

Kræsjkurs i STAT101

Noen anbefalinger

- Regn mange(5-10) oppgavesett til eksamen.
- Legg vekt på å forstå hva formlene brukes til, det vil si når, og hvordan?
- Lær sammenhengen mellom fordelingene og tema i pensum.
- På eksamen vil noen formler være tilgjengelig sammen med oppgavesettet
- Selv om du ikke trenger å pugge alle formlene, er der en fordel å husk dem utenat.
- Lær deg å bruke kalkulatoren din,

Statistikk: Læren om å samle inn, **presentere og analysere data**

Typiske anvendelser

- **Beskrive og oppsummere** data fra et utvalg
 - Eksempel: Tabeller, gjennomsnitt, osv.
- **Teste hypoteser** (dvs. trekke slutning fra et utvalg til en populasjon)
 - Eksempel: Gruppforskjeller
- Gjøre analyser i data for å **avdekke mønstre eller strukturer**
 - Eksempel: Hvilke personlighetstrekk hører sammen?

Eksperiment → → Statistisk modell → → Analyse
(observasjon) (Beskrive og tolke resultatet)

Tallmaterialet

Observerte data:

$$X : x_1, x_2, x_3, \dots, x_n$$

Målemodell (**store talls lov**)

Uavhengige malinger med same forventet verdi og standardavvik.

Diskriptiv statistikk

$$\mu \approx \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \sigma \approx s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Tallmaterialet

Tallmaterialet kan bestå av:

- Få observasjoner
- Mange observasjoner:

- 1) Gruppert materiale: Mange observasjonsverdier, men få antall observasjoner.
- 2) Klasseinndelt materiale: Mange observasjoner og mange variabelverdier.

Systematisering:

- Hyppighetstabell(frekvenstabell)
- Stolpediagram (få observasjoner), Histogram (mange observasjoner)

Sentralmål: Disse viser hvordan observasjoner er konsentrert.

- Gjennomsnitt (middelverdi) $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n) = \frac{1}{n} \sum_{i=1}^n X_i$

- Median (Md: den midterste observasjonen)

Hvor ligger medianen? $m=(n+1)/2$

n: odde tall: Md= den midterste observasjonen

n: Jamn tall : Md = gjennomsnitt av de to midterste observasjonene

- Empirisk Varians: $s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

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

Stokastisk forsøk, utfall, utfallsrom

2.1 Sannsynlighet

Et **eksperiment** er en situasjon som involverer sjanse eller sannsynlighet som fører til resultater som heter utfall. Et **utfall** er resultatet av en enkelt prøve av et eksperiment. En **hendelse** eller en begivenhet er én eller flere utfall av et eksperiment eller et forsøk. Utfallsrom er mengden av alle mulige utfall i en hendelse.

Et Stokastisk forsøk (tilfeldig eksperiment) har følgende egenskaper:

- 1) Resultatet kan ikke forutsies.
- 2) Utfallsrommet kan angis.
- 3) Forsøket kan gjentas.

$$P(A) = \frac{n_A}{n_\Omega} = \frac{g}{m}$$

Stokastisk forsøk, utfall, utfallsrom

Et **utfallsrom** U inneholder alle mulige utfall som et forsøk kan få: $U = \{u_1, u_2, \dots, u_n\}$

Hvert utfall u_i har en sannsynlighet $P(u_i)$ for å framstå. For sannsynlighetene gjelder at

- $0 \leq P(u_i) \leq 1$
- $P(u_1) + P(u_2) + \dots + P(u_n) = 1$ eller $\sum_{i=1}^n P(u_i) = 1$

Et forsøk som tilfredsstiller disse kravene har en *sannsynlighetsmodell*.

De enkelte utfallene i utfallsrommet og deres sannsynlighet utgjør en **sannsynlighetsmodell**. **Sannsynlighetsfordeling** anvendes i statistikk for å beskrive hvordan stokastiske variable, for eksempel tilfeldige utvalg, fordeler seg.

Sannsynlighetsmodell:

u_1	u_2	\dots	u_n
$P(u_1)$	$P(u_2)$	\dots	$P(u_n)$

Hvis alle utfallene har lik sannsynlighet kalles det en *uniform sannsynlighetsmodell*.

For eksempel mynkast eller terningskast.

Eksempler:

- A. En mynt kast $U = \{M, K\}$ og to mynt kast: $U = \{MM, MK, KM, KK\}$
- B. En terningskast? $U = \{1, 2, 3, 4, 5, 6\}$
- C. Tre barnsfamilie $U = \{GGG, GGI, GJG, GJJ, JJJ, JJG, JGJ, JGG\}$

Stokastisk forsøk, utfall, utfallsrom

Eksempel 2.1

Hva er sannsynligheten for en 2 barnsfamilie har 2 gutter?

Utfallsrommet: $U = \{GG, GJ, JG, JJ\}$ og hendelsen $A = \{GG\}$ gir $P(A) = \frac{n_A}{n_U} = \frac{g}{m} = \frac{1}{4}$

Addisjon setningen

La A og B være to hendelser. Vi ønsker å bestemme sannsynligheten for at hendelsen A eller B forekommer. Summen $P(A) + P(B)$ får vi med sannsynlighetene for alle utfallene som er med i A eller i B eller i begge, men de som er i kommer med to ganger.

Addisjonssetning kan da skrives som:

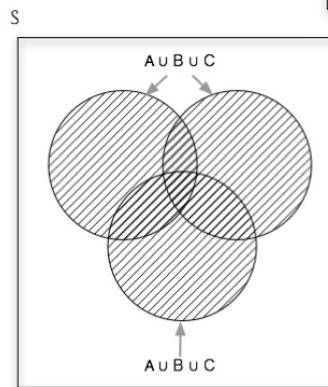
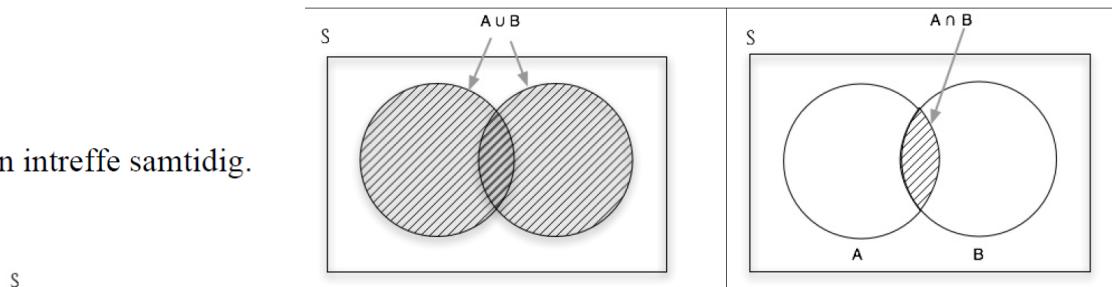
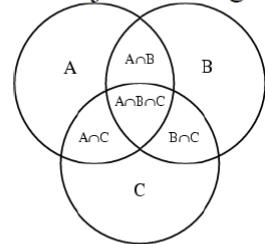
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

To hendelser er disjunkte dersom begge hendelsene ikke kan intreffe samtidig.

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

(disjunkte hendelser: $A \cap B = \emptyset$)

Addisjonssetning for 3 hendelser blir det:



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

2.6 Betinget sannsynlighet

Betinget sannsynlighet er sannsynligheten for at noe skal skje gitt at en annen ting har skjedd. Sannsynligheten for en hendelse må ofte modifiseres dersom en får oppgitt informasjon om en annen hendelse. Notasjonen som brukes er $P(A | B)$ som angir sannsynligheten for A gitt at B har inntruffet.

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

Sannsynligheten for Hendelsen A gitt B:

Dette kan medføre: $P(A \cap B) = P(A | B)P(B)$

2.7 Bayes lov

Ved å bruke $P(A \cap B) = P(A | B)P(B)$ får vi:

$$\text{Sannsynligheten for Hendelsen B gitt A: } P(B | A) = \frac{P(A | B)P(B)}{P(A)} \quad (\text{Bayes regel})$$

91% for at en 70 år gammel kvinne skal bli minst 75 år

A = "kvinnen blir minst 70 år"

B = "kvinnen blir minst 75 år"

2.7.1 Uavhengige hendelser

Hendelsene A og B sies å være uavhengige dersom en opplysning om at A har inntruffet ikke endrer sannsynligheten for at B skal inntreffe. Dette kan formuleres slik:

$P(A)$ - sannsynlighet for hendelse A

$P(B)$ - sannsynlighet for hendelse B

$P(A | B)$ - sannsynligheten for hendelse A gitt at B har inntruffet.

$P(B | A)$ - sannsynligheten for hendelse B gitt at A har inntruffet.

Dersom A og B er uavhengige gjelder:

$$P(A | B) = P(A) \quad , \quad P(B | A) = P(B) \quad \text{og} \quad P(A \cap B) = P(A)P(B)$$

Lov om total sannsynlighet

2.8 Lov om total sannsynlighet

Betrakt de disjunkte hendelsene :

$$A_1 \cup A_2 \cup \dots \cup A_n = \Omega \text{ (utfallsrommet, det vil si: } P(A_1) + P(A_2) + \dots + P(A_n) = 1)$$
$$A_1 \cap A_2 \cap \dots \cap A_n = \emptyset$$



Da gjelder det:

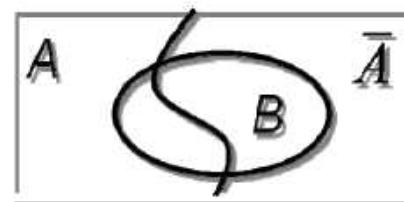
$$P(B) = P(B | A_1)P(A_1) + P(B | A_2)P(A_2) + \dots + P(B | A_n)P(A_n)$$



Dersom A og B ikke er disjunkte hendelser, gjelder det:

Total sannsynlighet :

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



Oppgave 2.8

På en videregående skole er det 20 lærere hvorav 12 er kvinner.

Rektoren har utnevnt et lærerutvalg som skal undersøke trivsel på skolen blant elevene.

Vi tenker oss en person blant lærerne valgt ut tilfeldig.

Sannsynligheten for at personen er med i utvalget gitt at det er en kvinne er 50% mens tilsvarende for menn er 25%. Vi definerer følgende hendelser:

K = Personen er en kvinne

M = Personen er en mann

U = Personen er i utvalget

- a) Sett opp sannsynlighetene $P(K)$, $P(M)$, $P(U | K)$ og $P(U | M)$.
- b) Finn sannsynligheten for at en tilfeldig lærer er i utvalget.
- c) Finn sannsynligheten for at et medlem av utvalget er en kvinne.
- d) Finn sannsynligheten for at en lærer som ikke er i utvalget er en mann.

2.8

a) $P(K) = \frac{12}{20} = 0.60 = 60\%$ $P(M) = 1 - P(K) = 0.40 = 40\%$

$P(U|K) = 50\%$ $P(U|M) = 25\%$

b) $P(U) = P(U|K).P(K) + P(U|M).P(M) = 0.50 \cdot 0.60 + 0.25 \cdot 0.4 = 0.4 = 40\%$

c) $P(K|U) = \frac{P(K \cap U)}{P(U)} = \frac{P(U|K).P(K)}{P(U)} = \frac{0.50 \cdot 0.60}{0.40} = \underline{\underline{0.75 = 75\%}}$

d) $P(M|\bar{U}) = \frac{P(M \cap \bar{U})}{P(\bar{U})} = \frac{P(M) - P(M \cap U)}{1 - P(U)} = \frac{0.40 - 0.25 \cdot 0.40}{1 - 0.40} = \frac{0.30}{0.60} = \underline{\underline{0.50 = 50\%}}$

Diskrete fordelinger

3.1 Sannsynlighetsmodell

- En sannsynlighetsmodell for et tilfeldig forsøk gir sannsynligheten for hvert enkelt utfall i utfallsrommet
- Sannsynligheten for alle utfall i utfallsrommet er til sammen 1.
- Sannsynligheten for hvert enkelt utfall er mellom 0 og 1.

a) **Binomisk fordeling**
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots, n.$$

$$E(X) = np \quad Var(X) = np(1-p)$$

b) **Hypergeometrisk fordeling**
$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, \dots$$

$$E(X) = np \quad Var(X) = \frac{N-n}{N-1} np(1-p) \quad \text{der } p = \frac{M}{N}$$

c) **Poissonsfordeling**

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-(\lambda t)} \quad x = 0, 1, 2, \dots, (\infty) \quad E(X) = \lambda t, \quad Var(X) = \lambda t$$

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0, 1, 2, \dots, (\infty) \quad E(X) = \lambda, \quad Var(X) = \lambda$$

Forventet Verdi og varians

Forventet verdi $\mu = E(X) = \sum_{\text{all } x} x_i p(x_i)$.

Regneregler for forventet verdi

$$\begin{aligned}E(b) &= b \\E(aX + b) &= aE(X) + b \\E(aX_1 + a_2 X_2) &= aE(X_1) + a_2 E(X_2)\end{aligned}$$

Variansen til en stokastisk variabel er et ikke-negativt tall som gir en pekepinn på hvor stor spredning verdiene av tilfeldige variable er sannsynlig å være, jo større variansen, jo mer spredt observasjonene i gjennomsnitt.

Angivelse av variansen gir et inntrykk av hvor nært konsentrert rundt den forventede verdien fordelingen er, det er et mål på "spredning" av en fordeling om dens gjennomsnittsverdi.

Avvik er symbolisert $Var(X)$ eller σ^2 . Variansen til den tilfeldige variabelen X er definert å være:

$$\sigma^2 = Var(X) = E((x_i - \mu)^2)$$

$$\sigma^2 = Var(X) = \sum_{i=1}^n (x_i - \mu)^2 p(x_i) = \sum_{i=1}^n x_i^2 p(x_i) - \mu^2$$

Regneregler for varians

$$\begin{aligned}Var(b) &= 0 \\Var(aX + b) &= a^2 Var(X) \\Var(aX_1 + a_2 X_2) &= a_1^2 Var(X_1) + a_2^2 Var(X_2)\end{aligned}$$

Oppgave 3.1

La X være antall støvsuger som blir hos Elkjøp i Rådal.
Samstsynlighetsfordelingen for X er gitt ved:

X	0	1	2	3	4
$P(X=x)$	0,1	0,4	0,1	0,2	p

- a) Bestem p .

$$\text{Regn ut } P(X \leq 1) \text{ og } P(1 \leq X \leq 3).$$

Regn ut forventning til X og varians til X .

Vi observerer så salget for to uker. La i denne forbindelse X_1 og X_2 stå for antall solgte solgte støvsugere i henholdsvis uke 1 og uke 2. Vi antar at X_1 og X_2 er uavhengige stokastiske variable.

- b) Regn ut $P((X_1 = 2) \cap (X_2 = 1))$, og deretter $P(X_1 + X_2 = 3)$.

- c) Firmaet er interessert i å studere summen og forskjellen på to uker. Regn i den forbindelse ut $E(X_1 + X_2)$, $\text{Var}(X_1 + X_2)$, $E(X_1 - X_2)$ og $\text{Var}(X_1 - X_2)$.

$$\text{a)} \quad p = 1 - (0,1 + 0,4 + 0,1 + 0,2) = \underline{\underline{0,2}}$$

$$P(X \leq 1) = \underline{\underline{0,5}}, \quad P(1 \leq X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = \underline{\underline{0,7}},$$

$$E(X) = \sum_{\text{alle } x} x_i p(x_i) = 0,1 \cdot 1 + 0,4 + 2 \cdot 0,1 + 3 \cdot 0,2 + 4 \cdot 0,2 = \underline{\underline{2}},$$

$$\begin{aligned} \sigma^2 &= \text{Var}(X) = \sum_{i=1}^n (x_i - \mu)^2 p(x_i) = \sum_{i=1}^n x_i^2 p(x_i) - \mu^2 \\ &= 0^2 \cdot 0,1 + 1^2 \cdot 0,4 + 2^2 \cdot 0,1 + 3^2 \cdot 0,2 + 4^2 \cdot 0,2 - 2^2 = \underline{\underline{1,8}} \end{aligned}$$

X	0	1	2	3	4
$P(X=x)$	0,1	0,4	0,1	0,2	p

- b) $P((X_1 = 2) \cap (X_2 = 1)) = P(X_1 = 2) \cdot P(X_2 = 1) = \underline{\underline{0,04}}$

$$\begin{aligned} P(X_1 + X_2 = 3) &= P((X_1 = 0) \cap (X_2 = 3)) + P((X_1 = 1) \cap (X_2 = 2)) \\ &\quad + P((X_1 = 2) \cap (X_2 = 1)) + P((X_1 = 3) \cap (X_2 = 0)) \\ &= 0,12 \end{aligned}$$

$$\begin{aligned} \text{b)} \quad E(X_1 + X_2) &= E(X_1) + E(X_2) = \underline{\underline{4}}, \quad \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = \underline{\underline{3,6}} \\ E(X_1 - X_2) &= E(X_1) - E(X_2) = \underline{\underline{0}}, \quad \text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2) = \underline{\underline{3,6}} \end{aligned}$$

Poisson fordeling

Antall ganger hendelsen A inntreffer (X) er poissonfordelt hvis:

- 1) Antall hendelser A er uavhengige av hverandre.
- 2) Forventet antall A pr. tidsenhet (eller volumenhet, lengdeenhet,...) er konstant ($= \lambda$).
- 3) A kan ikke inntreffe flere ganger helt samtidig/oppå hverandre.

I løpet av de neste t tidsenhetene vil vi observere A, X ganger.

Da er X poissonfordelt (skrivemåte: $X \sim PO(\lambda t)$ og hvis $t = 1$, skrives $X \sim PO(\lambda)$), med sannsynlighetsfordeling:

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-(\lambda t)} \quad x = 0, 1, 2, \dots (\infty) \quad E(X) = \lambda t, \quad Var(X) = \lambda t$$

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0, 1, 2, \dots (\infty) \quad E(X) = \lambda, \quad Var(X) = \lambda$$

Eks.

Eksempel 3.3

Et bilutleiebyrå som disponerer i alt 4 biler, leier ut biler for en dag om gangen. Antall (X) biler som kunder ønsker å leie pr. dag, har vist seg å være Poisson-fordelt med sannsynlighetsfunksjon (punktssannsynligheter):

$$p(x) = P(X = x) = \frac{3^x}{x!} e^{-3} \quad x = 0, 1, 2, \dots$$

Beregn sannsynlighetene for at henholdsvis 0, 1, 2, 3 og 4 biler vil være uteid en gitt dag.

Løsn.

$$P(X = 0) = \frac{3^0}{0!} e^{-3} = e^{-3} \approx 0.0498 \quad P(X = 1) = \frac{3^1}{1!} e^{-3} = e^{-3} \approx 0.1494$$

$$P(X = 2) = \frac{3^2}{2!} e^{-3} \approx 0.224 \quad P(X = 3) = \frac{3^3}{3!} e^{-3} \approx 0.224$$

$$P(X = 4) = 1 - P(X \leq 3) \approx 0.3528 \quad (\text{husk at } \sum_{i=0}^4 P(X = i) = 1)$$

Bemerkt at for å bestemme $P(X = 4)$ kan vi ikke benytte formelen for Poisson-fordeling siden summen av sannsynlighetene skal være like 1.

$\text{PO}(\lambda) \rightarrow \text{Bin}(n, p)$

3.6 Tilnærming av binomisk fordeling til Poisson fordeling

Binomisk fordeling kan tilnærmes Poisson-fordelingen når n er stor og p er liten.

$$\lim_{k \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} (np)^x n^{n-x} (1-p)^{n-x} = \frac{e^{-np} (np)^x}{x!} = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\lim_{n \rightarrow \infty} [1 - (\lambda/n)]^n = e^{-\lambda}$$



Eks. 3.3

Det er 2 defekter i et parti med 1000 komponenter.

Vi velger 10 komponenter tilfeldig.

Hva er sannsynligheten for at minst 1 av disse er defekte?



Dette er en binomisk fordeling: $\text{Bin}(n=10, p = \frac{2}{1000} = 0,002)$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{10}{0} (0,002)^0 (0,998)^{10} \approx 0,019821 \approx 2\%$$

Vi kan også benytte tilnærming til en Poisson-fordeling siden p er så liten.

$$\text{Bin}(n, p) \xrightarrow{p \text{ liten}} \text{PO}(\lambda = np)$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{(0,02)^0}{0!} e^{-0,02} = 1 - e^{-0,02} \approx 0,019801 \approx 2\%$$

Kontinuerlige fordelinger

Sannsynlighetsfunksjonen $f(x)$ tilfredsstiller $\int_{-\infty}^x f(x)dx = 1$ der $f(x) \geq 0$. Den kumulative

fordelingsfunksjonen gir oversikt over sannsynlighetsfordeling for en stokastisk variabel. For en stokastisk variabel X , med sannsynlighetsfunksjonen $P(x)$, defineres den kumulative fordelingsfunksjonen $F(x)$ som: $F(x) = P(X \leq x)$

Fordelingsfunksjon $F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$ og bemerk: $P(a \leq x \leq b) = \int_a^b f(x)dx$

Forventet verdi:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx .$$

Regneregler:

$$E(b) = b$$

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

$$E(aX_1 + a_2 X_2) = aE(X_1) + a_2 E(X_2)$$

Varians – kontinuerlige fordelinger

$$\sigma^2 = Var(X) = E((x_i - \mu)^2)$$

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx \right)^2$$

Varians:

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx \right)^2$$

Regneregler for varians

$$Var(b) = 0$$

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

$$Var(a_1 X_1 + a_2 X_2) = a_1^2 Var(X_1) + a_2^2 Var(X_2)$$

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

Her er G fordelingsfunksjonen for g , hvor $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ der g er også en normalfordeling og kalles **standard normalfordeling** og har $\mu=0$ og $\sigma=1$.

 $X \in N(\mu, \sigma)$ kan standardiseres ved $Z = \frac{X - \mu}{\sigma}$

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

- $P(X < a) = G\left(\frac{a - \mu}{\sigma}\right)$
- $P(X > a) = 1 - G\left(\frac{a - \mu}{\sigma}\right)$
- $P(a \leq X \leq b) = G\left(\frac{b - \mu}{\sigma}\right) - G\left(\frac{a - \mu}{\sigma}\right)$
- $G(-a) = 1 - G(a)$

4.7 Sentralgrensesetning

Dersom X_1, X_2, \dots, X_n er uavhengige identisk fordelt stokastiske variable med forventning μ og varians σ^2 , , det vil si: $X_i \in N(\mu, \sigma)$, $i = 1, 2, \dots, n$
 så er $S = \sum_{i=1}^n X_i$ samt middelverdien $\bar{X} = \frac{S}{n}$ tilnærmet normalfordelt dersom n er tilstrekkelig stor.

$$\boxed{\begin{aligned} S &\in N(n\mu, \sigma\sqrt{n}) , i = 1, 2, \dots, n & \text{der } S = \sum_{i=1}^n X_i \\ \bar{X} &\in N(\mu, \frac{\sigma}{\sqrt{n}}) , i = 1, 2, \dots, n & \text{der } \bar{X} = \frac{S}{n} \end{aligned}}$$

Det vil si:

$$P(S \leq b) = G\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right)$$

Bevis

$$P(\bar{X} \leq b) = G\left(\frac{b - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$\begin{aligned} E(S) &= E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) = \underbrace{\mu + \mu + \dots + \mu}_{n \text{ ganger}} = n\mu \\ Var(S) &= Var(X_1 + X_2 + \dots + X_n) = Var(X_1) + Var(X_2) + \dots + Var(X_n) \\ &= \underbrace{\sigma^2 + \sigma^2 + \dots + \sigma^2}_{n \text{ ganger}} = n\sigma^2 \end{aligned}$$

Dermed standardavviket er $SD(S) = \sqrt{n} \cdot \sigma$

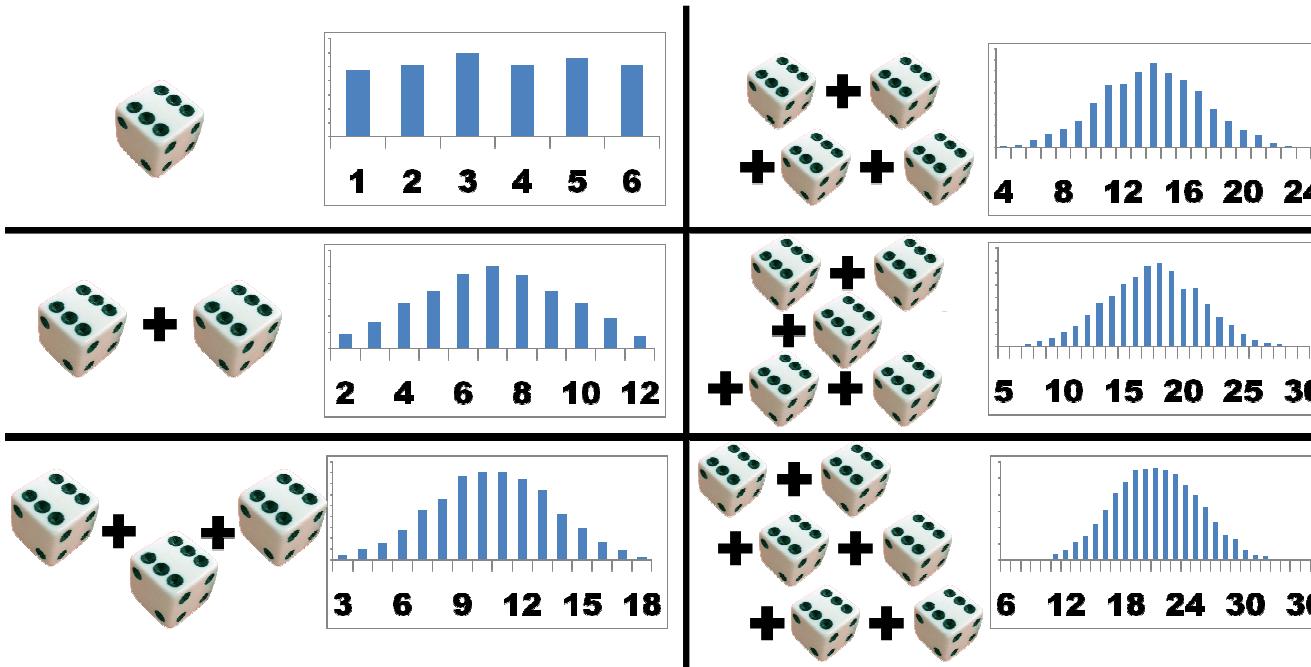
Eksempel – Sentralgrense teoremet

From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of *UCLA* against *USC* follows a distribution that has mean $\mu = 2.4$ and standard deviation $\sigma = 2.0$. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

We are given that $\mu = 2.4$, $\sigma = 2$, $n = 100$. There are 250 tickets available, so the 100 students will be able to purchase the tickets they want if all together ask for less than 250 tickets. The probability for that is $P(T < 250) = P(z < \frac{250 - 100(2.4)}{\sqrt{100}2}) = P(z < 0.5) = 0.6915$.

Normalfordeling og tilnærming

Diskret fordeling → Normalfordeling



Estimering av en ukjent parameter

Tre krav til estimatorer

- Estimatoren skal være forventningsrett,
- Estimatoren skal ha minst mulig varians (evt. standardavvik)
- Estimatoren sin varians (evt. standardavvik) skal gå mot null når størrelsen på utvalget øker.

$\hat{\mu} = \bar{X}$ Gjetter på at utvalget representerer virkeligheten.

$\hat{\mu}$ = estimatoren for μ

$$E(\hat{\mu}) = \mu \quad \bar{X} = \hat{\mu} \quad E(\bar{X}) = \mu$$

$$\hat{p} = \frac{X}{n} \quad X \sim B(n, p)$$

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = E\left(\frac{1}{n} \cdot X\right) = \frac{1}{n} E(X) = \frac{1}{n} n \cdot p = p$$

Estimering av en ukjent parameter

Forventningsrette estimatorer:

Anta at man skal estimere en eller annen parameter, t.d. θ , ved hjelp av estimatoren $\hat{\theta}$.

Siden $\hat{\theta}$ er en funksjon av utvalget, er den selv en tilfeldig variabel, og dermed har den også en forventning.

Dersom $E(\hat{\theta}) = \theta$, sier vi at estimatoren er forventningsrett. Denne egenskapen betyr at i det lange løp vil du verken underestimere eller overestimere θ dersom du bruker $\hat{\theta}$. Du gjør med andre ord ingen systematiske feil.

Eks: $E(\bar{X}) = \mu$

Dersom man velger $\bar{X} = \hat{\mu}$ får man et forventningsrett estimat av μ .

Standardavvik til estimatorene

Estimatorene har en usikkerhet, representert ved deres standardavvik. Er dette stort, er estimatoren usikker og dermed dårlig.

$$Var(\bar{X}) = Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

$$SD(\bar{X}) = SD(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

Normalfordeling og tilnærming

4.8 Oppsummering

Normalfordelingen er interessant fordi:



1. Den *forekommer ofte* i naturen. (Naturlig variasjon, avvik i målinger osv.)
2. Andre fordelinger, som er vanskelige å regne med eller lage tabeller for, kan *tilnærmes* med normalfordeling:

Binomisk fordeling $Bin(n, p)$, når $np \geq 10$ og $np(1 - p) \geq 10$

Hypergeometrisk fordeling $Bin(M, N, n)$, når $n\theta(1 - \theta) \geq 10$ der $\theta = \frac{M}{N}$

Poisson-fordeling $Po(\lambda)$ når $\lambda \geq 10$

3. *Sentralgrenseteoremet* sier at fordelingen av *summer* og *gjennomsnittet* i uttrukne stikkprøver er tilnærmet normalfordelt under visse betingelser. (Selv om den originale fordelingen *ikke* er normal!)

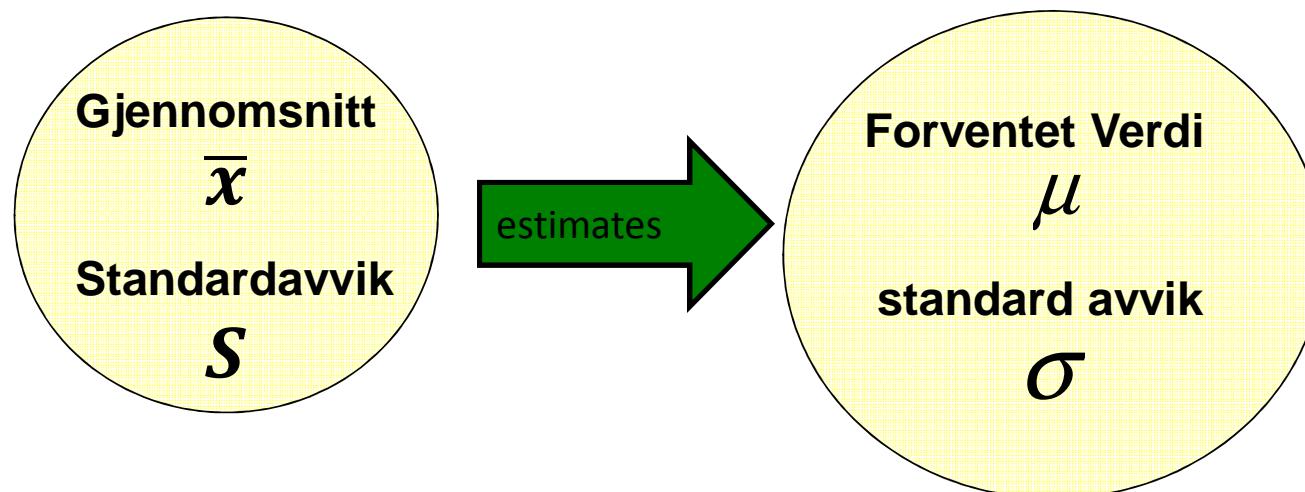
$$S \in N(n\mu, \sigma\sqrt{n}) , i = 1, 2, \dots, n \quad \text{der } S = \sum_{i=1}^n X_i$$

$$\bar{X} \in N(\mu, \frac{\sigma}{\sqrt{n}}) , i = 1, 2, \dots, n \quad \text{der } \bar{X} = \frac{S}{n}$$

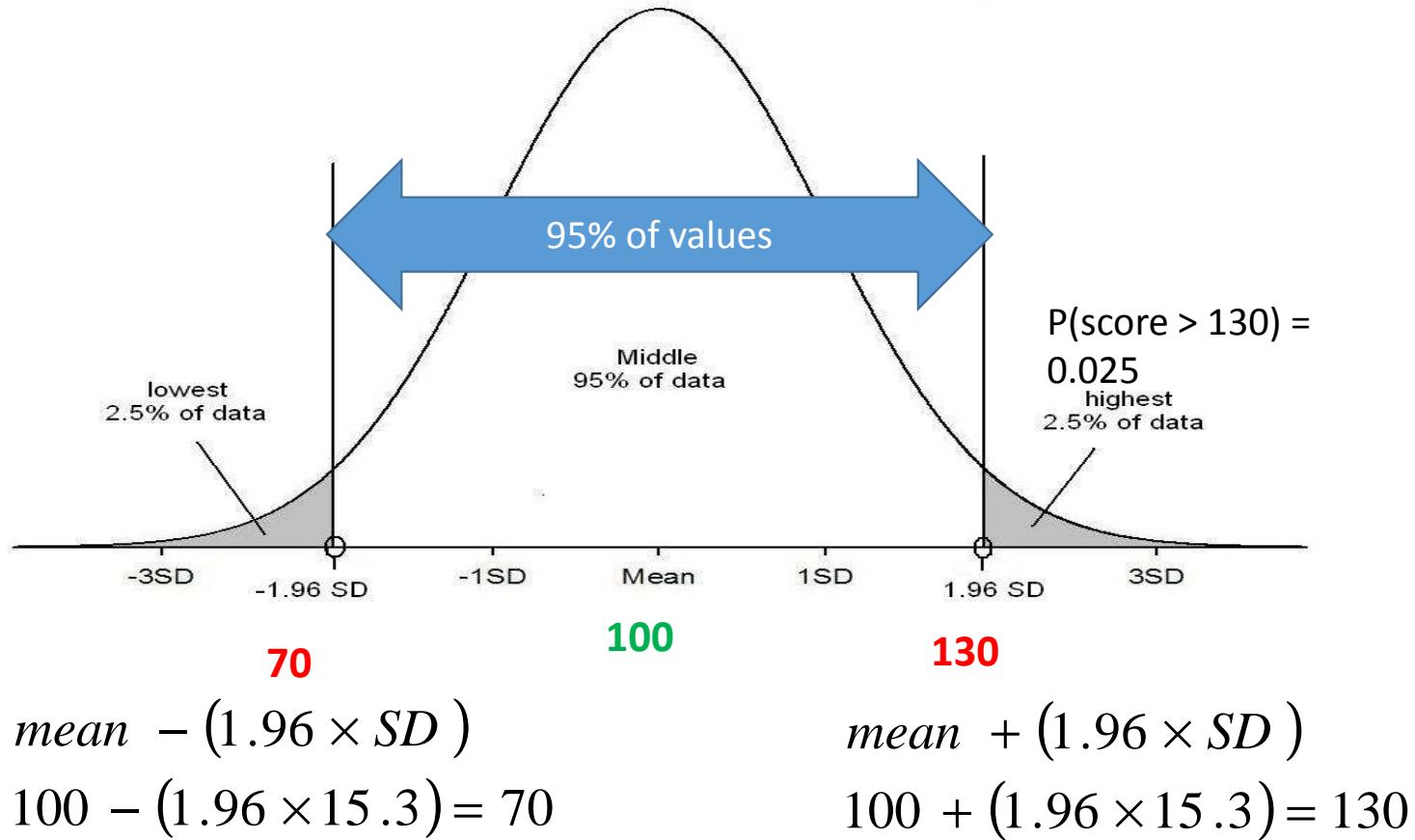
Punkt estimering

Måle data brukes til å estimerer ukjente parameter.

Parameterne presenterer karakteristikkene til populasjonen



95% $1.96 \times SD$'s from the mean



95% of people have an IQ between 70 and 130

Estimering og konfidensintervall

Sentralgrenseteoremet

Dersom X_1, X_2, \dots, X_n er uavhengige identisk fordelte stokastiske variable med forventning μ

og standardavvik σ er $\sum_{i=1}^n X_i$ tilnærmet normalfordelt $N(n\mu, \sigma\sqrt{n})$ når n er stor

og $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ er tilnærmet normalfordelt $N(\mu, \frac{\sigma}{\sqrt{n}})$ når n er stor

Estimering av p i binomisk fordeling.

$$\text{Punktestimator } \hat{p} = \frac{X}{n}, \quad E(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

For store verdier av n har vi : (i følge sentralgrenseteoremet)

Tilnærmet konfidensintervall: $I = \left[\hat{p} - u_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$

Hypotesetesting: Testvariabel: $\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ er tilnærmet

standardnormalfordelt under $H_0: p = p_0$ for store verdier av n ($np_0 > 10$)

Intervallestimering av μ og σ

(formlene er eksakte for normalfordeling, tilnærmet riktige for andre fordelinger når n stor (i følge sentralgrenseteoremet))

$I = \left[\bar{x} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$	Konfidensintervall for μ når σ er kjent
$I = \left[\bar{x} - t_{\frac{\alpha}{2}, (n-1)} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, (n-1)} \frac{s}{\sqrt{n}} \right]$	Konfidensintervall for μ når σ er ukjent

Viktige begrep i hypotesetesting

- Nullhypotese
- Alternativ hypotese
- Testobservator
- Kritisk område og Kritisk verdi
- Signifikansnivå
- p-verdi
- Type I og Type II feil

Hypotesetesting

Hypotesetest = Beslutningsregel

- Du må velge en av to konkurrerende hypoteser.
 - Nullhypotesen H_0 inneholder alltid =
 - Alternativhypotesen H_1 inneholder $<$, $>$ eller \neq
- Du må velge H_0 eller H_1 . Du antar H_0 er sann
- Beslutningsgrunnlaget er stikkprøven.
- Dersom stikkprøven ser "uvanlig" ut, så velger vi H_1
- Dersom stikkprøven ikke er uvanlig, så velger vi H_0

P-verdi og signifikansnivå i hypotesetesting

- Sannsynligheten for at Nullhypotesen er sann.
- p-verdien er jo sannsynligheten for at du skulle funnet et resultat som var like eller mer ekstremt, dersom nullhypotesen var sann.
- Hvis p-verdien er lavere enn signifikansnivået, f.eks. 5 %, så forkaster man nullhypotesen.

Signifikansnivå: Sannsynligheten å forkaste H_0 men det viser seg at den var sann (forkastingsfeil).

- Signifikansnivået angis med α
- Det er sannsynligheten for at testobservatoren havner i forkastningsområdet dersom H_0 er sann

p-verdien

- Antar at H_0 er sann
- Regner ut testobservatoren for testen
- *p*-verdien er sannsynligheten for å få en verdi på testobservatoren som er *minst like uvanlig* som den du fikk
- Nullhypotesen forkastes hvis *p*-verdien er mindre enn signifikansnivået

To metoder for hypotesetesting

Hypotesetesting

- Konklusjonen av testen er enten
 - ① Forkast nullhypotesen, eller
 - ② Ikke forkast nullhypotesen

Den tradisjonelle metoden

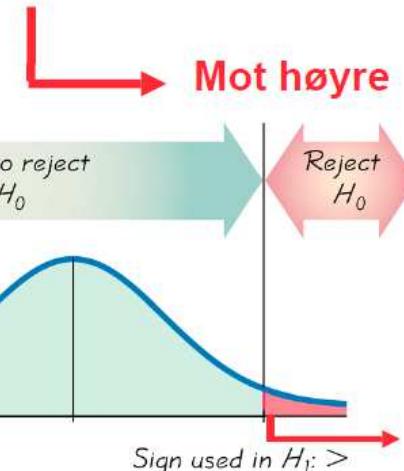
- ① Forkast nullhypotesen dersom testobservatoren er i forkastningsområdet
- ② Ikke forkast nullhypotesen dersom testobservatoren ikke er i forkastningsområdet

p-verdi metoden

- ① *p*-verdien er mindre enn signifikansnivået $\alpha \rightarrow$ Forkast H_0
- ② *p*-verdien er større enn signifikansnivået $\alpha \rightarrow$ Ikke forkast H_0

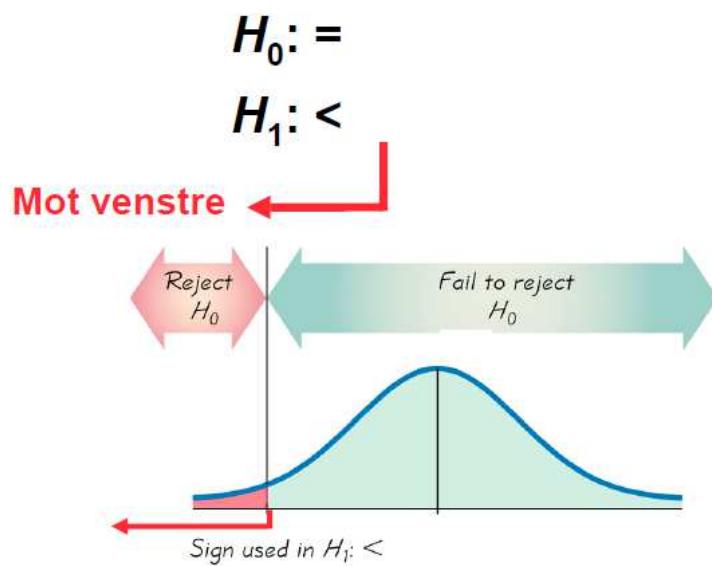
$$H_0: =$$

$$H_1: >$$



Venstresidig test

Venstresidig hypotesetest



To metoder for hypotesetesting

Hypotesetesting

- Konklusjonen av testen er enten
 - 1 Forkast nullhypotesen, eller
 - 2 Ikke forkast nullhypotesen

Den tradisjonelle metoden

- 1 Forkast nullhypotesen dersom testobservatoren er i forkastningsområdet
- 2 Ikke forkast nullhypotesen dersom testobservatoren ikke er i forkastningsområdet

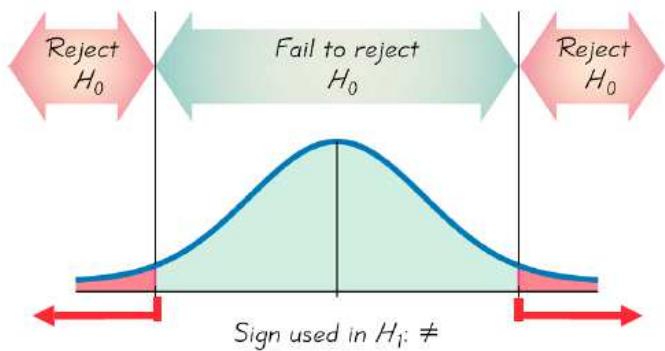
p-verdi metoden

- 1 p-verdien er mindre enn signifikansnivået $\alpha \rightarrow$ Forkast H_0
- 2 p-verdien er større enn signifikansnivået $\alpha \rightarrow$ Ikke forkast H_0

Tosidig hypotesetest

$H_0:$ = α likt fordelt i de to halene
 $H_1:$ \neq til forkastningsområdet

Betyr mindre enn eller større enn



Hypotesetesting - Hypotesetesting for μ når σ er kjent

Kvantilen u_α bestemmes altså fullt og helt av α . Kvantiltabell:

α	0,001	0,005	0,01	0,025	0,05	0,10
u_α	3,090	2,576	2,326	1,960	1,645	1,282

Hvis svaret på om begivenheten finner sted blir JA, så har vi en positiv test og vi har statistisk belegg for å si at H_1 er riktig, altså H_0 forkastes.

I motsatt fall, altså at svaret blir NEI, så har vi en negativ test og kan ikke påstå at H_1 er riktig. Dermed beholder vi nullhypotesen H_0 .

Alternativ 2: Identifiser en testobservator for testen: $u = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ (de σ er kjent)

σ kjent	u-test	Forkastningsregel	Test variabel (testobservator)
Høyresidig test 1) $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ signifikansnivå α	Forkast H_0 hvis $\bar{x} > K$ der $K = \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > u_\alpha$	
Venstresidig test 2) $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ signifikansnivå α	Forkast H_0 hvis $\bar{x} < K$ der $K = \mu_0 - u_\alpha \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -u_\alpha$	
Tosidig test 3) $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ signifikansnivå α	Forkast H_0 hvis $\bar{x} < K_1$ eller $\bar{x} > K_2$ der $K_1 = \mu_0 - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ og $K_2 = \mu_0 + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $ u = \left \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right > u_{\alpha/2}$	

6.2 Hypotesetesting for μ når σ er ukjent

Fremgangsmåten blir da den samme med 2 forskjeller:

1. Vi erstatter σ med S (estimert standardavvik)
2. Isteden for $t_{\alpha/2, (n-1)}$ benyttes da kvantil for t -fordelingen med $n-1$ frihetsgrader, $t_{\alpha/2, (n-1)}$.
altså kvantilen avhenger ikke bare av α , men også av n .

Siden σ er ukjent må vi bruke Student t-fordeling med T som test variabel (testobservator) istedelenfor Gaussfordelingen med U som variabel.

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

For å finne T intervallet trenger vi α -kvantilet til t-fordelingen, $t_{\alpha/2, (n-1)}$.

Verdien av $t_{\alpha/2, (n-1)}$ avhenger av antall frihetsgrader ($n-1$).

$$1 - \alpha = P(-t_{\alpha/2, (n-1)} < T < t_{\alpha/2, (n-1)})$$

σ ukjent	t-test	Forkastningsregel
Høyresidig test		
1) $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ signifikansnivå α	$\bar{x} > K$ der $K = \mu_0 + t_{\alpha, (n-1)} \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} > t_{\alpha, n-1}$
Vennstresidig test		
2) $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ signifikansnivå α	$\bar{x} < K$ der $K = \mu_0 - t_{\alpha, (n-1)} \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} < -t_{\alpha, n-1}$
Tosidig test		
3) $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ signifikansnivå α	$\bar{x} < K_1$ eller $\bar{x} > K_2$ $der K_1 = \mu_0 - t_{\alpha/2, (n-1)} \frac{\sigma}{\sqrt{n}}$ $og K_2 = \mu_0 + t_{\alpha/2, (n-1)} \frac{\sigma}{\sqrt{n}}$	Forkast H_0 hvis $ t = \left \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \right > t_{\alpha/2, n-1}$

Formal Statement of the Model

- General regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

1. β_0 , and β_1 are parameters
2. X is a known constant
3. Deviations ε are independent $N(0, \sigma^2)$

Stokastisk forsøk, utfall, utfallsrom

Regresjonsanalyse

- Vi bruker tilgjengelige data for å estimere verdier på konstantleddet og stigningsforholdet

$$\hat{Y} = b_0 + b_1 X \text{ hvor } \hat{Y} = \text{anslått (predikert)} \\ \text{verdi på Y}$$

- Forskjellen mellom faktisk og predikert verdi på X er feilreddet

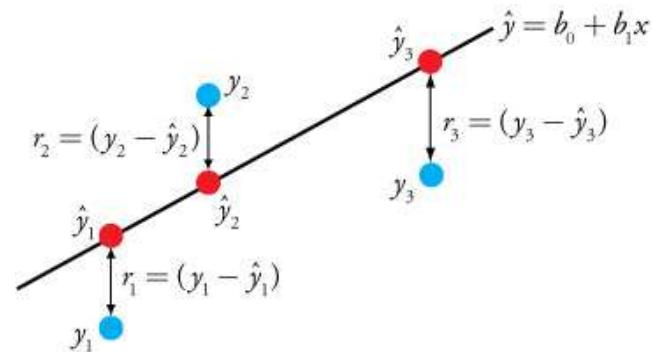
$$e = Y - \hat{Y}$$

Meaning of Regression Coefficients

- The values of the regression parameters β_0 , and β_1 are not known. We estimate them from data.
- β_1 indicates the change in the mean response per unit increase in X.

Residual = Observed value – predicted value

$$e = y - \hat{y}$$



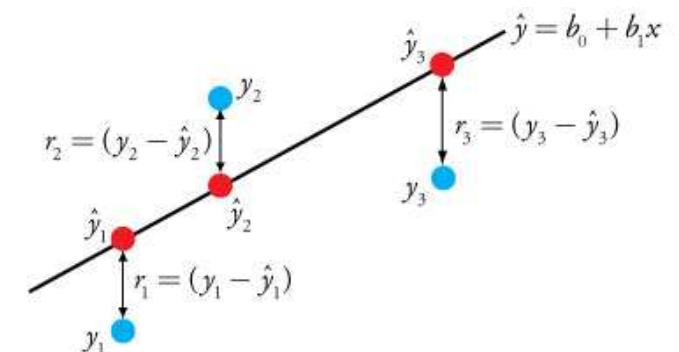
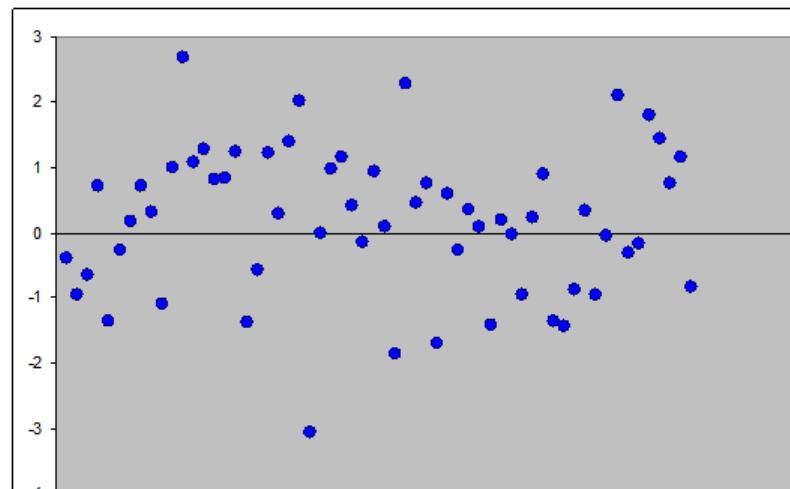
Forutsetningene for regresjon

Residual = Observed value – predicted value

$$e = y - \hat{y}$$

1. Residualene har et gjennomsnitt på 0 i populasjonen.
2. Residualene har lik varians med alle X

- "Residualene" bør se noenlunde slik ut:



Lineær regresjon

Er det lineær sammenheng mellom uavhengig variable(forklaringsvariabel) x og responsvariabel y?

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT.
x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM

$$y = a + bx$$



Formler

Regresjon

$$\hat{y} = a + b x, \quad \text{der } b = r \frac{s_y}{s_x}, \quad a = \bar{y} - b \bar{x}$$

$$s^2 = \frac{1}{n-2} \sum e_i^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2.$$

Standardfeil

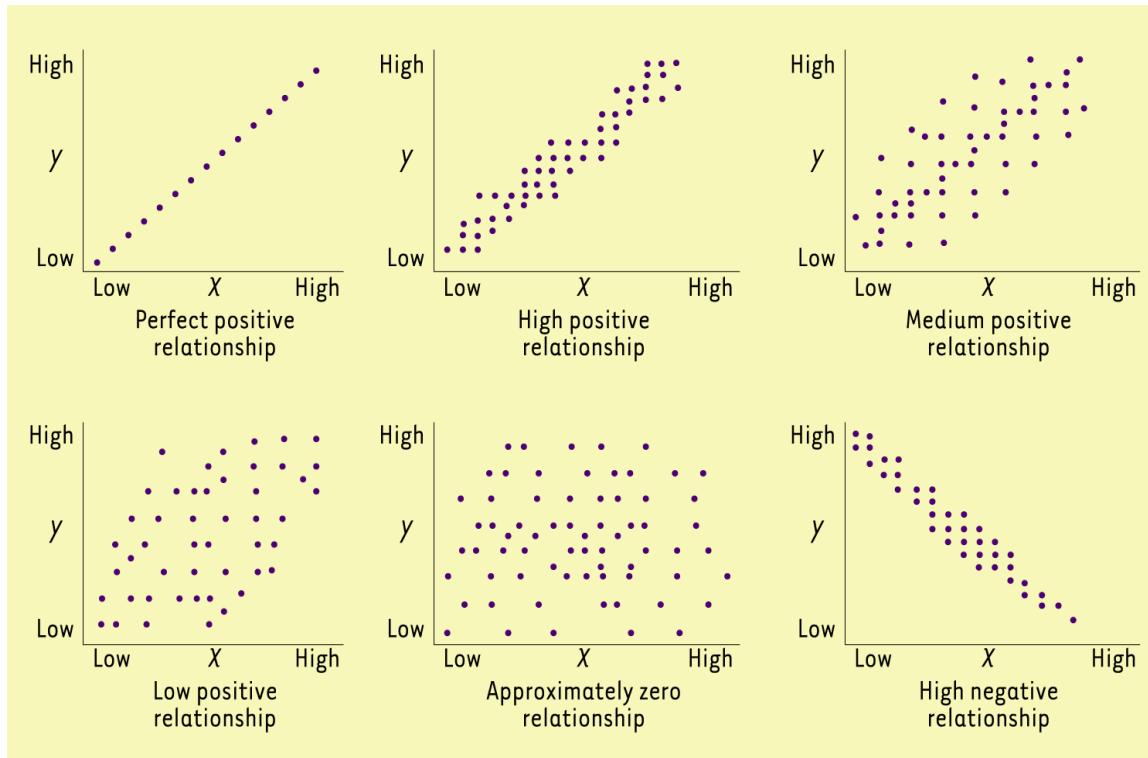
$$\text{SE}_a = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$\text{SE}_b = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$\text{SE}_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

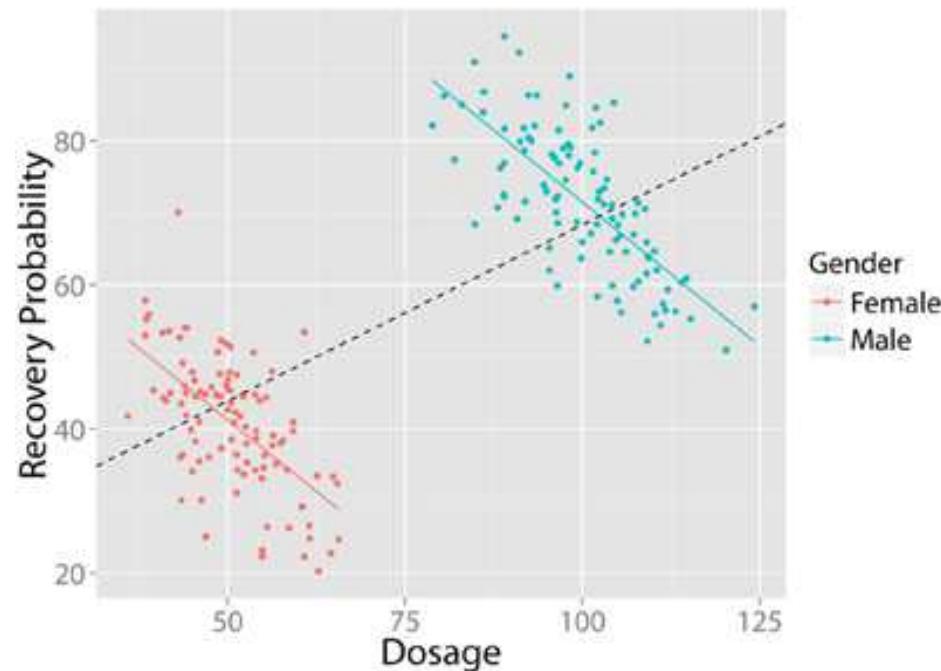
$$\text{SE}_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Korrelasjon



Simpsons paradox

Example of Simpson's Paradox. Despite the fact that there exists a negative relationship between dosage and recovery in both males and females, when grouped together, there exists a positive relationship



Simpsons paradoks

Tallene lyver ikke, men de kan av og til skjule sannheten. Ikke sjeldent finner vi bare små eller ingen endringer eller forskjeller når vi studerer utviklingen av et fenomen over tid eller sammenlikner ulike grupper. Men dette er ofte bare tilsynelatende. Om vi går «bak» totaltallene eller gjennomsnittet, finner vi ofte store endringer eller forskjeller. Helheten varierer mindre enn delene.

Eksempel [rediger | rediger kilde]

Tenk at det utføres en undersøkelse på hva ingeniører og lærer tjener fordelt mellom kjønn. Utvalget består av 20 mannlige og 5 kvinnelige ingeniører, og 15 mannlige og 10 kvinnelige lærere. Anta videre at de fire gruppene har følgende gjennomsnittlig lønn:

	Menn	Kvinner
Ingeniør	800 000,-	860 000,-
Lærer	400 000,-	430 000,-
Gjennomsnitt for kjønn	628 571,-	573 333,-

Fra tabellen over ser vi at kvinnelige ingeniører tjener gjennomsnittlig bedre enn mannlige ingeniører, og at kvinnelige lærer tjener gjennomsnittlig bedre enn mannlige lærere, paradokset er at menn tjener gjennomsnittlig bedre enn kvinner. Årsaken er at det er flere menn enn kvinner som er ingeniører.

For å komme ut av dette paradokset må analysen standardiseres. I dette tilfellet vil det være å analysere gjennomsnittslønnen mellom kjønnene hvor forholdet mellom antall lærer og ingeniører holdes likt for begge kjønn. [\[1\]](#)

Residuals:

	Min	1Q	Median	3Q	Max
	-2.25074	-0.57409	0.04147	0.52315	2.31959

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept) β_0	0.16127	0.09218	1.75	0.0833 .
x β_1	0.47169	SE $\hat{\beta}_1$	14.66	<2e-16 ***

Residual standard error: 0.917 on 98 degrees of freedom

Multiple R-squared: 0.6867, Adjusted R-squared:
0.6835

F-statistic: 214.8 on 1 and 98 DF, p-value: < 2.2e-16

- c. Bruk utskriften til å lage et 95% konfidensintervall for β_1 .
- d. Lag et 95% prediksjonsintervall for en ny observasjon gjort med $x^* = 3.2$. Du får oppgitt at $\bar{x} = -0.291$ og $\sum(x_i - \bar{x})^2 = 812$.

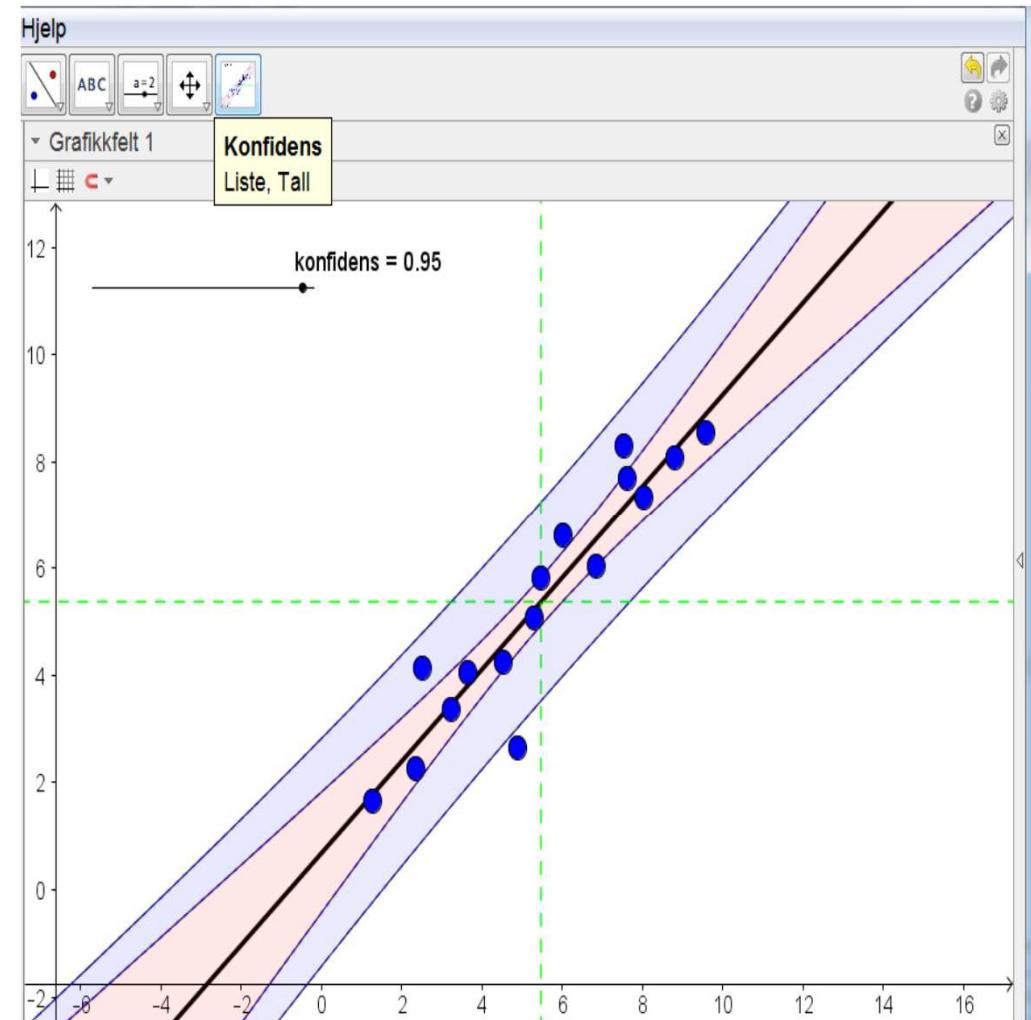
Stokastisk forsøk, utfall, utfallsrom

- c. Konfidensintervallet er $\hat{\beta}_1 \pm t_{0.025,98} SE_{\hat{\beta}_1} = 0.47169 \pm 1.984 * 0.03218 = [0.408, 0.536]$. Vi finner $SE_{\hat{\beta}_1} = 0.03218$ i R-utskriften og bruker $t_{0.025,100} = 1.984$ som er det nærmeste vi finner df=98 i Tabell D.
- d. Først må vi finne $\hat{y} = 0.16127 + 0.47169 * 3.2 = 1.67$. Et 95% prediksjonsintervall blir $\hat{y} \pm t_{100,0.025} SE_{\hat{y}} = 1.67 \pm 1.984 * 0.93 = [-0.175, 3.515]$. Her bruker vi $t_{0.025,100} = 1.984$ som er det nærmeste vi finner df=98 i Tabell D. I utregning av SE

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 0.917 \sqrt{1 + \frac{1}{100} + \frac{(3.2 - (-0.291))^2}{812}} = 0.93$$

Har vi brukt $s = 0.917$ fra R-utskriften.

konfidensintervall for β_1 og
prediksjonsintervall for y



3. Et firma ønsker å teste en ny blodtrykksreduserende medisin mot en eksisterende medisin. Følgende tabell viser reduksjonene 10 ulike pasienter oppnådde med bruk av legemidlene. (5 pasienter prøvde «Eksisterende» og 5 andre pasienter prøvde «Nytt»). Vi ønsker å teste om den nye medisinen (i forventning) gir større reduksjon i blodtrykk enn den gamle.

Eksisterende x_1	Nytt x_2
3.04	5.59
3.48	4.52
4.09	7.88
3.68	7.5
5.45	3.20

- a. Finn gjennomsnitt og standardavvik for målingene x_1 og x_2 .
- b. Formuler nullhypotese og alternativ hypotese. Beregn t-observatoren for to-utvalgs t-testen med ulik varians i de to gruppene.
- c. Hvilken konklusjon gir testen på 5% signifikansnivå? Hvilke antagelser bygger testen på?

I slike forsøk vil man ofte la samme pasient prøve begge medisiner. Anta nå at dette ble gjort i tabellen over, slik at de to tallene i hver rekke svarer til samme pasient (og at det dermed kun er 5 ulike pasienter involvert totalt). Følgende R utskrift viser resultatet fra en paret t-test for dataene i tabellen over.

```
One Sample t-test

data: x2 - x1
t = 1.5827, df = 4, p-value = 0.09433
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
-0.6210895      Inf
sample estimates:
mean of x
1.79
```

- d. Forklar hvordan du beregner t-observatoren for den parede t-testen. Hva blir konklusjonen på testen når du bruker 5% signifikansnivå?
- e. Hva betyr det at testen har 5% signifikansnivå?

To grupper

3.

- a. Gjennomsnitt for x_1 og x_2 er henholdsvis 3.95 og 5.74. Standardavvik for x_1 og x_2 er henholdsvis 0.92 og 1.98.
- b. Nullhypotese og alternativ hypotese: $H_0: \mu_1 - \mu_2 = 0$ og $H_a: \mu_1 - \mu_2 < 0$, t-observatoren for to-utvalgs t-testen med ulik varians i de to gruppene:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3.95 - 5.74}{\sqrt{\frac{0.92^2}{5} + \frac{1.98^2}{5}}} = -1.835$$

- c. To-utvalgs t-testen har $\min(5-1, 5-1) = 4$ frihetsgrader. Kritisk verdi for testen fra tabell D er -2.132, dvs. vi beholder H_0 . Med ord: vi kan ikke med utgangspunkt i data (og denne testen) påstå at ny medisin har bedre virking enn den eksisterende. Antagelser som testen bygger på: observasjonene er uavhengige og normalfordelte, men μ og σ tillates forskjellige i de to gruppene.
- d. Beregning av t-observatoren for den parede t-testen: Det man gjør i praksis er å beregne $y = x_2 - x_1$ for hver av de 5 linjene i tabellen. Basert på disse 5 observasjonene gjør man en vanlig ett-utvalgs t-test (det er dette R-utskriften viser). Fordi $p\text{-value} = 0.09433 < 0.05$ beholder vi H_0 . Konklusjonen blir den same som i b).
- e. At testen har 5% signifikansnivå betyr at når H_0 er sann ($\mu_1 - \mu_2 = 0$) forkaster testen H_0 kun i 5% av tilfellene (under gjentatt sampling fra populasjon av pasienter).
- f. En p-verdi på 0.09433 betyr at sannsynligheten for å observere en t som er større enn $t = 1.5827$, gitt at H_0 er sann, er 0.09433. «Sannsynlighet» skal her beregnes under gjentatt sampling av pasienter fra populasjonen.

$$\text{To-utvalgs } t\text{-observator: } t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\min(n_1 - 1, n_2 - 1))$$

$$\text{Pooled to-utvalgs } t\text{-observator: } t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

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

Eksisterende	Nytt
x_1	x_2
3.04	5.59
3.48	4.52
4.09	7.88
3.68	7.5
5.45	3.20

One Sample t-test

```
data: x2 - x1
t = 1.5827, df = 4, p-value = 0.09433
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
-0.6210895 Inf
sample estimates:
mean of x
1.79
```

Eksamens H2014 – 1d)

n	k	.10	.15	.20	.25	.30	.35	.40	.45	.50
10	0	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7		.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8			.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9					.0001	.0005	.0016	.0042	.0098
	10							.0001	.0003	.0010