	H_1: $\mu_1 < \mu_2$	Forkast H_0 hvis $T < -t_\alpha$	$p < \alpha$ hvis $T < -t_\alpha$
<u>Test 3:</u>	H_0: $\mu_1 = \mu_2$	H_1: $\mu_1 \ne \mu_2$	Forkast H_0 hvis $\lvert T\rvert > t_{\alpha/2}$	$p < \alpha$ hvis $\lvert T\rvert > t_{\alpha/2}$

<u>Appendix:</u> Oversikt over statistikkfunksjoner på kalkulator hp 30S

[mode] - 1) STAT - 1-VAR: gir $\overline{x}$ og S_x mm.

[mode] - 1) STAT - 2-VAR: gir **r**, **a** og **b** mm. (korrelasjon og regresjon)

[mode] - 1) STAT - CLR-DATA: nullstiller statistikkminnet

[PRB] gir kombinatorikk: nPr = $P_{n,r}$ nCr = $C_{n,r} = \binom{n}{r}$ n! (fakultet)