Ny og bedre versjon

MAT110

Statistikk 1

Kompendium 2019, del 2

Per Kristian Rekdal

Bård-Inge Pettersen

Høgskolen i Molde
Vitenskapelig høgskole i logistikk

Innhold

Kapittel 4

Kontinuerlige stokastiske fordelinger og CLT

4.1 Kontinuerlig fordeling

Motiverende eksempel: (diskret $\longrightarrow$ kontinuerlig fordeling)

Hustadmarmor AS er en stor produksjonsfabrikk i Elnesvågen i Fræna kommune. Bedriften produserer og leverer "*slurry*". Slurry er en flytende hvit væske produsert fra kalkstein og marmor som brukes i fyllmasseindustrien over hele verden, blant annet i papir og tannkrem, se figur 4.1.

Anta at Hustadmarmor bruker rundt 1 000 MWh med strøm per dag, men at dette varierer noe dra dag til dag.

Figur 4.1: Hustadmarmor. Slurry. Strøm.

Eksperiment:

Hustadmarmor måler det daglige forbruket av strøm.

i) Inndeling 1:

Hustadmarmor måler det daglige forbruket av strøm aller dager i ett år, dvs. $n = 365$ dager.
Deler målingene inn i 4 intervall:
$600 - 800$ MWh, $800 - 1000$ MWh, $1000 - 1200$ MWh, $1200 - 1400$ MWh.

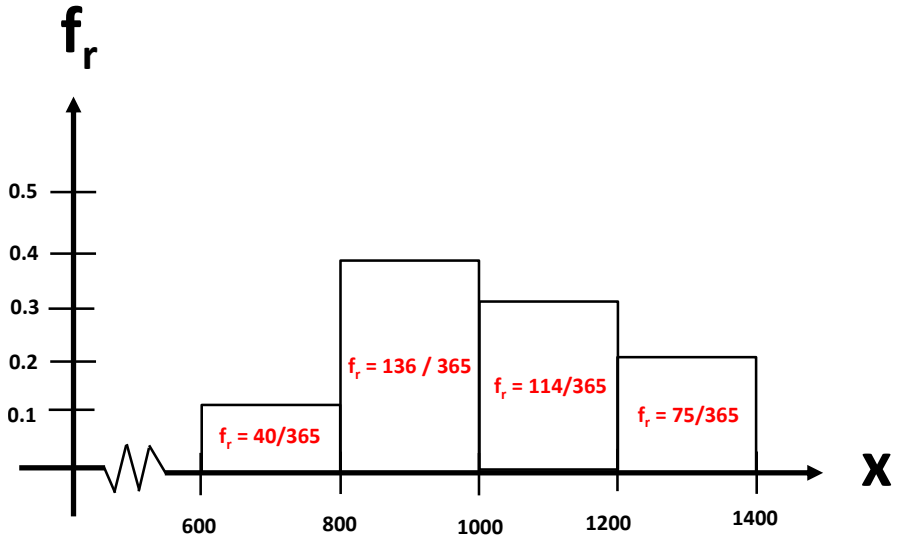
Relativ frekvensen for disse intervallene er:

$$f_r(n_1) = \frac{n_1}{n} = \frac{40}{365} = 0.1096 \tag{4.1}$$

$$f_r(n_2) = \frac{n_2}{n} = \frac{136}{365} = 0.3726 \tag{4.2}$$

$$f_r(n_3) = \frac{n_3}{n} = \frac{114}{365} = 0.3123 \tag{4.3}$$

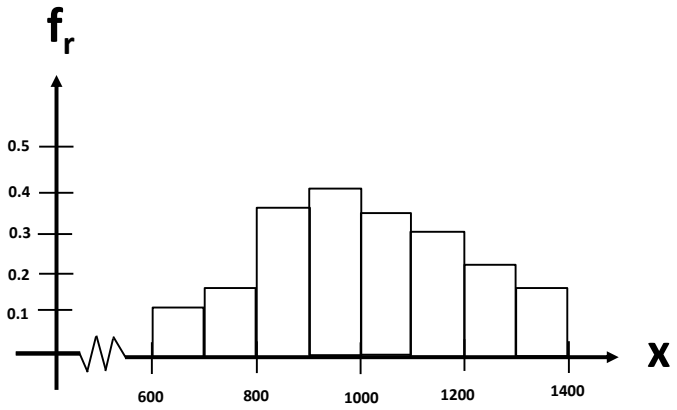
$$f_r(n_4) = \frac{n_4}{n} = \frac{75}{365} = 0.2055 \tag{4.4}$$



Figur 4.2: Relativ frekvens. Strømforbruk.

ii) Inndeling 2:

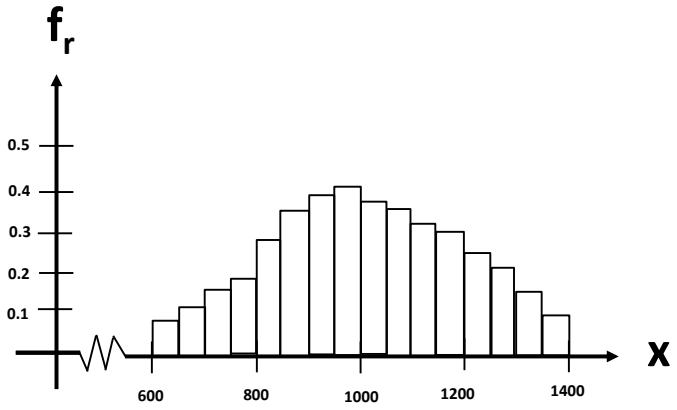
Måler et daglige forbruket av strøm aller dager i to år, dvs. $n = 2 \cdot 365 = 730$ dager.
Deler målingene inn i 8 intervall.



Figur 4.3: Relativ frekvens. Strømforbruk.

ii) Inndeling 3:

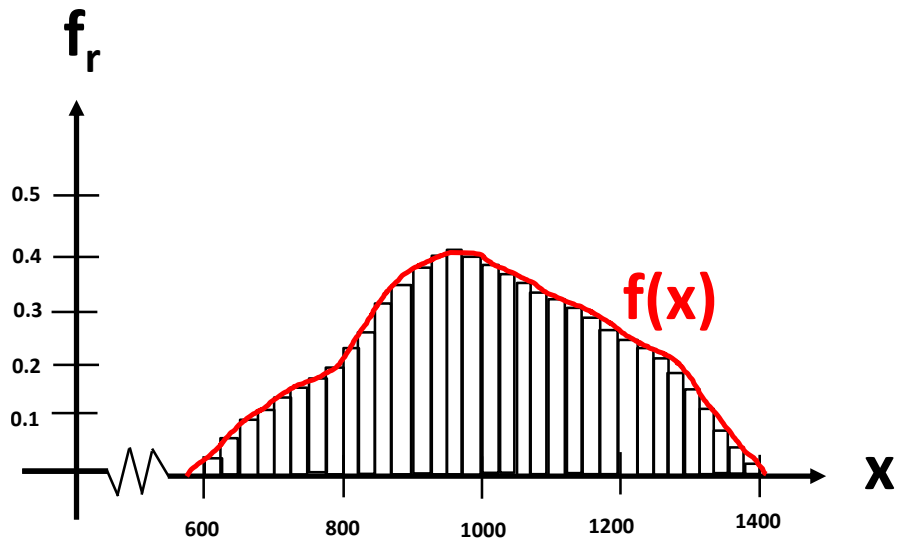
Måler et daglige forbruket av strøm aller dager i to år, dvs. $n = 4 \cdot 365 = 1460$ dager.
Deler målingene inn i 16 intervall.



Figur 4.4: Relativ frekvens. Strømforbruk.

iv) Inndeling 4:

Deler målingene inn i 32 intervall.



Figur 4.5: Relativ frekvens og tetthetsfunksjon. Strømforbruk.

Når vi deler den relative frekvensen inn i finere og finere "maskevidder" så danner det en "glatt" kruve, en kontinuerlig kurve.

Denne kontinuerlige kurven kalles *tetthetsfunksjonen*:

$$f(x) \;=\; \text{tettehetsfunksjon} \tag{4.5}$$

På samme måte som at $f_r(n_i)$ er rett normalisert

$$\underline{\sum_{i=1}^{4} f_r(n_i)} = f_r(n_1) + f_r(n_2) + f_r(n_3) + f_r(n_4) = \frac{n_1}{n} + \frac{n_2}{n} + \frac{n_3}{n} + \frac{n_4}{n} \tag{4.6}$$

$$= \frac{40}{365} + \frac{136}{365} + \frac{114}{365} + \frac{75}{365} = \frac{365}{365} = \underline{1} \tag{4.7}$$

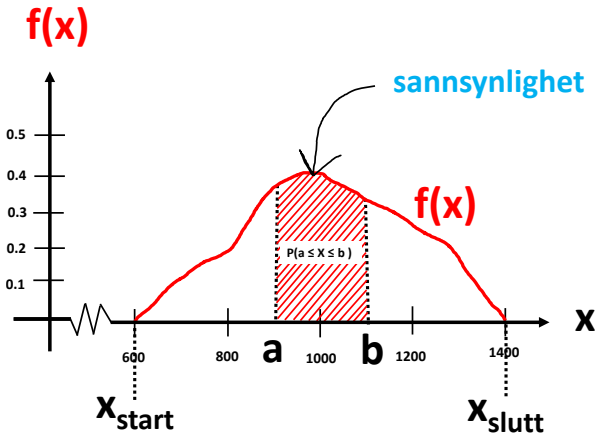
så er også $f(x)$ rett normalisert

$$\int_{x_{\text{start}}}^{x_{\text{slutt}}} f(x)\,dx = 1 \tag{4.8}$$

Sannsynligheten regnes ut via et integral: [1]

$$\boxed{P\big(a \le X \le b\big) = \int_a^b f(x)\,dx} \tag{4.10}$$

hvor $P\big(a \le X \le b\big)$ = sannsynligheten for at strømforbruket ligger mellom a og b.



Figur 4.6: Relativ frekvens og tetthetsfunksjon. Strømforbruk.

[1] Analogt, for en diskret sannsynlighetfordeling er:

$$P\big(a \le X \le b\big) = \sum_{i=a}^{b} P(X = x_i) \tag{4.9}$$

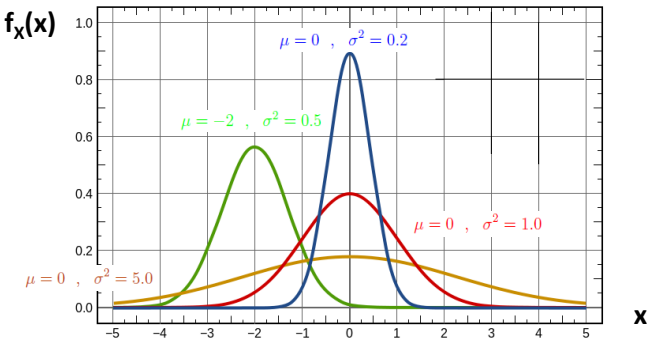
4.2 Normalfordelingen (kontinuerlig)

Det viser seg at svært mange fenomen og målinger har sannsynlighetsfordelinger som kan beskrives av en spesiell, glatt og fin funksjon.

Med god tilnærmelse kan denne glatte, fine og spesielle fordelingen beskrive sannsynlighetfordelingen til for eksempel blodtrykk, reisetid, målingsfeiler og høyde på personer osv.

Denne spesielle fordelingen kalles *normalfordelingen*. [2]
Eksempel på denne glatte og fine fordelingen ser du i figur 4.7.
Normalfordelingen har en helt sentral rolle i statistikken, blant annet fordi:

- beskriver fordelingen til svært mange fenomen med god tilnæremelse
- har gode matematiske egenskaper som gjør den den "enkel" å regne på
- det viser seg at dersom man har mange like fordelinger, ikke nødvendigvis normalfordelinger, så er gjennomsnittet av dem normalfordelt [3]



Figur 4.7: Ulike **normal**fordelinger.

[2] Fordelingen kalles også en *Gauss* -fordeling.
[3] Dette kalles *sentralgrensesetningen* (CLT) og er noe vi skal se på senere. Denne setningen er helt grunnleggende i statistikk.

Eksempel: ($\overbrace{\textbf{binominal}}^{\text{diskret}}$ fordeling $\approx$ $\overbrace{\textbf{normal}}^{\text{kont.}}$fordeling)

I dette eksemplet skal vi vise, under visse betingelser, at en binomialfordeling med god tilnærmelse kan beskrives av en normalfordeling.

Anta at X er binomialfordelt:

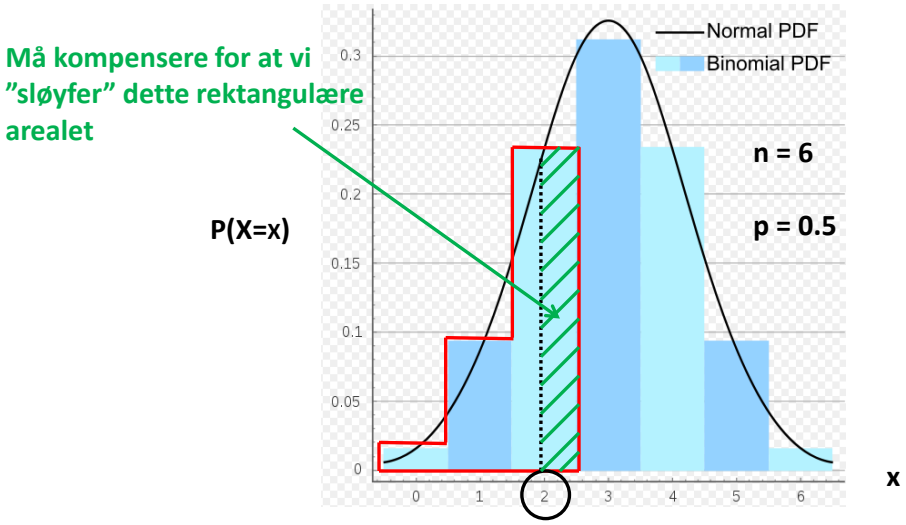
$$X \sim \text{Bin}\big[\, n = 6\,,\, p = 0.5 \,\big] \tag{4.11}$$

a) Finn den eksakte verdien til $P(X \leq 2)$.

b) Dersom man regner ut $P(X \leq 2)$ via en tilpasset normalfordeling med $E[X] = np$ og $Var[X] = np(1-p)$, så får man:

$$P(X \leq 2) \approx 0.3409 \quad \text{(tilnærmet)} \tag{4.12}$$

Sammenlign svarene i oppgave **a** og **b** og kommenter svaret.



Figur 4.8: **Binomisk** fordeling (diskret) og **normal**fordelingen (kontinuerlig).

a) Eksakt svar: ($\overbrace{\textbf{binominal}}^{\text{diskret}}$ fordeling)

$$\underline{\underline{P(X \leq 2)}} \quad = \quad \text{summen av arealene til de 3 søylene markert med rødt i figur (4.8)}$$

$$= \quad P(X=0) \; + \; P(X=1) \; + \; P(X=2) \tag{4.13}$$

$$= \quad \binom{n}{0} p^0 (1-p)^{n-0} \; + \; \binom{n}{1} p^1 (1-p)^{n-1} \; + \; \binom{n}{2} p^2 (1-p)^{n-2} \tag{4.14}$$

$$= \quad 0.0156250 \; + \; 0.093750 \; + \; 0.234375 \quad = \quad \underline{\underline{0.34375}} \quad \text{(eksakt)} \tag{4.15}$$

b) Sammenligner:

$$P(X \leq 2) \quad = \quad 0.34375 \qquad \text{(eksakt)} \tag{4.16}$$
$$P(X \leq 2) \quad \approx \quad 0.3409 \qquad \text{(tilnærmet)} \tag{4.17}$$

Dette er en god tilnærmelse.
Det viser seg at under visse betingelser, nemlig:

$$n \cdot p\,(1-p) \quad \gtrsim \quad 5 \tag{4.18}$$

så kan en binomialfordeling med tilnærmelse beskrives av en normalfordeling.

■

Definisjon: ($\overbrace{\text{normal}}^{\text{kont.}}$fordeling [4], $X \sim \overbrace{N[\mu, \sigma]}^{\text{2 param.}}$)

Tetthetsfunksjonen til en normalfordeling er: [5]

$$f_X(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \tag{4.19}$$

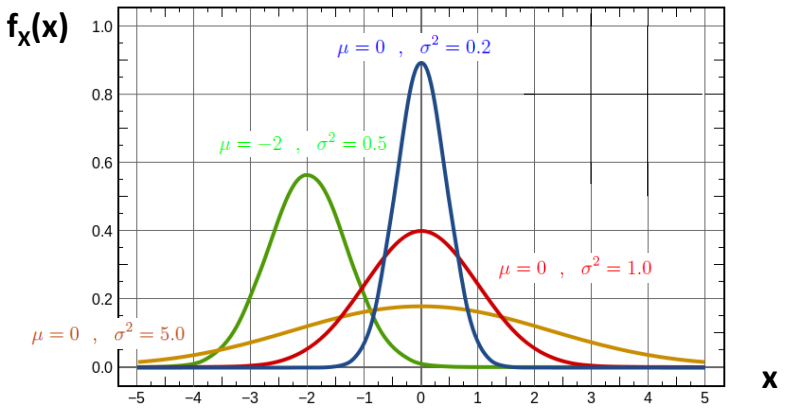
hvor [6]

$$\mu = E[X] = \text{forventingen av } X \quad (\text{ lokalisering av toppunkt }) \tag{4.20}$$

$$\sigma^2 = Var[X] = \text{variansen av } X \tag{4.21}$$

$$\sigma = \sqrt{\sigma^2} = \text{standardavviket av } X \quad (\text{ halvbredden av kurven }) \tag{4.22}$$

■



Figur 4.9: **Normal**fordelinger.

[4] Normalfordelingen kalles også en *Gauss*-fordeling.
[5] Lign.(4.19) er en tetthetsfunksjon.
[6] μ leses "my". σ leses "sigma".

Noen egenskaper til normalfordelingen:

1. $f_X(x)$ er **symmetrisk** om forventingsverdien $E[X] = \mu$

2. Arealet under kurven er alltid lik 1. Vi sier at normalfordelingen er **normert** til 1: [7]

$$\int_{-\infty}^{\infty} f_X(x)\, dx \;=\; 1 \qquad (4.23)$$

3. Forventingsverdien $E[X] = \mu =$ lokalisering til toppunktet, med andre ord tyngdepunktet.

4. Standardavviket $\sigma = \sqrt{Var[X]}$ sier noe om "bredden", dvs. spredningen, til normalfordelingen, se figur (4.9).

Sannsynligheten finner man ved å integrere:

$$P\big(a \leq X \leq b\big) \;=\; \int_a^b f(x)dx \qquad (4.24)$$

PS:
Dette er et statistikkurs, ikke et matematikkurs. Derfor skal vi ikke å regne ut integralet i lign.(4.24) her. Vi skal bare gjøre et $\underbrace{\text{tabelloppslag}}_{\text{NB!}}$ hvor integralene er regnet ut.

[7] Dette er helt analogt med betingelsen i lign.(1.29) $\sum_{i=1}^{n} p_i = 1$.

4.2.1 Standardisering

For å slippe å integrere så skal vi nå først gjøre en omskalering som heter *standardisering*.
Etter at normalfordelingen har blitt standardisert så ligger alt klart til å gjøre tabelloppslag.

Spesieltilfellet med $E[X] = 0$ og $Var[X] = 1$ gjør at den generelle normalfordelingen i lign.(4.19) reduserer seg til det som kalles den **standardiserte** normalfordelingen. Denne standard normalfordelingen spiller en så sentral rolle at vi formulerer den i en egen setning:

Definisjon: (standardisert $\overbrace{\text{normal}}^{\text{kont.}}$fordeling, $X \sim N[\mu = 0, \sigma = 1]$)

La X være en *kontinuerlig* stokastisk variabel. Dersom

$$E[X] = \mu = 0 \quad , \quad Var[X] = \sigma^2 = 1 \tag{4.25}$$

så vil den generelle normalfordelingen i lign.(4.19) redusere seg til

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \tag{4.26}$$

Dette kalles den **standardiserte** normalfordelingen.

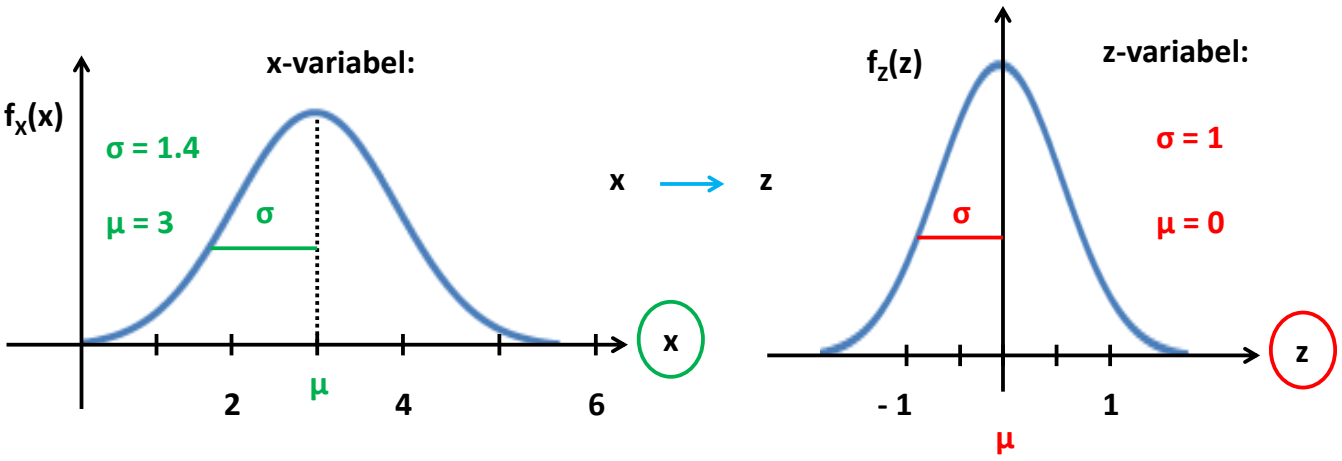
■

4.2.2 Standardisering = omskalering

En hvilken som helst generell normalfordeling $f_X(x)$ (med $\mu \neq 0$ og $\sigma \neq 1$) kan skaleres om til en **standard** normalfordeling $f_Z(z)$ (med $\mu = 0$ og $\sigma = 1$). Dette innser vi ved å betrakte følgende sammenheng mellom Z og X:

$$Z = \frac{X - \mu}{\sigma} \tag{4.27}$$

Et illustrativt eksempel på effekten av denne OMSKALERINGEN er:



Figur 4.10: Tetthetsfunksjonen $f_X(x)$ med X-variabelen og $f_Z(z)$ med Z-variabelen.

Som figuren illustrerer, den nye OMSKALERTE variabelen har følgende effekt:

- fordelingen (dvs. grafen) flyttes slik at den blir symmetrisk om y-aksen, dvs. forventingen endres fra 3 til $\mu = 0$

- fordelingen (dvs. grafen) blitt smalere og høyere (!) siden standardavviket endres fra 1.4 til $\sigma = 1$

- det totale arealet under grafen er forsatt normert til 1, også etter omskaleringen [8]

[8] Dette er kanskje ikke så lett å se med det blotte øyne, men det kan vises matematisk.

Matematisk så betyr det faktum at $\mu = 0$ og $\sigma = 1$ i den nye omskalerte variabelen, følgende:

$$\boxed{E[Z] \;=\; 0 \qquad , \qquad Var[Z] \;=\; 1} \tag{4.28}$$

Bevis:

Forventning og varians for Z: (bruker regnereglene i lign.(2.64)-(2.71))

$$\underline{\underline{E[Z]}} \;=\; E\left[\frac{X-\mu}{\sigma}\right] \;=\; \frac{1}{\sigma}\,E[X-\mu] \;=\; \underbrace{E[X]}_{=\mu} - \mu \;=\; \mu - \mu \;=\; \underline{\underline{0}} \tag{4.29}$$

$$\underline{\underline{Var[Z]}} \;=\; Var\left[\frac{X-\mu}{\sigma}\right] \;=\; \frac{1}{\sigma^2}\,Var[X-\mu] \tag{4.30}$$

$$=\; \frac{1}{\sigma^2}\left(\underbrace{Var[X]}_{=\sigma^2} - \underbrace{Var[\mu]}_{=0}\right) \;=\; \frac{1}{\sigma^2}\left(\sigma^2 - 0\right) \;=\; \frac{\sigma^2}{\sigma^2} \;=\; \underline{\underline{1}} \tag{4.31}$$

■

4.2.3 Sammenhengen mellom $P(Z \leq z)$ og $G(z)$

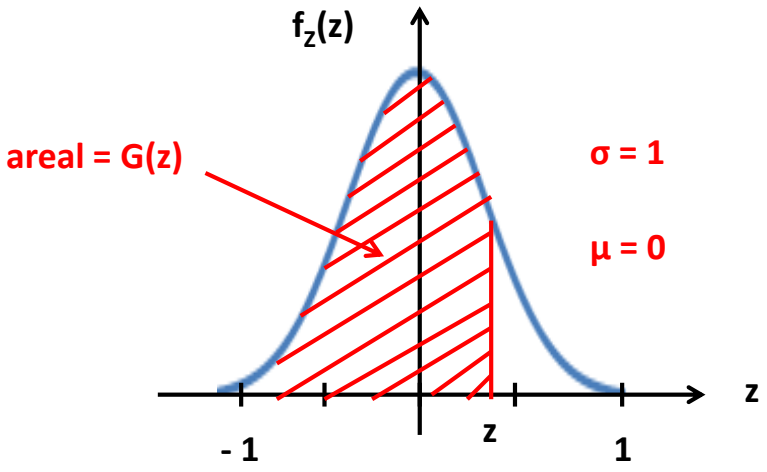
La oss se på en bestemt verdi $Z = z$ som illustrert i figur (4.11). Arealet til venstre for denne verdien z representerer en sannsynlighet:

arealet til venstre for z	$=$	sannsynligheten for at den stokastiske variabelen Z har verdier mindre eller lik z	
	$=$	$P(Z \leq z)$	(4.32)

Arealet til venstre for z er det samme som integralet under grafen $f_Z(z)$: [9]

$$\underline{\underline{P(Z \leq z)}} = \int_{\text{start}}^{\text{stopp}} = \overbrace{\int_{-\infty}^{z} f_Z(s)\, ds}^{=G(z)} \overset{\text{def.}}{=} \underline{\underline{\overbrace{G(z)}^{\text{tabelloppslag}}}} \tag{4.33}$$

Gauss-integralet $G(z)$ behøver ikke regnes ut av oss. Det er tabelloppslag. Tabell finnes på side 206.



Figur 4.11: Arealet som vist representerer sannsynligheten $P(Z \leq z)$.

[9] Dette integralet er et *Gaussintegral*. Derfor velges bokstaven G.
(En del engelske bøker bruker notasjonen $\Phi(z)$ istedet for $G(z)$.)

$P(Z \leq z) = G(z)$

Tabell 4. Normalkurven (arealtabell)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	–	–	–	–	–	–	–	–	–

Eksempel: ($P(X \leq 4)$, $\overbrace{\text{normal}}^{\text{kont.}}$fordeling)

Anta at $\overbrace{X \sim N[\mu = 3, \sigma = 1.4]}^{\text{generell normalfordeling}}$, dvs. anta at X er normalfordelt med $E[X] = \mu = 3$ og $\sqrt{Var[X]} = \sigma = 1.4$.

Hva er $P(X \leq 4)$?

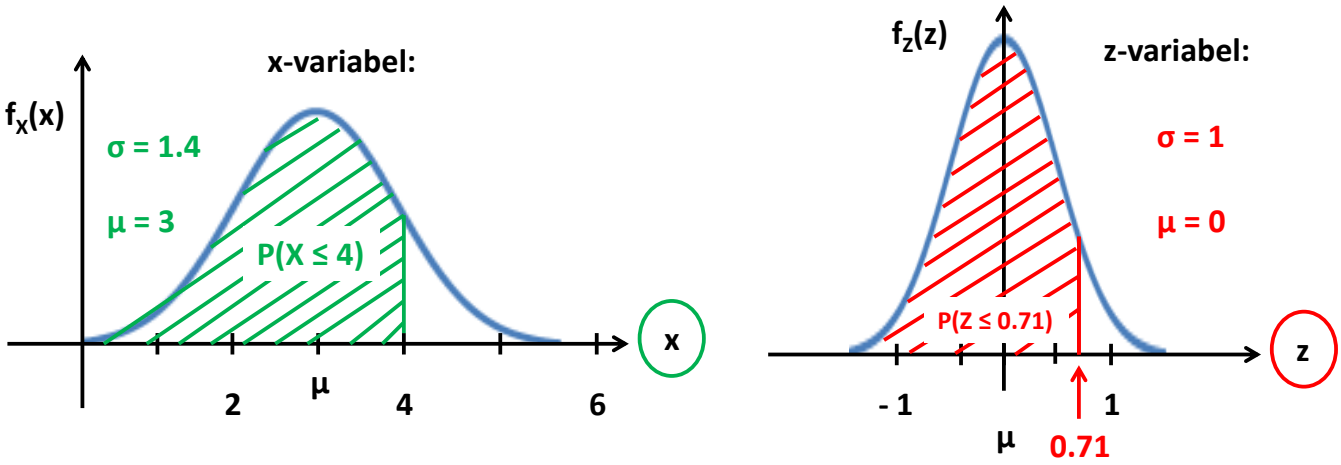
Løsning:

Matematisk vet vi at det er lov å trekke fra et tall på begge sider av et ulikhetstegn, dvs. ulikheten forblir den samme. Det er også lov å dele på samme positive tall på begge sider. Dermed:

$$P(X \leq 4) \overset{\text{standardiser}}{=} P\left(\frac{X-\mu}{\sigma} \leq \frac{4-\mu}{\sigma}\right) \tag{4.34}$$

$$= P\left(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \underbrace{\frac{4-3}{1.4}}_{\approx 0.71}\right) \tag{4.35}$$

$$= P(Z \leq 0.71) = G(0.71) \overset{\text{tabell}}{=} 0.7611 \tag{4.36}$$



Figur 4.12: Sannsynlighetene $P(X \leq 4)$ og $P(Z \leq 0.71)$ er like, dvs. det grønne og det røde arealet er likt.

■

Setning: ($P(X \leq x)$, $P(Z \leq z)$ og $G(z)$)

La $\overbrace{X \sim N[\mu, \sigma]}^{\text{generell n.-fordeling}}$ og $\overbrace{Z \sim N[\mu = 0, \sigma = 1]}^{\text{standard n.-fordeling}}$, dvs. X og Z være kontinuerlige stokastiske variabeler relatert på følgende måte:

$$\boxed{Z = \frac{X - \mu}{\sigma}} \tag{4.37}$$

Da er den *kumulative* sannsynlighetsfordelingen gitt ved:

$$\boxed{\underbrace{P(X \leq x)}_{\text{integral}} = \underbrace{P(Z \leq z) = G(z)}_{\text{tabelloppslag}}} \tag{4.38}$$

hvor [10]

$$\underbrace{G(z) = \int_{-\infty}^{z} f_Z(z)\, dz}_{\text{tabelloppslag for normalfordeling}} \tag{4.39}$$

og $f_Z(z)$ er gitt ved lign.(4.26).

■

[10] Dette integralet $G(z)$, Gaussintegralet, behøver ikke regnes ut. Det er tabelloppslag. Tabell finnes på side 206.

Setning: (egenskap til $\overbrace{G(z)}^{\text{tabelloppslag}}$)

Fra figur (4.13) innser vi at

$$G(z) \quad + \quad G(-z) \quad = \quad \text{arealet under hele kurven} \quad = \quad 1 \tag{4.40}$$

Altså vi kan skrive:

$$\boxed{G(-z) \quad = \quad 1 \quad - \quad G(z)} \tag{4.41}$$

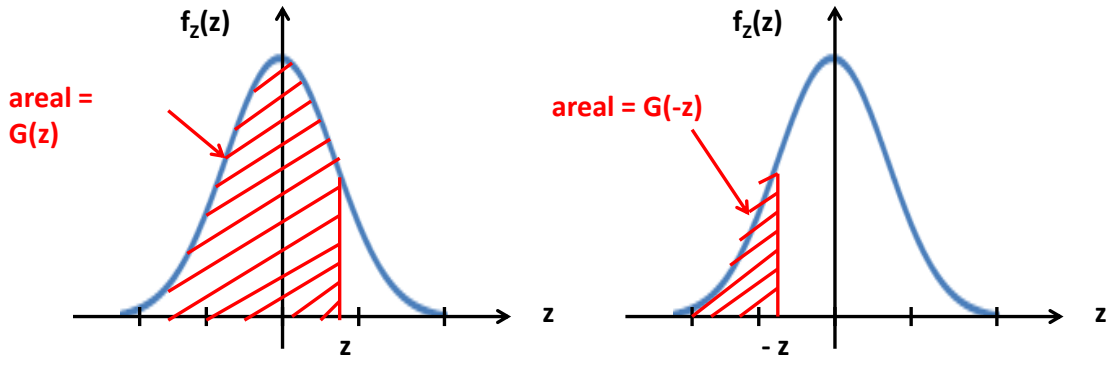
eller ekvivalent

$$P(Z \leq -z) \quad = \quad 1 \quad - \quad P(Z \leq z) \tag{4.42}$$

■

Dette betyr at vi kun behøver $G(z)$ med positivt argument.

$G(z)$ (med positivt argument) er som tidligere nevnt tabelloppslag.
En slik tabell finnes på side 206.



Figur 4.13: Arealene representerer Gauss-integralene $G(z)$ og $G(-z)$.

Eksempel: ($\overbrace{\text{normal}}^{\text{kont.}}$fordeling , logistikk)

En bedrift som produserer rør som settes sammen til gassrørledninger. Bedriften har en maskin som produserer rør, som til en bestemt rørledning skal være omtrent 9 meter lang. Lengden til rør produsert av maskinen er med god tilnærmelse beskrevet av en normalfordeling. Anta at denne maskinen produserer rør som har en forventning

$$\mu = 9 \text{ meter} \tag{4.43}$$

og standardavvik

$$\sigma = 0.1 \text{ meter} \tag{4.44}$$

a) Hvordan vil du definere den stokastiske variabelen i dette problemet?

b) Hva er sannsynligheten for at et tilfeldig valgt rør er 9.2 meter eller lengere?

c) Hva er sannsynligheten for at lengden et tilfeldig valgt rør ligger mellom 8.9 meter og 9.1 meter?

Figur 4.14: Gassrør som lagres etter produksjon.

Du jobber i bedriften som ansvarlig for leveranse av gassrør til store prosjekt. Herunder kommer også tilhørende logistikk og kvalitetssikring. Ett av flere mål på kvalitet i denne sammenheng er presisjonen av lengden på rørene.

Bedriften har fått i oppdrag fra A/S Norske Shell å produsere 11.25 kilometer med gassrør for Shell på Nyhamna i Aukra. Shell krever en feilmargin på ± 5 meter. Dette betyr at Shell kan leve med at den totale lengden av rørledningen er 5 meter lengere eller kortere enn spesifisert verdi.

d) Hva er forventet totallengde av rørledningen?

e) Hva er variansen til totallengden? Anta at lengden til hvert rør er uavhengige.

f) Hva er sannsynligheten for at totallengden blir for lang eller for kort?

g) Dersom standardavviket til lengden av et gitt rør, σ, kunne justeres til en annen verdi, hvilken verdi måte det justeres til for at sannsynligheten for at rør ledningen blir for lang eller for kort, skal bli 1 %?

Figur 4.15: Gassrør på Nyhamna i Aukra.

Løsning:

a) I situasjonen som beskrevet i oppgaven er det hensiktsmessig å definere den stokastiske variabelen:

$$\underline{\underline{X \;=\; \text{lengden av et tilfeldig valgt rør}}} \tag{4.45}$$

b) Sannsynligheten for at et tilfeldig valgt rør er 9.2 meter eller lengere:

$$\underline{\underline{P\big(X \geq 9.2\big)}} \overset{\text{standardiser}}{=} P\Bigg(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \geq \frac{9.2-\mu}{\sigma}\Bigg) \;=\; P\Bigg(Z \geq \underbrace{\frac{9.2-9}{0.1}}_{=2}\Bigg) \tag{4.46}$$

$$\overset{\text{lign.(4.42)}}{=} \;1 \;-\; P(Z \leq 2) \overset{\text{lign.(4.39)}}{\approx} \;1 \;-\; \underbrace{G(2)}_{\text{se tabell}} \tag{4.47}$$

$$= \;1 \;-\; 0.9772 \;=\; \underline{\underline{0.023}} \tag{4.48}$$

c) Sannsynligheten for at lengden et tilfeldig valgt rør ligger mellom 8.9 meter og 9.1 meter?

$$\underline{P\big(8.9 \leq X \leq 9.1\big)} \;=\; P(X \leq 9.1) \;-\; P(X \leq 8.9) \tag{4.49}$$

$$= \;P\Bigg(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{9.1-\mu}{\sigma}\Bigg) \;-\; P\Bigg(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{8.9-\mu}{\sigma}\Bigg) \tag{4.50}$$

$$= \;P\Bigg(Z \leq \underbrace{\frac{9.1-9}{0.1}}_{=1}\Bigg) \;-\; P\Bigg(Z \leq \underbrace{\frac{8.9-9}{0.1}}_{=-1}\Bigg) \tag{4.51}$$

$$= \;P(Z \leq 1) \;-\; \Big(1 - P(Z \leq 1)\Big) \tag{4.52}$$

$$= \;\underline{2\,P(Z \leq 1) \;-\; 1} \tag{4.53}$$

$$\underline{\underline{P\big(8.9 \leq X \leq 9.1\big)}} \quad = \quad 2\,P(Z \leq 1) \;-\; 1 \tag{4.54}$$

$$\overset{\text{lign.(4.39)}}{=} \quad 2\,\underbrace{G(1)}_{\text{se tabell}} \;-\; 1 \tag{4.55}$$

$$= \quad 2 \cdot 0.8413 \;-\; 1 \;=\; \underline{\underline{0.6826}} \tag{4.56}$$

d) Shell skal ha 11 250 meter med rør. Hvert rør er 9 meter, dvs. man trenger totalt $\frac{11\,250}{9} = 1250$ rør. Forventet totallengde av rørledningen blir dermed:

$$\underline{\underline{E\big[\,X_{\text{total}}\,\big]}} \quad = \quad E\big[\,X_1 \;+\; X_2 \;+\; \ldots \;+\; X_{1149} \;+\; X_{1150}\,\big] \tag{4.57}$$

$$\overset{\text{lign.(2.67)}}{=} \quad \overbrace{E[X_1] \;+\; E[X_2] \;+\; \ldots \;+\; E[X_{1149}] \;+\; E[X_{1250}]}^{=\ 1250\ \text{stk.}} \tag{4.58}$$

$$= \quad 1250 \cdot \overbrace{E[X]}^{=\mu=9} \tag{4.59}$$

$$= \quad 1250 \cdot 9 \;\; \text{meter} \tag{4.60}$$

$$= \quad \underline{\underline{11\,250 \;\; \text{meter}}} \tag{4.61}$$

e) Siden lengden til hvert rør er uavhengige, så gjelder:

$$\underline{\underline{Var\big[\,X_{\text{total}}\,\big]}} = Var\big[\,X_1 \;+\; X_2 \;+\; \ldots \;+\; X_{1249} \;+\; X_{1250}\,\big] \tag{4.62}$$

$$\overset{\text{uavh.}}{=} \overbrace{Var[X_1] \;+\; Var[X_2] \;+\; \ldots \;+\; Var[X_{1249}] \;+\; Var[X_{1250}]}^{=\ 1250\ \text{stk.}} \tag{4.63}$$

$$= 1250 \cdot \overbrace{Var[X]}^{=0.1^2} \tag{4.64}$$

$$= 1250 \cdot 0.1^2 \;\; \text{meter}^2 \tag{4.65}$$

$$= \underline{\underline{12.5 \;\; \text{meter}^2}} \tag{4.66}$$

med tilhørende standardavvik

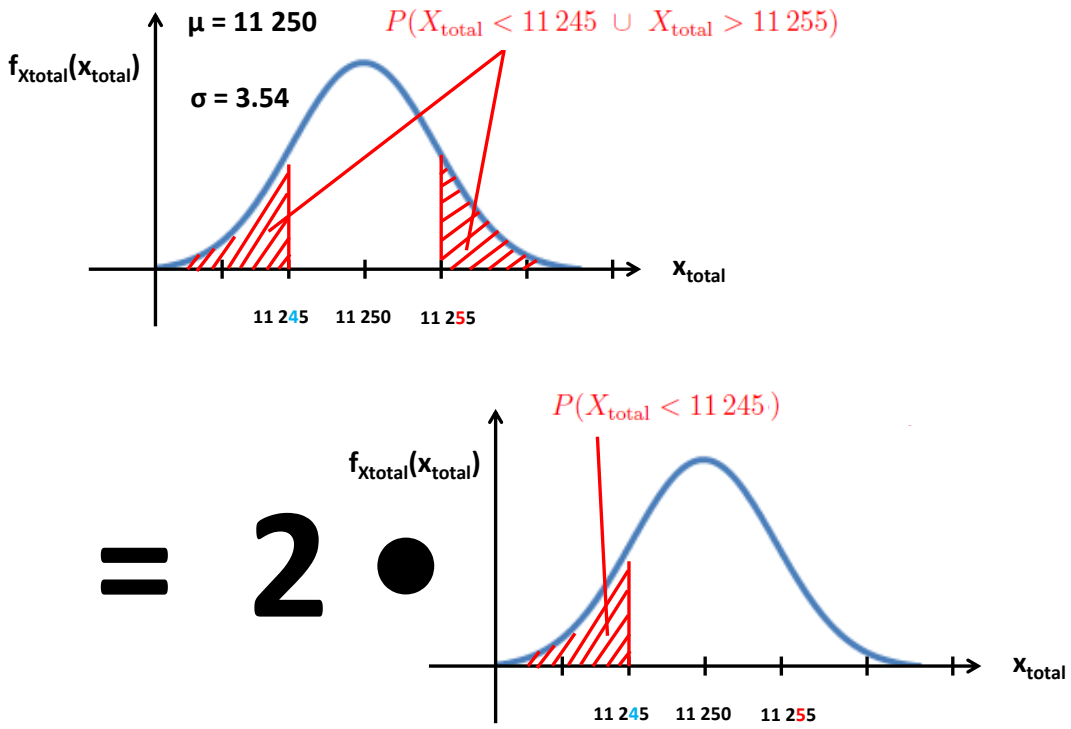
$$\underline{\underline{\sigma[X_{\text{total}}]}} = \sqrt{Var[X_{\text{total}}]} = \sqrt{12.5} \approx \underline{\underline{3.54 \text{ meter.}}} \tag{4.67}$$

f) Sannsynligheten for at blir totallengden blir for lang eller for kort:

$$P\big(X_{\text{total}} < 11\,245 \overbrace{\cup}^{\text{eller}} X_{\text{total}} > 11\,255\big) = P(X_{\text{total}} < 11\,245) + P\big(X_{\text{total}} > 11\,255\big) \tag{4.68}$$

$$= 2 \cdot P(X_{\text{total}} < 11\,245) \tag{4.69}$$

hvor disse to overgangene kan forklares via figur (4.16).



Figur 4.16: Visualisering av lign.(4.66).

$$\underline{P\big(\, X_{\text{total}} < 11\,245 \overbrace{\cup}^{\text{eller}} X_{\text{total}} > 11\,255 \,\big)} \quad = \quad 2 \cdot P\big(X_{\text{total}} < 11\,245\big) \tag{4.70}$$

$$= \quad 2 \cdot P\Bigg(\underbrace{\frac{X - \mu}{\sigma}}_{= Z} \leq \underbrace{\frac{11\,245 - 11\,250}{3.54}}_{= -\,1.41} \Bigg)$$

$$= \quad 2 \cdot P\big(Z \leq -1.41\big) \tag{4.71}$$

$$= \quad 2 \cdot \Big(1 - \underbrace{P(Z \leq 1.41)}_{= 0.9207} \Big) \tag{4.72}$$

$$= \quad 2 \cdot \Big(1 - 0.9208 \Big) \tag{4.73}$$

$$\approx \quad \underline{\underline{0.16}} \tag{4.74}$$

g) Vi skal finne det standardavviket σ for lengden av et gitt rør, som gjør at sannsynligheten for at rør ledningen blir for lang eller for kort, skal bli 1 %:

$$P(\text{for kort} \quad \text{eller} \quad \text{for lang}) \quad = \quad 0.01 \tag{4.75}$$

$$\underbrace{P(X_{\text{total}} < 11\,245 \ \cup \ X_{\text{total}} > 11\,255)}_{\overset{\text{spes. add.}}{=} P(X_{\text{total}}<11\,245) \ + \ P(X_{\text{total}}>11\,255)} \quad = \quad 0.01 \tag{4.76}$$

Siden begivenheten "for kort" og begivenheten "for lang" ikke overlapper, dvs. de er disjunkte så kan vi bruke den spesielle addisjons setning:

$$P(X_{\text{total}} < 11\,245) \quad + \quad P(X_{\text{total}} > 11\,255) \quad = \quad 0.01 \tag{4.77}$$

Siden normalfordelingen er symmetrisk så innser f.eks. fra figur (4.16) at:

$$P(X_{\text{total}} < 11\,245) \quad = \quad P(X_{\text{total}} > 11\,255) \tag{4.78}$$

dvs. sannsynligheten for at rørledningen er "for kort" er den samme som sannsynligheten for at den er "for lang", $P(\text{for kort}) \quad = \quad \text{P(for lang)}$. Lign.(4.77) gir dermed:

$$2 \cdot P(X_{\text{total}} < 11\,245) \quad = \quad 0.01 \tag{4.79}$$

$$P(X_{\text{total}} < 11\,245) \quad = \quad \frac{0.01}{2} \tag{4.80}$$

$$P\left(\underbrace{\frac{X_{\text{total}} - \mu_{\text{total}}}{\sigma[X_{\text{total}}]}}_{= Z_{\text{total}}} \leq \frac{11\,245 - \mu_{\text{total}}}{\sigma[X_{\text{total}}]}\right) \quad \overset{\text{standardiser}}{=} \quad \frac{0.01}{2} \tag{4.81}$$

$$P\left(Z_{\text{total}} \leq \frac{11\,245 - 11\,250}{\sqrt{1250 \cdot \sigma^2}}\right) \quad = \quad 0.005 \tag{4.82}$$

$$\underbrace{P\left(Z_{\text{total}} \leq -\frac{0.1414}{\sigma}\right)}_{= \ 1 \ - \ P\left(Z_{\text{total}} \leq \frac{0.1414}{\sigma}\right)} \quad = \quad 0.005$$

hvor vi har brukt at standardavviket for den totale rørledningen er $\sigma[X_{\text{total}}] = \sqrt{Var[X_{\text{total}}]} = \sqrt{1250 \cdot \sigma^2}$.

$$P\left(Z_{\text{total}} < \frac{0.1414}{\sigma}\right) \quad = \quad \underbrace{1 \; - \; 0.005}_{= \; 0.9950} \tag{4.83}$$

Ved "**omvendt tilbabelloppslag**": tallet 0.9950 ligger midt mellom 0.9949 og 0.9951, se arealtabellen på side 206. Dette tilsvarer at *argumentet* er 2.575. Dermed:

$$\frac{0.1414}{\sigma} \quad = \quad 2.575 \tag{4.84}$$

$$\underline{\sigma} \quad = \quad \frac{0.1414}{2.575} \quad = \quad \underline{0.055} \tag{4.85}$$

Konklusjon:

Dersom standardavviket ("usikkerheten"/spredningen) til lengden av et gitt rør reduseres til $\underline{\underline{\sigma = 0.055 \text{ meter}}}$ (= 5.5 cm) så er det 1 % sannsynlighet for at dem totale rørlednigen er for kort eller for lang.

■

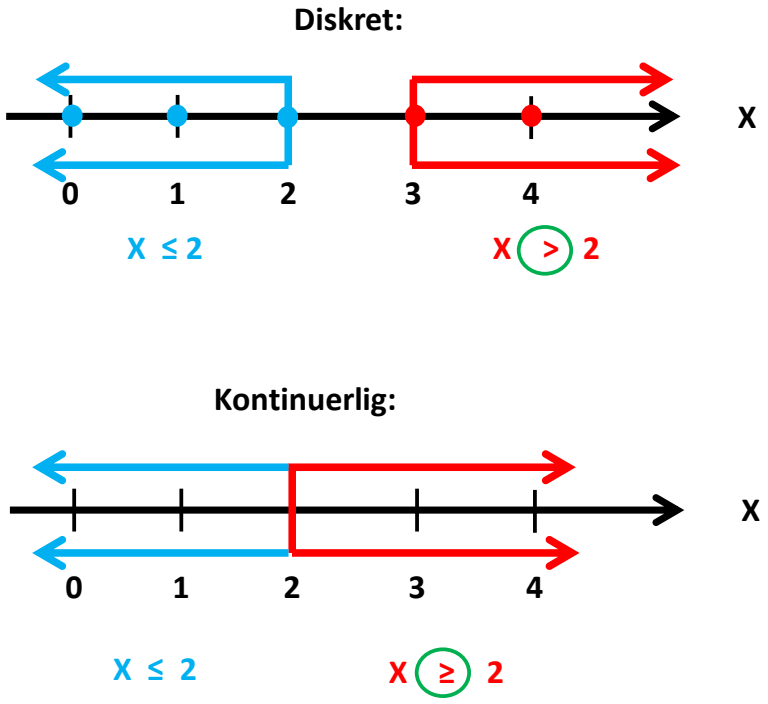
4.2.4 Diskret vs kontinuerlig fordeling: en viktig forskjell

En viktig forskjell mellom en diskret og kontinuerlig sannsynlighetsfordeling er følgende:

$$P(X \leq 2) \overset{\text{disk.}}{=} 1 - P(X \overbrace{\geq}^{!} 3) \qquad \text{(diskret)} \tag{4.86}$$

$$P(X \leq 2) \overset{\text{kont.}}{=} 1 - P(X \underbrace{\geq}_{!} 2) \qquad \text{(kontinuerlig)} \tag{4.87}$$

I det kontinuerlige tilfellet, siden integralet over kun et punkt er null, $\int_a^a f_X(x)\,dx = 0$, så kan vi både ha mindre enn "**eller lik**", $\leq$, på venstre siden i lign.(4.86), og også "**eller lik**" på høyre side. Det kan vi ikke i det diskrete tilfellet. [11]



Figur 4.17: Diskret vs kontinuerlig fordeling: en viktig forskjell.

[11] I det kontinuerlige tilfellet så kan man også dette skrives på følgende måte:

$$\underline{\underline{P(X \leq x)}} \overset{\text{kont.}}{=} P(X < x) + \underbrace{P(X = x)}_{\overset{\text{kont.}}{=} 0} = \underline{\underline{P(X < x)}} \qquad \text{(kontinuerlig)} \tag{4.88}$$

4.2.5 Standardavvik σ og %-vis areal

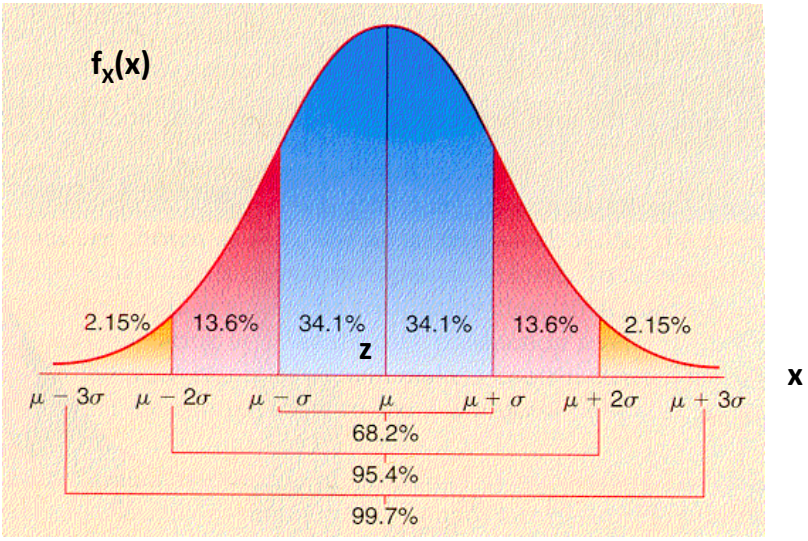
Som vi har lært tidligere så er varians og standardavvik et mål på spredning. For en normalfordeling med et gitt standardavvik σ, så dekker intervallet

$$\mu - \sigma \ \leq \ X \ \leq \ \mu + \sigma \quad \text{hele } 68.2\ \% \text{ av arealet} \tag{4.89}$$

under sannsynlighetfordelingen $f_X(x)$. Tilsvarende dekker intervallet

$$\mu - 2\sigma \ \leq \ X \ \leq \ \mu + 2\sigma \quad \text{hele } 95.4\ \% \text{ av arealet} \tag{4.90}$$

under sannsynlighetfordelingen $f_X(x)$. Dette er illustrert i denne figuren:



Figur 4.18: Standardavvik σ og %-vis areal for en normalfordeling.

Eksempel: ($\overbrace{\text{normal}}^{\text{kont.}}$fordeling , økonomi)

Anta at du er økonomisjef i et selskap som produserer kjøleskapmotorer. For ha kontroll på økonomien til selskapet så må selskapet vite litt om levetiden til disse motorene. Det viser set at levetiden er normalfordelt med forventet levetid på $\underbrace{19.4 \text{ år}}_{=\mu}$. Standardavviket er $\underbrace{4.3 \text{ år}}_{=\sigma}$.

a) Hva er sannsynligheten for at en motor fungerer i 12 år eller mindre?

b) Hva er sannsynligheten for at en motor fungerer i 25 år eller mer?

c) Hva er sannsynligheten for at en motor fungerer mer enn 10 år, men mindre enn 20 år?

Figur 4.19: Produksjon av kjøleskap og kjøleskapmotorer.

For å fremme salget av sine motorerønsker selskapet å gå ut med en levetidsgaranti.
Denne garantien går ut på at kunden kostnadfritt får ny motor dersom den ryker innenfor garantitiden.

d) Hvor mange års garanti kan selskapet gå ut med dersom de ikke ønsker å erstatte mer enn 2 % av motorene?

Selskapet tjener totalt 1200 NOK på en moter som holder hele garantitiden. For motorer som ryker innenfor garantitiden så taper selskapet 4500 NOK.

e) **i**) Dersom selskapet opererer med 12 år garanti på motorene, hva er da forventet fortjeneste på salg av en motor?

ii) Hvordan vil du tolke resultatet i forrige oppgave?

Løsning:

La [12]

$$X \;=\; \text{levetiden til en } \textit{tilfeldig} \text{ valgt motor} \tag{4.91}$$

Siden forventet levetid på 19.4 år med standardavvik på 4.3 år, så har vi:

$$\mu \;=\; 19.4 \qquad , \qquad \sigma \;=\; 4.3 \tag{4.92}$$

a) Sannsynligheten for at en motor fungerer i 12 år eller mindre:

$$\underline{\underline{P(X \le 12)}} \;=\; P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \le \frac{12-\mu}{\sigma} \Big) \;=\; P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \le \underbrace{\frac{12-19.4}{4.3}}_{=-1.72} \Big) \tag{4.93}$$

$$= \;=\; P(Z \le -1.72) \;=\; 1 \;-\; \underbrace{P(Z \le 1.72)}_{=G(1.72)} \;=\; 1 \;-\; \underbrace{G(1.72)}_{\text{se tabell}} \tag{4.94}$$

$$= \;\; 1 \;-\; \underbrace{0.9573}_{\text{se tabell}} \;=\; \underline{\underline{0.043}} \tag{4.95}$$

[12] X er altså en kontinuerlig stokastisk variabel.

b) Sannsynligheten for at en motor fungerer i 25 år eller mer:

$$\underline{\underline{P(X \geq 25)}} = 1 - P(X \leq 25) \tag{4.96}$$

$$= 1 - P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{25-\mu}{\sigma}\Big) = 1 - P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \underbrace{\frac{25-19.4}{4.3}}_{=1.30}\Big) \tag{4.97}$$

$$= = 1 - P(Z \leq 1.30) = 1 - \underbrace{P(Z \leq 1.30)}_{=G(1.30)} = 1 - \underbrace{G(1.72)}_{\text{se tabell}} \tag{4.98}$$

$$= 1 - \underbrace{0.9032}_{\text{se tabell}} = \underline{\underline{0.097}} \tag{4.99}$$

c) Sannynligheten for at en motor fungerer mer enn 10 år, men mindre enn 20 år:

$$\underline{\underline{P(10 \leq X \leq 20)}} = P(X \leq 20) - P(X \leq 10) \tag{4.100}$$

$$= P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{20-\mu}{\sigma}\Big) - P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \frac{10-\mu}{\sigma}\Big) \tag{4.101}$$

$$= P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \underbrace{\frac{20-19.4}{4.3}}_{=0.14}\Big) - P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \underbrace{\frac{10-19.4}{4.3}}_{=-2.19}\Big) \tag{4.102}$$

$$= P(Z \leq 0.14) - P(Z \leq -2.19) \tag{4.103}$$

$$= \underbrace{P(Z \leq 0.14)}_{=G(0.14)} - (1 - \underbrace{P(Z \leq 2.19)}_{=G(2.19)}) \tag{4.104}$$

$$= 0.5557 - (1 - \underbrace{0.9857}_{\text{se tabell}}) = \underline{\underline{0.5414}} \tag{4.105}$$

d) La X_g være den ukjente garantitiden som vi skal finne. Dersom selskapet ikke ønsker å erstatte mer enn 2 % av motorene, så betyr det:

$$P(X \leq X_g) \quad = \quad 0.02 \tag{4.106}$$

La oss standardisere denne:

$$P\left(\underbrace{\frac{X-\mu}{\sigma}}_{=Z} \leq \underbrace{\frac{X_g-\mu}{\sigma}}_{\text{negativ}} \right) \quad = \quad 0.02 \tag{4.107}$$

$$1 - P\left(Z \leq \underbrace{\frac{\mu - X_g}{\sigma}}_{\text{positiv}} \right) \quad = \quad 0.02 \tag{4.108}$$

$$P\left(Z \leq \frac{\mu - X_g}{\sigma} \right) \quad = \quad 0.98 \tag{4.109}$$

Ved "**omvendt tilbabelloppslag**" ser vi at 0.9798 er det som er nærmest 0.98, se arealtabellen side 206. Dette tilsvarer at *argumentet* er 2.05. Dermed:

$$Z \quad = \quad 2.05 \tag{4.110}$$

$$\frac{\mu - X_g}{\sigma} \quad = \quad 2.05 \tag{4.111}$$

$$X_g \quad = \quad 10.6 \tag{4.112}$$

Konklusjon:

Dersom selskapet ikke ønsker å erstatte mer enn 2 % av motorene, så kan selskapet max gå ut med en levetidsgaranti på 10.6 år.

e) **i)** La

$$F \;=\; \text{fortjenesten til selskapet ved salg av en } \textit{tilfeldig} \text{ valgt motor} \tag{4.113}$$

Utfallet til F er enten 1200 kroner eller -4500 kroner, dvs. $\Omega = \{\,-4500, 1200\,\}$.
Fra kap.(2) og lign.(2.17) vet vi at forventning er:

$$\underline{E(F)} \overset{\text{lign.(2.17)}}{=} \sum_{i=1}^{m} f_i \cdot P(F = f_i) \tag{4.114}$$

$$= \underline{1200 \cdot \overbrace{P(F = 1200)}^{\text{må finne denne}} \;+\; (-4500) \cdot \overbrace{P(F = -4500)}^{\text{må finne denne}}} \tag{4.115}$$

Men sannsynlighten for å tjene 1200 kroner, $P(F = 1200)$, er lik sannsynligheten for at motoren holder i hele garantitiden. Altså:

$$\underline{P(F = 1200)} \;=\; \text{sanns. for å tjene 1200 kroner ved salg av en tilfeldig motor} \tag{4.116}$$

$$= \text{sanns. for at motoren holder i hele garantitiden på 12 år} \tag{4.117}$$

$$= P(X \geq 12) \tag{4.118}$$

$$= 1 \;-\; P(X \leq 12) \tag{4.119}$$

$$= 1 \;-\; P\left(\underbrace{\frac{X - \mu}{\sigma}}_{=Z} \leq \frac{12 - \mu}{\sigma} \right) \tag{4.120}$$

$$= 1 \;-\; P\left(Z \leq \underbrace{\frac{12 - 19.3}{4.3}}_{=-1.72} \right) \tag{4.121}$$

$$= 1 \;-\; P(Z \leq -1.72) \tag{4.122}$$

$$= \cancel{1} - \left(\cancel{1} - P(Z \geq 1.72) \right) \tag{4.123}$$

$$= P(Z \geq 1.72) \;=\; \underline{0.9573} \tag{4.124}$$

Tilsvarende: sannsynlighten for å tjene 1200 kroner, $P(F = 1200)$, er like sannsynligheten for at motoren holder i hele garantitiden. Altså:

$$\underline{P(F=-4500)} = \text{sanns. for å tjene -4500 kroner ved salg av en tilfeldig motor} \quad (4.125)$$

$$= \text{sanns. for at motoren } \textit{ikke} \text{ holder i hele garantitiden på 12 år} \quad (4.126)$$

$$= P(X \leq 12) \quad (4.127)$$

$$= 1 - \underbrace{P(X \geq 12)}_{=0.9573} \quad (4.128)$$

$$= 1 - 0.9573 = \underline{0.0427} \quad (4.129)$$

Innsatt i lign.(4.115):

$$\underline{\underline{E(F)}} \overset{\text{lign.(4.115)}}{=} 1200 \cdot \overbrace{P(F=1200)}^{0.9573} + (-4500) \cdot \overbrace{P(F=-4500)}^{0.0427} \quad (4.130)$$

$$= 1200 \cdot 0.9573 + (-4500) \cdot 0.0427 = \underline{\underline{956.6}} \quad (4.131)$$

e) **ii**) Tolking:

$$\underline{\underline{E[F]}} = \text{gjennomsnittlig fortjeneste per solgte motor } \textit{i det lange løp} \quad (4.132)$$

■

4.3 Oversikt: Bin, Hyp, Poi og N

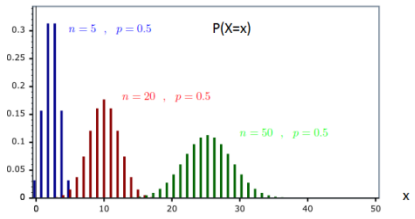
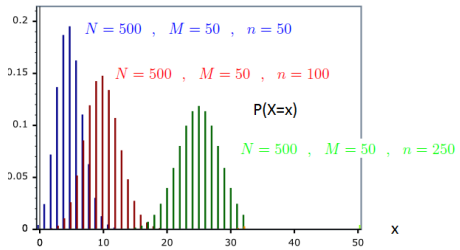
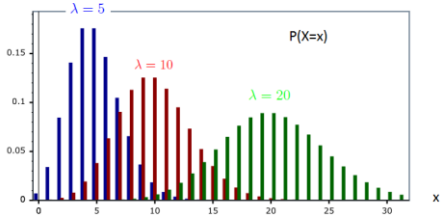
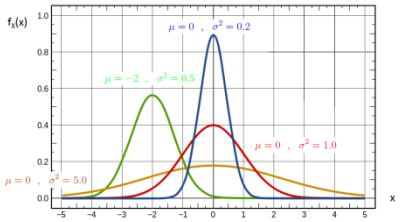
Vi har i dette kapitlet lært om fire sannsynlighetsfordelinger:

1. Binomisk fordeling
2. Den hypergeometriske fordeling
3. Poissonfordelingen
4. Normalfording

På de to neste sidene er det laget en oversikt over disse fire fordelingene. Særlig viktig er kommentarene. Disse kommentarene som sier blant annet noe om for hvilke situasjoner de respektive fordelingene beskriver.

Når man støter på en situasjon hvor man skal avgjøre hvilken sannsynlighetsfordeling som er kan være aktuell for å beskrive/modellere en situasjon så kan en slik oversikt være til stor hjelp.

Bin[n , p]	Hyp[N , M , n]	Poi[λ]	N[μ , σ]
2 param.	3 param.	1 param.	2 param.
diskret	diskret	diskret	kontinuerlig
$P(X=x) \overset{\text{def.}}{=} \binom{n}{x} p^x (1-p)^{n-x}$	$P(X=x) \overset{\text{def.}}{=} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$	$P(X=x) \overset{\text{def.}}{=} \frac{\lambda^x}{x!} e^{-\lambda}$	$f_X(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$
$E[X] = n \cdot p$	$E[X] = n \cdot \frac{M}{N}$	$E[X] = \lambda$	$E[X] = \mu$
$Var[X] = n \cdot p\,(1-p)$	$Var[X] = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$	$Var[X] = \lambda$	$Var[X] = \sigma^2$

Bin[n , p]	Hyp[N , M , n]	Poi[λ]	N[μ , σ]
1) 2 mulig utfall 2) samme ”p” for ”suksess” 3) uavhengige 4) n antall forsøk	1) x antall ”suksesser” / ”spesielle” 2) N antall i grunnmengden 3) M antall ”spesielle” 4) n antall trukne elementer	1) x antall begivenheter innenfor en gitt tid 2) λ = rate	1) Tetthetsfunksjon $f_X(x)$ 2) Gausskurve
- kjenner ikke fordelingen i urnen - m / tilbakelegging - teller opp antall ”suksesser”	- kjenner fordelingen i urnen - u / tilbakelegging - teller opp antall ”suksesser”	- rate (konstant) - antall begivenheter innenfor en gitt tid eller gitt rom - telleforsøk - ”loven om skjeldne begivenh.”	- under bestemte betingelser vil mange diskrete og kontinuerige fordelinger med god tilnæring være normalfordelt (f.eks. ”CLT”)
			

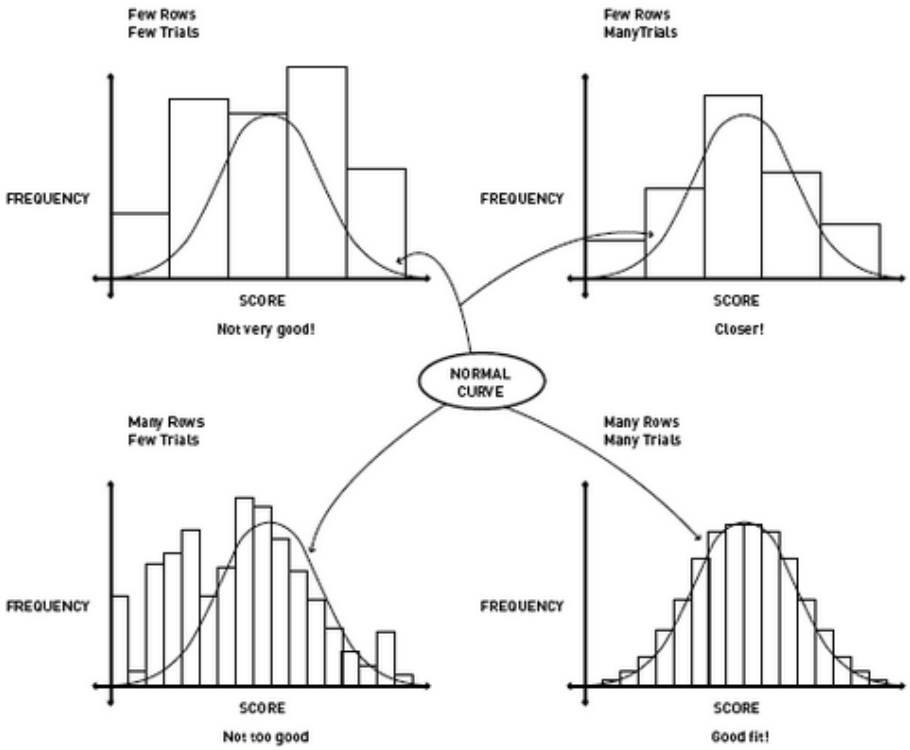
4.4 Sentralgrensesetningen

Sentralgrenseteoremet er et sentralt teorem (=læresetning) innen matematisk statistikk og sannsynlighetsteori. Teoremet sier at en

$$\text{sum av uavhengige og } \overbrace{\text{identisk fordelte stokastiske variabler}}^{\text{diskret ELLER kont.}}$$

går mot en **normal** fordeling når antallet går mot uendelig.

!



Figur 4.20: Visualisering av sentralgrenseteoremet ("CLT").

Eksempel: (n stk. terninger , sentralgrensesetningen)

I dette eksemplet skal vi se på terningkast. Med flere terninger. Vi definerer:

$$n = \text{antall terninger i et terningkast} \tag{4.133}$$

$$\overline{X} = \text{gjennomsnittet av antall øyne med } n \text{ terninger} \tag{4.134}$$

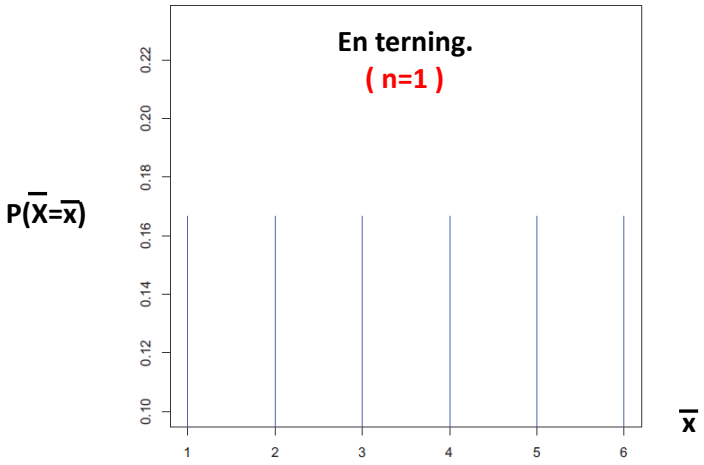
Her er n bare en konstant. $\overline{X}$ er en stokastisk variabel.

Først skal vi gjøre et forsøk med kun èn terning, $n = 1$. Deretter gjør vi et nytt forsøk, men denne gangen med 2 terninger, $n = 2$. Deretter atter et nytt forsøk med 3 terninger $n = 3$ osv.

Figur 4.21: Terningkast.

i) $\mathbf{n = 1}$:

Èn terning.
Med bare èn terning, $n = 1$, så er åpenbart $\overline{x} = x_i$. Sannsynlighetsfordelingen til gjennomsnittet $\overline{x}$, dvs. $P(\overline{X} = \overline{x}) = 1/6 = 0.166...$, kan visualiseres slik:



Figur 4.22: $P(X = \overline{x})$ for *en* terning ($\mathbf{n = 1}$).

ii) $\mathbf{n = 2}$:

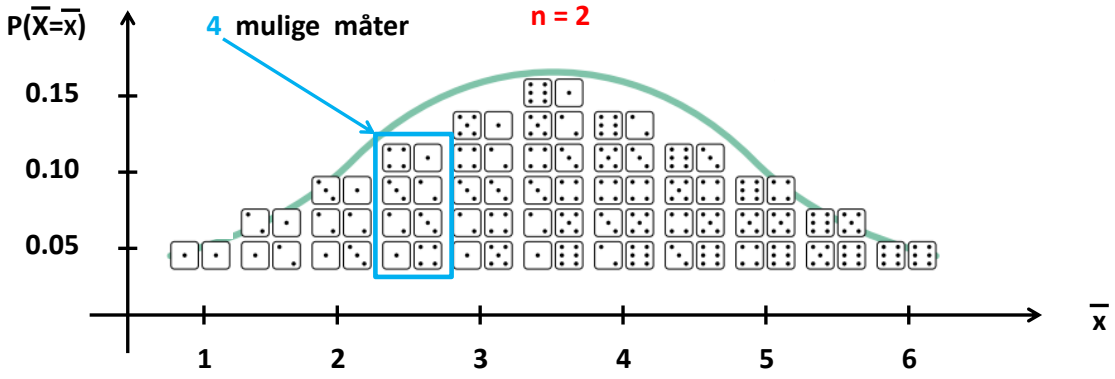
To terninger.
Dersom vi kaster to terninger, $n = 2$, så kan vi regne ut gjennomsnittet av antall øyne til teringkastene. F.eks., dersom utfallet blir 1 øyne og 4 øyne så er gjennomsnittet:

$$\overline{x} \overset{\text{lign.(5.21)}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{2}(1+4) = 2.5 \tag{4.135}$$

Sannsynligheten $P(\overline{X} = 2.5)$ for at våre $n = 2$ terninger skal gi et gjennomsnittet på $\overline{x} = 2.5$ kan finnes via urnemodellen siden dette er en **tellesituasjon**: Det er 4 mulige måter at å få gjennomsnittet $\overline{x} = 2.5$ på [13]. Totalt er det $6^2 = 36$ mulige utfall [14]. Sannsynligheten $P(\overline{X} = 2.5)$ er da gitt ved:

$$\underline{P(\overline{X} = 2.5)} = \frac{\text{antall } \textit{gunstige} \text{ utfall}}{\text{antall } \textit{mulige} \text{ utfall}} = \underline{\frac{4}{36}} \tag{4.136}$$

Tilsvarende kan vi regne ut $P(\overline{X} = \overline{x})$ for alle andre mulige utfall av $\overline{x}$.



Figur 4.23: $P(X = \overline{x})$ for man kaster $\mathbf{n = 2}$ terninger.

$P(X = \overline{x})$ for de forskjellige mulige verdiene av $\overline{x}$ finnes via tellemetoden:

$$P(\overline{X} = 1) = \frac{1}{36} \tag{4.137}$$

$$P(\overline{X} = 1.5) = \frac{2}{36} \tag{4.138}$$

$$P(\overline{X} = 2) = \frac{3}{36} \tag{4.139}$$

$$P(\overline{X} = 2.5) = \frac{4}{36} \tag{4.140}$$

$$P(\overline{X} = 3) = \frac{5}{36} \tag{4.141}$$

[13] Det er utfallene $(1, 4)$, $(4, 1)$, $(2, 3)$ og $(3, 2)$.

[14] Dette tilsvarer et ordnet utvalg med tilbakelegging, dvs. **situasjon 1**, se lign.(B.7) i kapittel B. Med $N = 6$ og $s = 2$ får man $N^s = 6^2 = 36$.

$$P(\overline{X} = 3.5) = \frac{6}{36} \tag{4.142}$$

$$P(\overline{X} = 4) = \frac{5}{36} \tag{4.143}$$

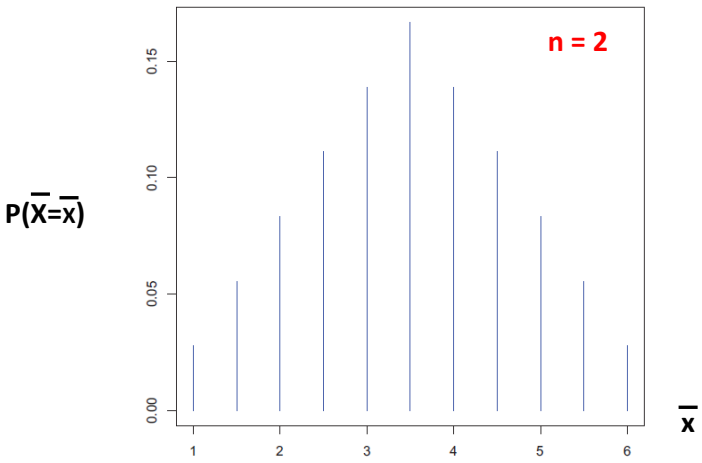
$$P(\overline{X} = 4.5) = \frac{4}{36} \tag{4.144}$$

$$P(\overline{X} = 5) = \frac{3}{36} \tag{4.145}$$

$$P(\overline{X} = 5.5) = \frac{2}{36} \tag{4.146}$$

$$P(\overline{X} = 6) = \frac{1}{36} \tag{4.147}$$

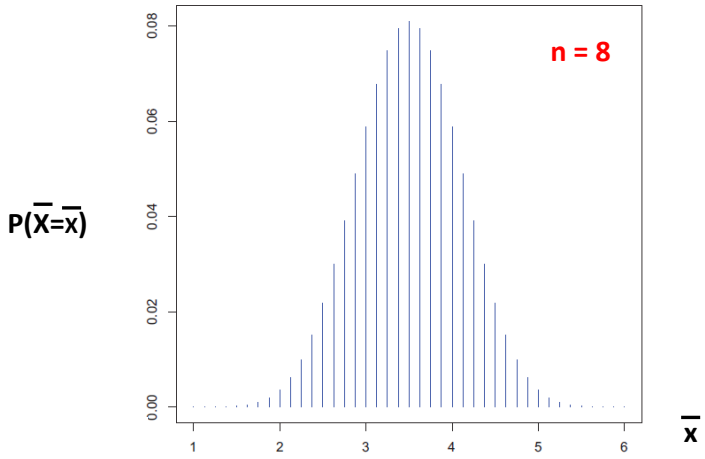
Sannsynlighetene i lign.(4.137)-(4.147) kan plottes i en figur:



Figur 4.24: $P(X = \overline{x})$ for *to* terninger, $\mathbf{n = 2}$.

iii) $n = 8$:

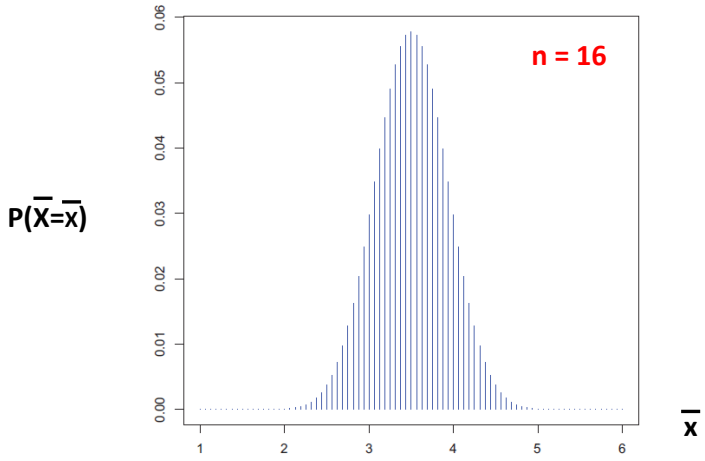
8 terninger.
Dersom vi kaster $n = 8$ terninger så kan vi regne ut sannsynlighetsfordelingen for gj.snittet $\overline{x}$ på samme måte som ovenfor. Resultatet blir:



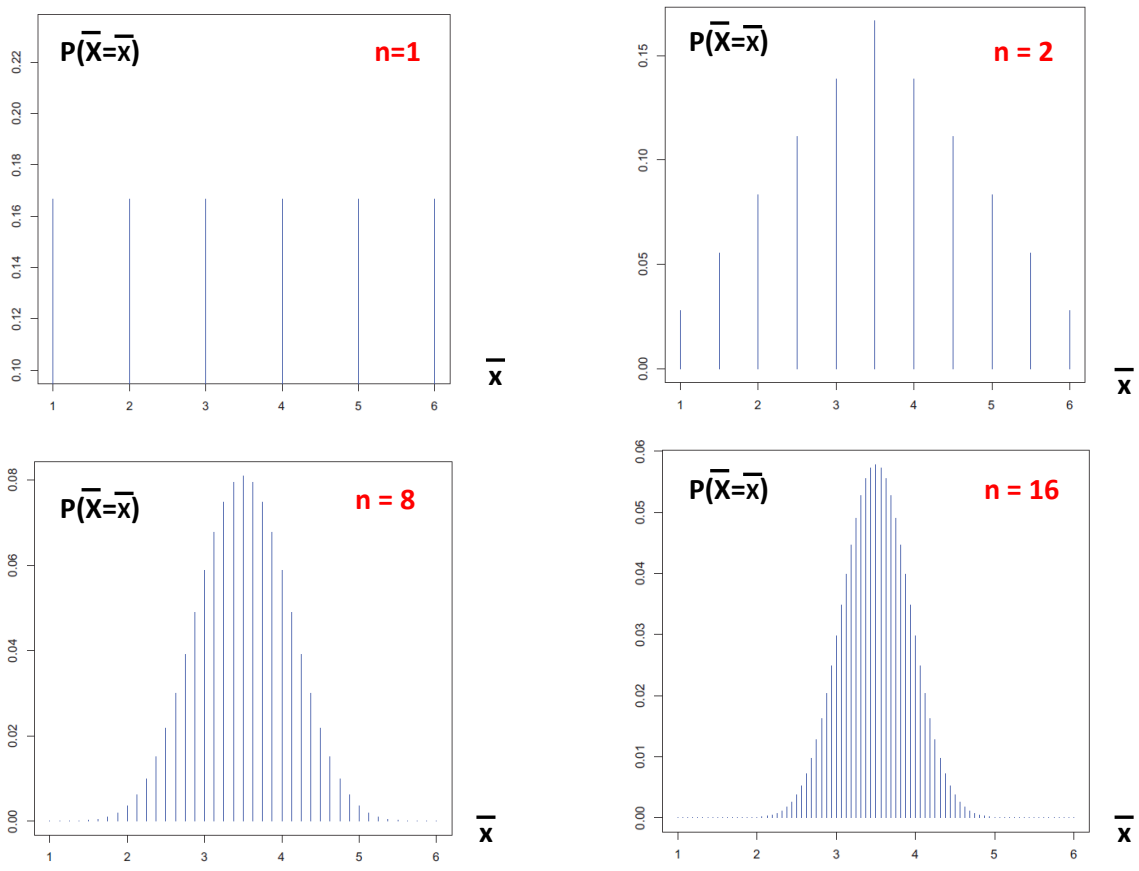
Figur 4.25: $P(X = \overline{x})$ for *åtte* terninger, $n = 8$.

iv) $n = 16$:

16 terninger.
Dersom vi kaster $n = 16$ terninger så blir sannsynlighetsfordelingen for gj.snittet $\overline{x}$:



Figur 4.26: $P(X = \overline{x})$ for 16 terninger, $n = 16$.



Figur 4.27: $P(X = \overline{x})$ for $\mathbf{n = 1, 2, 8, 16}$ terningkast.

Med forutseningene

- terningene er helt like [15]
- uavhengige terninger [16]

så ser vi at

- Lokaliseringsmål:
  Forventningen til gjennomsnittet av antall øyne $E[\overline{X}]$ er den samme som forventningen til antall øyne når man kun har èn terning $E[X_i]$:

$$\boxed{E[\overline{X}] \;=\; E[X_i] \;=\; 3.5 \qquad \text{(samme)} \qquad (4.148)}$$

- Spredningsmål:
  $P(\overline{X} = \overline{x})$ blir smalere og smalere,
  dvs. *variansen* til gjennomsnittet av terningkast $Var[\overline{X}]$ er **mindre** enn variansen til et enkeltstående terningkast $Var[X_i]$:

$$\boxed{Var[\overline{X}] \;<\; Var[X_i] \qquad \textbf{(mindre)} \qquad (4.149)}$$

[15] Hovedpoenget her er altså at alle **terningene** har samme sannsynlighetsfordeling. Altså at terningene er helt like. Hovedpoenget er ikke i denne sammenheng at sannsynligheten for et gitt utfall for en gitt terning også er det samme, $P(X = 1) = P(X = 2) = ...P(X = 6) = 1/6$.

[16] Altså at utfallet til hver enkelt terning er uavhengige av hverandre.

Setning: ("*CLT*" [17] , sentralgrensesetningen) (versjon 1)

La X_1, X_2, X_3, ... X_n være stokastiske variabler. Anta:

- alle X_i er uavhengige
- alle X_i har samme $\overbrace{\text{sannsynlighetsfordeling}}^{\text{diskret ELLER kont.}}$ $P(X = x_i)$,
  dvs. $E[X_i] = \mu$ og $Var[X_i] = \sigma^2$ for alle $i = 1, 2, 3, ..., n$.

Da gjelder at *gjennomsnittet* $\overline{X}$: (n = antall stok. var.)

$$\boxed{\overline{X} \;=\; \frac{X_1 \;+\; X_2 \;+\; X_3 + \;...\; +\; X_n}{n}} \qquad (4.150)$$

er normalfordelt i grensen når antall forsøk n **blir stor**:

$$\boxed{\overline{X} \;\overset{n=\text{stor}}{\approx}\; \mathrm{N}\left[\,\mu\,,\,\frac{\sigma}{\sqrt{n}}\,\right]} \qquad (4.151)$$

altså $P(\overline{X} = \overline{x})$ er normalfordelt med forventning og varians hhv.

$$\boxed{E[\,\overline{X}\,] \;=\; \mu \qquad \text{og} \qquad Var[\,\overline{X}\,] \;=\; \frac{\sigma^2}{n}} \qquad (4.152)$$

■

[17] På engelsk brukes ofte forkortelsen *CLT*, dvs. "*central limit theorem*".

Setning: ("*CLT*" [18] , sentralgrensesetningen) (versjon 2)

La X_1, X_2, X_3, ... X_n være stokastiske variabler. Anta:

- alle X_i er uavhengige
- alle X_i har samme $\overbrace{\text{sannsynlighetsfordeling}}^{\text{diskret ELLER kont.}}$ $P(X = x_i)$,
  dvs. $E[X_i] = \mu$ og $Var[X_i] = \sigma^2$ for alle $i = 1, 2, 3, ..., n$.

Da gjelder at summen Y_n: (n = antall stok. var.)

$$\boxed{Y \;=\; X_1 \;+\; X_2 \;+\; X_3 \;+\; ... \;+\; X_n} \qquad (4.153)$$

er normalfordelt i grensen når antall forsøk n **blir stor**:

$$\boxed{Y \;\overset{n=\text{stor}}{\approx}\; \mathrm{N}\Big[\, n\mu \,,\, \sqrt{n}\sigma \,\Big]} \qquad (4.154)$$

altså $P(Y = y)$ er normalfordelt med forventning og varians hhv.

$$\boxed{E[\,Y\,] \;=\; n\mu \qquad \text{og} \qquad Var[\,Y\,] \;=\; n\,\sigma^2} \qquad (4.155)$$

■

[18] På engelsk brukes ofte forkortelsen *CLT*, dvs. "*central limit theorem*".

Bevis:

Vi skal nå ikke bevise CLT, men vi skal vise hvorfor lign.(4.152) stemmer under de forutsetningene som CLT har.

1) **Forventing**:

$$\underline{E[\overline{X}]} = E\left[\frac{X_1 + X_2 + X_3 + \ ... \ + X_n}{n}\right] \tag{4.156}$$

$$= E\left[\frac{1}{n}\Big(X_1 + X_2 + X_3 + \ ... \ + X_n\Big)\right] \tag{4.157}$$

$$= \frac{1}{n}E\Big[X_1 + X_2 + X_3 + \ ... \ + X_n\Big] \tag{4.158}$$

$$= \underline{\frac{1}{n}\Big(E[X_1] + E[X_2] + E[X_3] + \ ... \ + E[X_n]\Big)} \tag{4.159}$$

Men i CLT antar vi at alle stokastiske variablene har samme forventingsverdi, dvs. $E[X_1] = E[X_2] = ... = E[X_n] = \mu$. Dermed:

$$\underline{\underline{E[\overline{X}]}} = \frac{1}{n}\Big(\overbrace{\mu + \mu + \mu + \ ... \ + \mu}^{n \text{ stk.}}\Big) \tag{4.160}$$

$$= \frac{1}{\cancel{n}}\cancel{n}\mu = \underline{\underline{\mu}} \tag{4.161}$$

og vi har vist den første sammenhengen i lign.(4.152).

2) **Varians**:

$$\underline{Var[\overline{X}]} = Var\left[\frac{X_1 + X_2 + X_3 + \ ... \ + X_n}{n}\right] \tag{4.162}$$

$$= Var\left[\frac{1}{n}\Big(X_1 + X_2 + X_3 + \ ... \ + X_n\Big)\right] \tag{4.163}$$

$$= \left(\frac{1}{n}\right)^2 Var\Big[X_1 + X_2 + X_3 + \ ... \ + X_n\Big] \tag{4.164}$$

$$\overset{\text{uavh.}}{=} \left(\frac{1}{n}\right)^2 \underline{\Big(Var[X_1] + Var[X_2] + Var[X_3] + \ ... \ + Var[X_n]\Big)} \tag{4.165}$$

hvor vi i siste overgang i lign.(4.165) har brukt antagelsen om at de stokastiske variablene er uavhengige.
Men i CLT antar vi også at alle stokastiske variablene samme varians,
dvs. $Var[X_1] = Var[X_2] = ... = Var[X_n] = \sigma^2$. Dermed:

$$\underline{\underline{Var[\overline{X}]}} = \left(\frac{1}{n}\right)^2 \Big(\overbrace{\sigma^2 + \sigma^2 + \sigma^2 + \ ... \ + \sigma^2}^{n \text{ stk.}}\Big) \tag{4.166}$$

$$= \frac{1}{n^{\not 2}}\not{n}\sigma^2 = \underline{\underline{\frac{\sigma^2}{n}}} \tag{4.167}$$

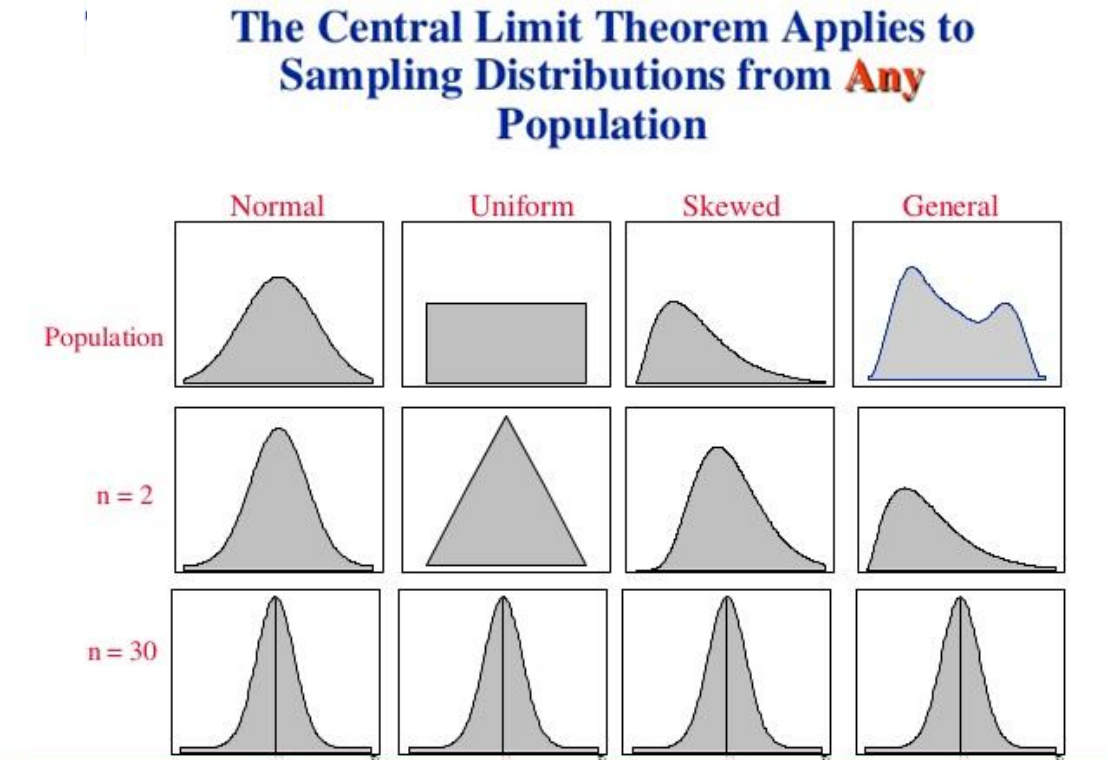
og vi har vist den andre sammenhengen i lign.(4.155).

■

Veldig forenkelt og kort:
Dersom alle stokastiske variabler X_i har samme vilkårlige fordeling P:

$$\overbrace{Y}^{\sim N} = \overbrace{X_1}^{\sim P} + \overbrace{X_2}^{\sim P} + \overbrace{X_3}^{\sim P} + ... + \overbrace{X_n}^{\sim P} \tag{4.168}$$

Visualisering av sentralgrenseteoremet: (n = antall stok. var.)



Figur 4.28: Sentralgrenseteoremet.

På "godt norsk" så betyr dette:

Sannsynlighetsfordelingen til gjennomsnittet av stokastiske variabeler
med **samme** sannsynlighetsfordeling (diskret ELLER kont.) vil, for store n ($n \gtrsim 30$), bli normalfordelt. (4.169)

Kommentar:

CLT sier også at forventingen til *gjennomsnittet* av tenringkast $E[\overline{X}]$ og forventningen til et *gitt* terningkast $E[X_i]$ er det **samme**

$$E[\overline{X}] = E[X_i] \qquad \text{(samme)} \qquad (4.170)$$

men standardavviket til *gjennomsnittet* av terningkast $Var[\overline{X}]$ er **mindre** enn standardavviket til *gjennomsnittet* av terningkast $Var[X_i]$

$$\sigma[\overline{X}] = \frac{\sigma[X_i]}{\sqrt{n}} \qquad \text{(mindre)} \qquad (4.171)$$

Med andre ord:

jo flere forsøk vi beregner gjennomsnittet av, jo mindre blir standardavviket, dvs. jo mindre blir "spredningen".

Kommentar:

Hvor stor n må være (n = antall forsøk) for at sentralgrensesetningen skal gjelde er avhengig av situasjonen. Men en **tommelfingerregel** er at vi bør ha

$$n \gtrsim 30 \tag{4.172}$$

dvs. antall forsøk bør være ca. 30 eller mer.

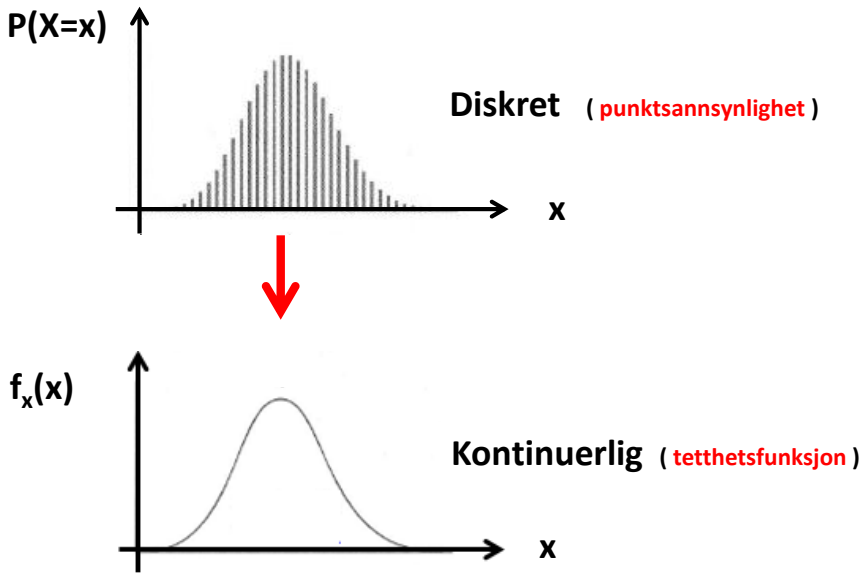
4.5 Diskrete fordelinger $\longrightarrow$ normalfordeling

Under visse betingelser er det slik at diskrete sannsynlighetfordelinger er tilnærmet normalfordelt.

F.eks.:

- binomisk fordeling (diskret)
- hypergeometrisk fordeling (diskret)
- Poissonfordeling (diskret)
- generell diskret fordeling $P(X = x)$

er, under visse betingelser, tilnærmet en $\underbrace{\text{normalfordeling}}_{\text{kontinuerlig}}$. [19]

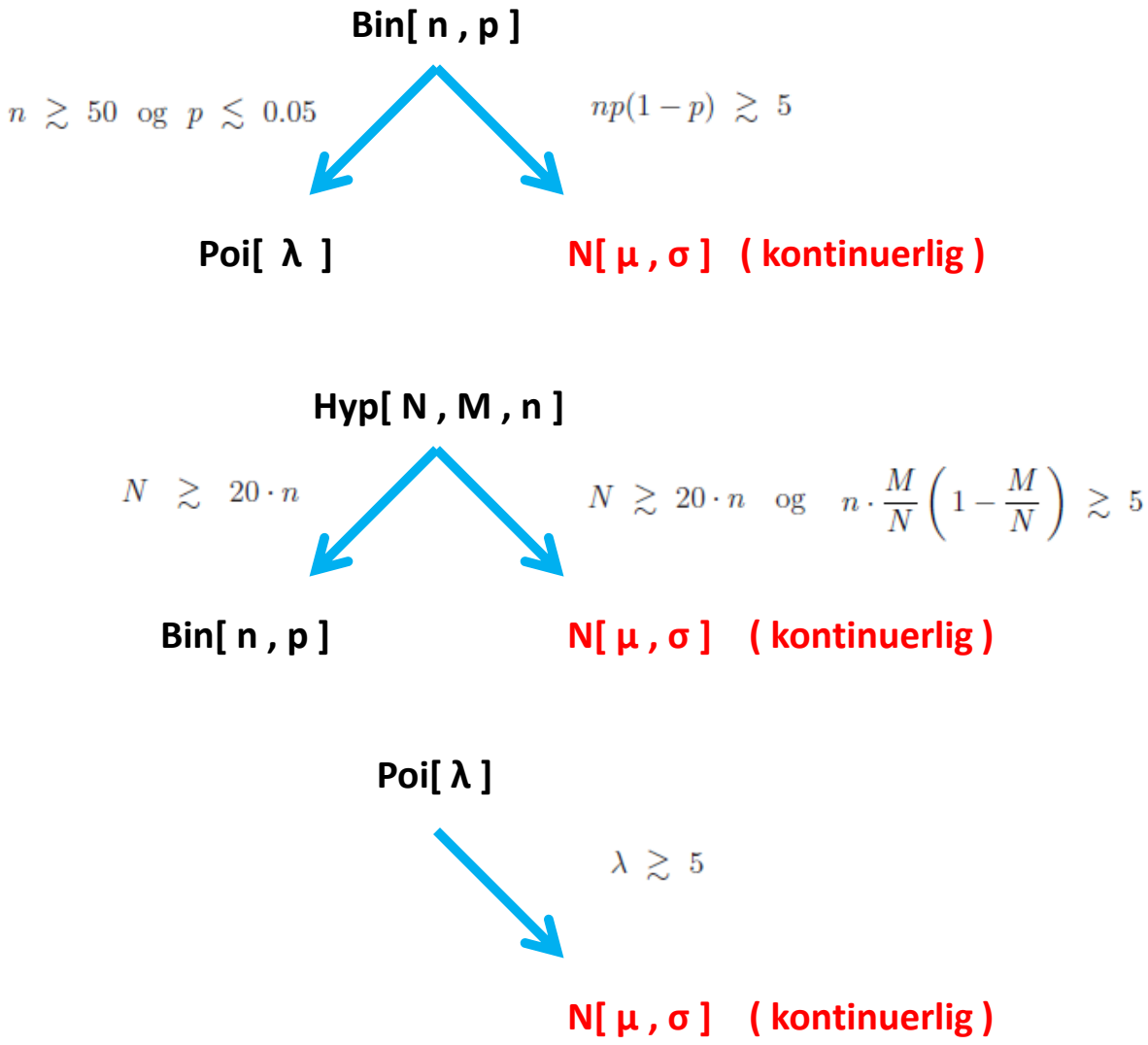


Figur 4.29: Fra diskret til kontinuerlig sannsynlighetsfordeling.

[19] Siden vi ønsker å tilnærme en diskret fordeling med en kontinuerlig normalfordeling så bør vi gjøre en *heltallskorreksjon*. Heltallkorreksjon er ikke pensum i MAT110.

4.5.1 Sammenheng: Bin, Hyp, Poi og N

Under visse betingelser er fordelingene vi har lært om om dette kurset tilnærmet like:



Figur 4.30: Sammenheng: Bin, Hyp, Poi og N.

4.6 Sum av uavhengige stokastiske variabler

Setning: (binomisk fordeling)

Anta at vi har uavhengige og binomisk fordelte stokastiske variabler
$X_1 \sim \text{Bin}[\, n_1, p\,]$, $X_2 \sim \text{Bin}[\, n_2, p\,]$ og $X_3 \sim \text{Bin}[\, n_3, p\,]$.
Da er også summen

$$Y \;=\; X_1 \;+\; X_2 \;+\; X_3 \tag{4.173}$$

binomisk fordelt:

$$\boxed{Y \;\sim\; \text{Bin}\big[\, n_Y \;,\; p\,\big]} \tag{4.174}$$

hvor

$$n_Y \;=\; n_1 + n_2 + n_3 \tag{4.175}$$

■

Du kan lese mer om dette i f.eks. wikipedia.

PS:
Det er ingen enkel eller generell sannsynlighetsfordeling for summen av uavhengige hypergeometriske stokastiske variabler.

Setning: (Poisson fordeling)

Anta at vi har uavhengige og Poisson fordelte stokastiske variabler
$X_1 \sim \text{Poi}[\,\lambda_1\,]$, $X_2 \sim \text{Poi}[\,\lambda_2\,]$ og $X_3 \sim \text{Poi}[\,\lambda_3\,]$
Da er også summen

$$Y \;=\; X_1 \;+\; X_2 \;+\; X_3 \tag{4.176}$$

Poisson fordelt:

$$\boxed{Y \;\sim\; \text{Poi}\big[\,\lambda_Y\,\big]} \tag{4.177}$$

hvor

$$\lambda_Y \;=\; \lambda_1 + \lambda_2 + \lambda_3 \tag{4.178}$$

■

Du kan lese mer om dette i f.eks. wikipedia.

Setning: (normalfordeling)

Anta at vi har uavhengige og normalfordelte stokastiske variabler
$X_1 \sim N[\,\mu_1, \sigma_1^2\,]$, $X_2 \sim N[\,\mu_2, \sigma_2^2\,]$ og $X_3 \sim N[\,\mu_3, \sigma_3^2\,]$
Da er også lineærkombinajonen

$$Y \;=\; aX_1 \;+\; bX_2 \;+\; cX_3 \tag{4.179}$$

normalfordelt:

$$\boxed{Y \;\sim\; N\big[\,\mu_Y \;,\; \sigma_Y^2\,\big]} \tag{4.180}$$

hvor

$$\mu_Y \;=\; a\mu_1 + b\mu_2 + c\mu_3 \tag{4.181}$$

$$\sigma_Y^2 \;=\; a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2 \tag{4.182}$$

■

Du kan lese mer om dette i f.eks. wikipedia.

For oversikten sin del så formulerer vi her sentralgrenseteoremet fra side 231 i samme stil, selv om vi har presentert CLT to ganger tidligere.
Vi formulerer altå CLT på en alternativ og likeverdig måte sammenlignet med lign.(4.151) og (4.154).

Setning: (CLT)

Anta at vi har n antall uavhengige og stokastiske variabler.
Anta videre at disse variablene har samme forventning og samme varians, dvs.

$$E[X_i] = \mu \qquad \text{og} \qquad Var[X_i] = \sigma^2 \quad , \quad i = 1, 2, 3, ..., n \tag{4.183}$$

Da vil også summen

$$Y \;=\; X_1 \;+\; X_2 \;+\; X_3 \;+\; ... \;+\; X_n \tag{4.184}$$

i grensen når $n \to \infty$, være normalfordelt:

$$\boxed{Y \;\sim\; N\big[\,\mu_Y \;,\; \sigma_Y^2\,\big]} \tag{4.185}$$

hvor

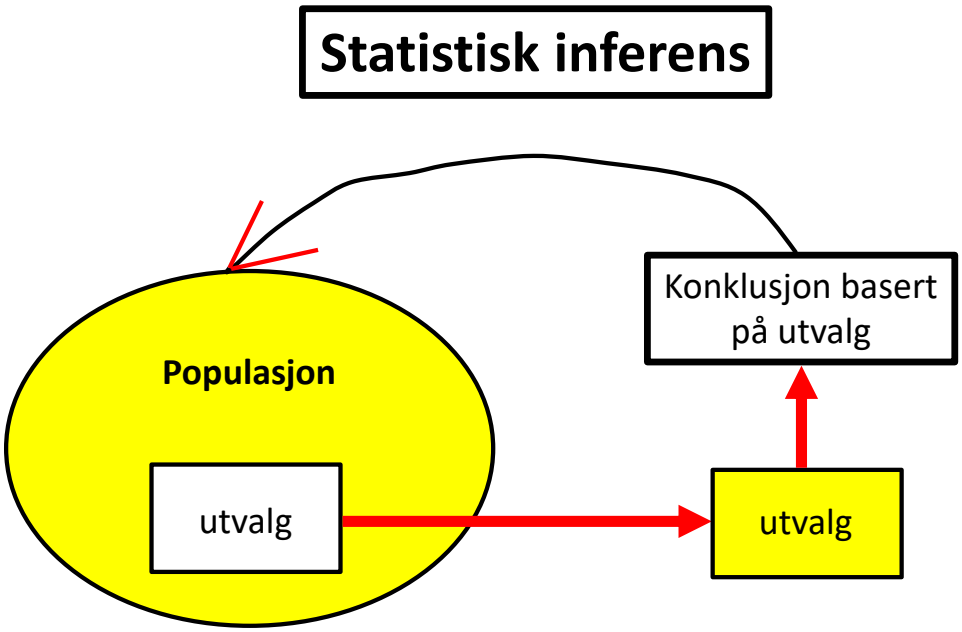
$$E[Y] \;=\; n\,\mu \tag{4.186}$$

$$Var[Y] \;=\; n\,\sigma^2 \tag{4.187}$$

■

Kapittel 5

Statistisk inferens



Figur 5.1: Statistisk inferens.

5.1 Fra sannsynlighetsteori til statistisk inferens

I dialogen mellom vennen til en alvorlig syk pasient og en sykesøster fra side 6, forsøkte vennen å få et estimat fra sykesøsteren på sannsynligheten p for at medisinen pasienten fikk ville fungere.

Definerer stokastisk variabel:

$$X = \text{ja/nei om en tilfeldig valgt pasient, av alle som har den aktuelle sykdommen, blir frisk eller ikke} \tag{5.1}$$

Matematisk kan vi f.eks. definere X slik:

$$X = \begin{cases} 1 & , \overbrace{\text{dersom den tilfeldig valgte pasienten blir frisk}}^{=p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=1-p} \end{cases} \tag{5.2}$$

Ja/nei eksperimenter, dvs. eksperimenter med kun to utfall, kalles et Bernoulli-forsøk. [1]
En tilhørende stokastisk variabel med verdiene 0 og 1 sies å være Bernoulli-fordelt:

$$X \sim \text{Ber}[p] \tag{5.3}$$

hvor

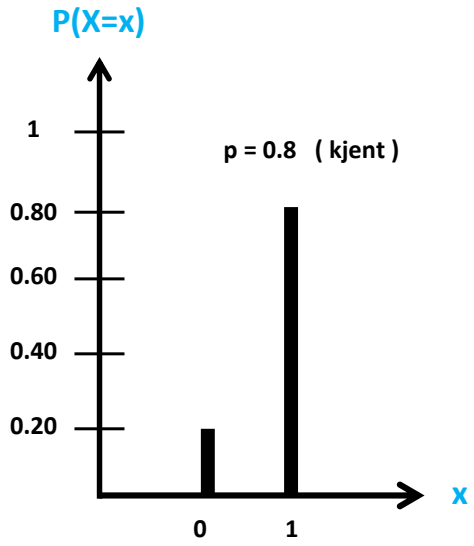
$$p = \underbrace{\text{suksess-sannsynlighet}}_{\text{ukjent } p \text{ samme for alle pasienter}} \tag{5.4}$$

[1] Binomisk eksperiment = n antall Bernoulli-forsøk: $\text{Ber}[p] = \text{Bin}[n = 1, p]$

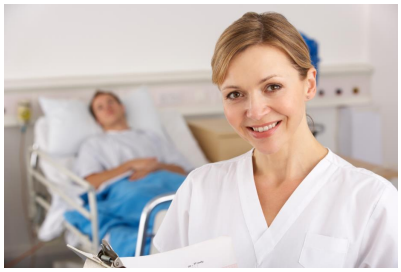
Ideell verden: sannsynlighetsteori (kjenner p)

Sannsynlighetsteori handler om hvilke konklusjoner man kan ta for alle pasienter med den sykdommen, gitt at man ideelt sett kjenner sannsynligheten p.

Anta f.eks. at $p = 0.8$.
Anta videre at vi ikke ser på alle pasienter som har sykdommen, men vi ser kun på f.eks. $n = 100$ tilfeldig utvalgte pasienter. Vi gir medisinen til disse 100 tilfeldig utvalgte pasientene.
Sannsynlighetsteori gir da at det forventes at 80 pasienter blir friske.



Figur 5.2: $X \sim \text{Ber}[\, p = 0.8\,]$.



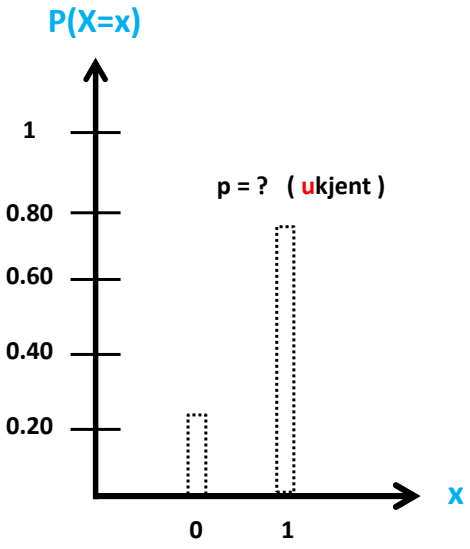
Figur 5.3: Sykesøster.

Problemet med sannsynlighetsteori (ukjent p)

I realiteten kjenner vi ikke sannsynligheten p - eller mer generelt:

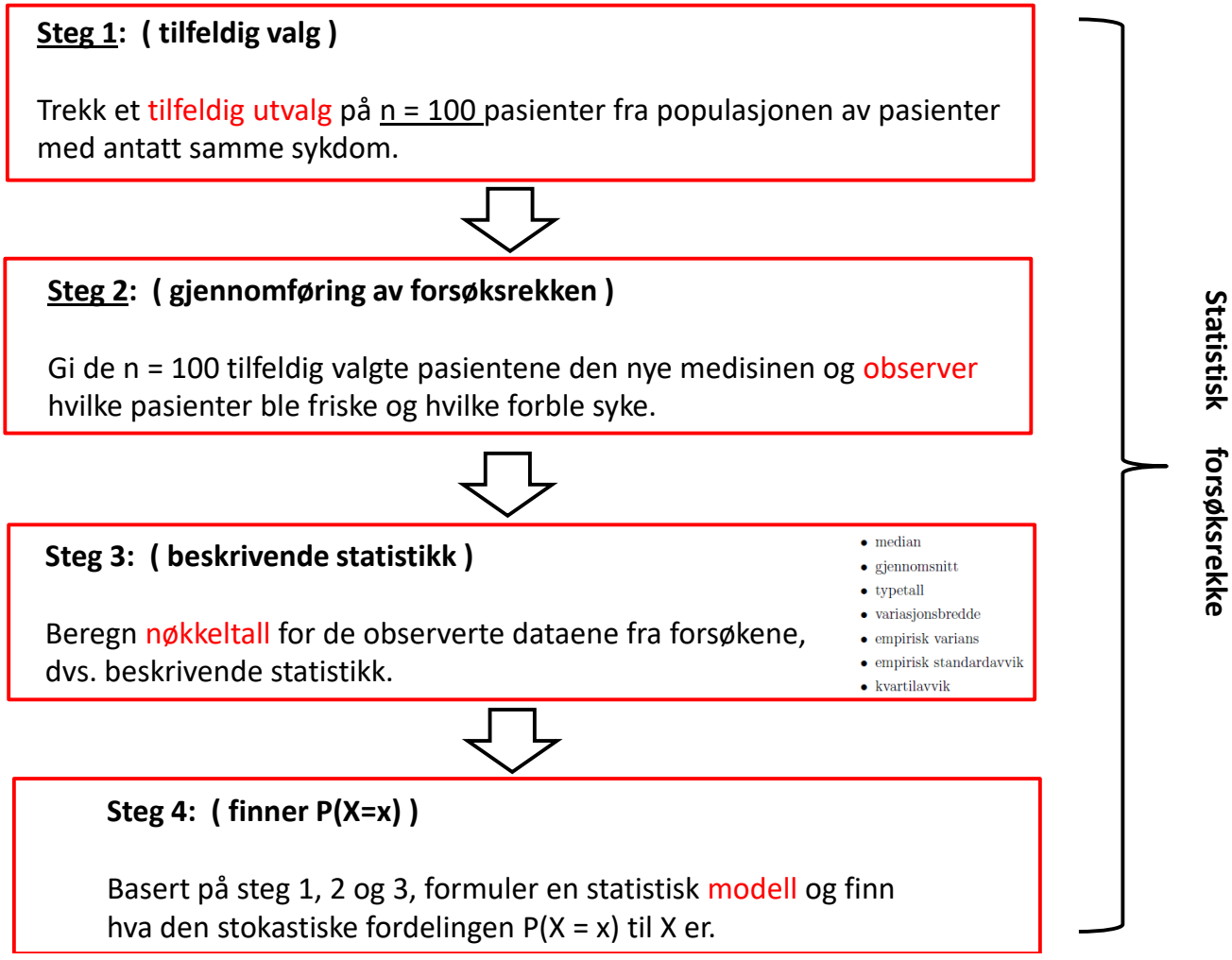
$$P(X = x) \;=\; \text{ukjent} \tag{5.5}$$

Uten sannsynligheten p kan vi ikke trekke konklusjoner om mengden av pasienter som kunne tatt medisinen.



Figur 5.4: $X \sim \text{Ber}[p]$, og p er ukjent.

Løsningen: statistisk forsøksrekke (4 steg)



Figur 5.5: 4 steg.

Definisjon: (populasjon)

populasjon = den totale mengden av objekter/data som vi ønsker å analysere

■

Definisjon: (utvalg)

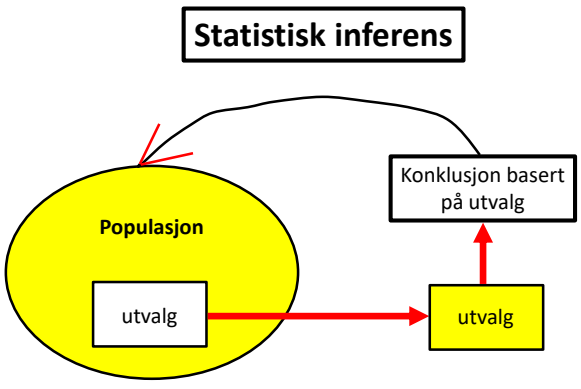
Utvalg = en delmengde av populasjonen,
dvs. en samling av data som er hentet fra en populasjon

■

Definisjon: (statistisk inferens)

Statistisk inferens = det å trekke konklusjoner om populasjonen basert på forsøksrekken som ble gjennomført

■



Figur 5.6: Statistisk inferens.

Eksempel på $\overbrace{\text{statistisk inferens}}^{\text{trekke konklusjoner}}$:

- Kap. 6: Estimering og konfidensintervaller
  Anta $n_1 = 88$ av de $n = 100$ tilfeldig valgte pasientene ble friske. Dersom vi repeterer forsøksrekken med 100 nye tilfeldig valgte pasienter, så kan vi regne på dette og finner at med 95 % *konfidens* [2] konkludere at mellom 69.88 % og 86.11 % av pasientene vil bli friske.

- Kap. 7: Hypotesetesting [3]
  Påstand om populasjonen:
  Legene som har utviklet medisinen har en hypotese, nullhypotese H_0, om at kvinner og menn reagerer likt på medisinen. Legene ønsker å undersøke om de har grunnlag for å forkaste H_0, mens den alternative hypotesen H_1 er at menn og kvinner reagerer forskjellig. Basert på forsøksrekken forkastes H_0 med 95 % sannsynlighet for at vi tar en korrekt konklusjon.

- Kap. 8: Regresjon
  Legene har en teori om at det er en *lineær* sammenheng mellom hvor tung pasienten er, X_1, og minimum dose pasienten bør ha for å bli frisk, X_2. Basert på forsøksrekken, kan legene predikere med 95 % sannsynlighet at dersom du er mann og veier 100 kilo, trenger du en dose på minst 500 mg.

[2] konfidens = sikkerhet/sannsynlighet

[3] Hypotese er en gjetning, antagelse eller forklaring som synes rimelig ut fra foreliggende kunnskap, og som man forsøker å avkrefte eller bekrefte.

5.2 Steg 1: Tilfeldig utvalg

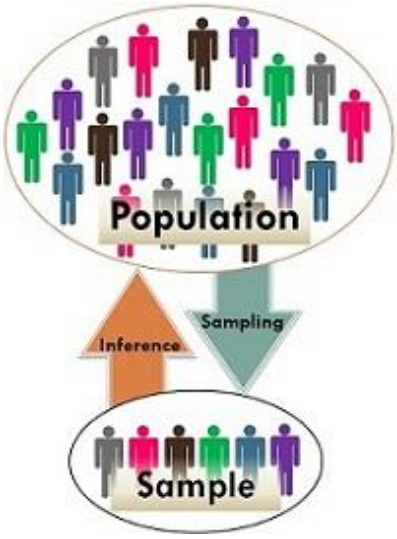
Målet er å kunne trekke konklusjoner om hvorvidt et nytt legemiddel fungerer eller ikke, basert på en statistisk forsøksrekke, dvs. vi ønsker å bruke statistisk inferens for å si noe om legemiddelets effekt.

Første steg i en statistisk forsøksrekke er å velge et *tilfeldig* utvalg fra populasjonen vi ønsker å studere. Men før vi trekker et tilfeldig utvalg, må vi ha en *populasjon* å trekke fra:

Eksempel: (populasjon)

populasjon = alle personer som har fått diagnosen til den gjeldende sykdommen

■



Figur 5.7: Populasjon.

En ideell strategi hadde vært å testet alle personene i populasjonen, men dette er som regel ikke gjennomførbart - pga. både praktiske og økonomiske årsaker. Dermed må vi ty til kun et utvalg av populasjonen, som skal gi oss grunnlaget for å si noe om den totale populasjonen.

I forrige avsnitt definerte vi at et utvalg kun er en delmengde av populasjonen, men hva mener vi med et *tilfeldig* utvalg?

Definisjon: (tilfeldig utvalg)

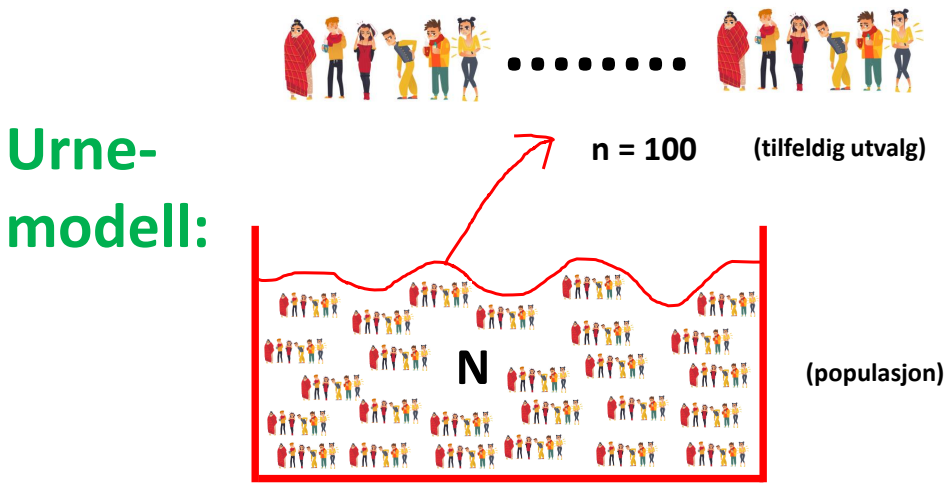
tilfeldig utvalg = et utvalg fra en populasjon som er valgt slik at det er lik sannsynlighet for at hvert enkelt objekt i populasjonen blir trekt, *uavhengig* av de andre objektene som har blitt trekt

■

Eksempel: (tilfeldig utvalg - legemiddel)

En analogi til tilfeldig utvalg er *urnemodellen*: [4]

Anta at individene i populasjonen representerer kulene i urnemodellen.
Vi trekker $n = 100$ tilfeldige pasienter.
Det er lik sannsynlighet for å trekke kulene/pasientene i urnen. Vi trekker uten tilbakelegging.



Figur 5.8: Urne.

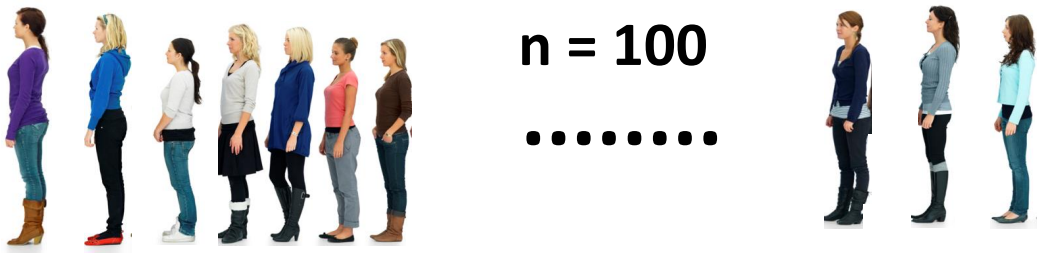
[4]Husk at urnemodellen forutsetter lik sannsynlighet for trekke de ulike kulene. Man kan si det forutsetter at kulene er "like store".

Forskjellen mellom et utvalg og et *tilfeldig* utvalg er svært essensielt å forstå.
Følgende eksempel belyser dette:

Eksempel: (ikke tilfeldig utvalg - legemiddel)

1. Anta vi trekte $n = 100$ pasienter som alle var kvinner. Vi har fremdeles et utvalg, men utvalget er *ikke* tilfeldig. Vi har bevisst sett vekk ifra alle menn fra populasjonen.

2. Anta at vi for hver trekte pasient inkluderte eventuelle slektninger som også hadde den samme sykdommen. Vi ville da fått en *avhengighet* mellom forsøkene som dermed ville kunne ledet til feilaktige konklusjoner for den *totale* populasjonen. Utvalget er dermed ikke et *tilfeldig* utvalg.



Figur 5.9: Trekker $n = 100$ pasienter som alle er kvinner. Ikke tilfeldig.

■

Eksempel: (tilfeldig utvalg - legemiddel)

Anta vi har gjennomført en

$$\text{tilfeldig trekning av } n = 100 \text{ pasienter}$$

fra listen over alle pasienter med gjeldende diagnose (populasjonen).
Resultatet finner du i tabell 5.1, hvor det ble trekt:

$$48 \text{ kvinner} \quad , \quad 52 \text{ menn} \tag{5.6}$$

Kvinnene er nummerert i rødt, mens mennene er nummerert i blått.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Tabell 5.1: Et tilfeldig utvalg av $n = 100$ pasienter - kvinner og menn.

■

Viktige konsekvenser: (av tilfeldig utvalg)

1. Representativ:

   Hensikten med å ha et *tilfeldig* utvalg er at forsøksrekken skal være *representativ* for populasjonen. Med et utvalg som ikke er tilfeldig, kan feil konklusjoner trekkes om populasjonen.

2. Gjensidig uavhengighet:

   Anta at vi har trukket et tilfeldig utvalg på $n = 100$ pasienter. Siden vi har gjort et *tilfeldig* utvalg, kan vi i tillegg anta at alle disse 100 variablene er *gjensidig uavhengige*, dvs.:

$$P(X_i = x_i | X_j = x_j) \;=\; P(X_i = x_i) \qquad\qquad i \neq j = 1, 2, 3, \ldots, 100 \qquad\qquad (5.7)$$

   Lign.(5.7) sier på godt norsk at resultatet fra forsøk nr. j ikke betyr noe for resultatet for *alle* de andre forsøkene.

■

5.2.1 Populasjonsvariabler og forsøksvariabler

Stokastisk variabeler:

$$X \overset{\text{lign.}(5.1)}{=} \begin{array}{l}\text{ja/nei om en tilfeldig valgt pasient,}\\ \text{av alle som har den aktuelle sykdommen,}\\ \text{blir frisk eller ikke}\end{array} \tag{5.8}$$

$$X_i = \text{om pasient nr. } i \text{ blir frisk eller ikke} \tag{5.9}$$

hvor

$$\text{populasjonsvariabel } X = \text{variabel for alle som har den aktuelle sykdommen} \tag{5.10}$$

$$\text{forsøksvariabel } X_i = \text{variabel for kun de } n = 100 \text{ tilfeldig valgte pasientene} \tag{5.11}$$

Forsøksvariablene $X_1, X_2, \ldots, X_{100}$ er uavhengige kopier av variabelen X i lign.(5.8), som representerer den totale populasjonen. Vi sier at variablene er i.i.d. eller *uavhengige identiske fordelte variabler* på godt norsk: [5] [6]

Definisjon: (i.i.d. - uavhengige identisk fordelte variabler)

Et sett med variabler $X_1, X_2, \ldots, X_n$ er $\overbrace{\text{uavhengig identisk fordelt}}^{\text{i.i.d.}}$, dersom:

1. X_i-ene har samme (= identisk) sannsynlighetsfordeling [7]
2. X_i-ene er gjensidig uavhengige

■

Figur 5.10: Uavhengighet.

[5] i.i.d. = independent identical distributed

[6] Antagelsen om i.i.d. ved statistiske forsøksrekker er helt fundamental for statistisk inferens og ligger dermed sentralt i bunn for alle de resterende kapitlene i faget (kap.6, 7 og 8).

[7] Dvs. samme $P(X = x)$ (diskrete variabler) eller $f_X(x)$ (kontinuerlige variabler).

Matematisk kan disse stokastiske variablene formuleres slik:

$$X_i = \begin{cases} 1 & , \overbrace{\text{dersom pasient nr. } i \text{ blir frisk}}^{=p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=1-p} \end{cases} \tag{5.12}$$

$$X \overset{\text{lign.(5.2)}}{=} \begin{cases} 1 & , \overbrace{\begin{array}{l}\text{dersom en tilfeldig valgt pasient} \\ \text{blant alle som har den aktuelle sykdommen,} \\ \text{blir frisk}\end{array}}^{=p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=1-p} \end{cases} \tag{5.13}$$

hvor

$$p = \underbrace{\text{suksess-sannsynlighet}}_{\underline{\text{u}}\text{kjent } p \text{ samme for alle pasienter}} \tag{5.14}$$

altså både X_i og X er Bernoulli-fordelt: [8]

$$X_1, X_2, X_3, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p] \tag{5.15}$$

hvor $n = 100$.

[8] Legg merke til at sannsynligheten for å bli frisk *ikke* er avhengig av i. Det betyr at vi antar at p er den samme for alle pasienter.

5.3 Steg 2: Gjennomføring av forsøksrekken

a) **Resultat**: (effekten Y_i - et tall mellom 0 og 1)

Anta at effekten av legemiddelet som pasientene tok i forsøksrekken kan måles fra skala fra 0 til 1, hvor 1 betyr helt frisk mens 0.5 betyr 50 % frisk. [9]

$$Y_i = \underbrace{\text{effekten av legemiddelet}}_{\text{tall mellom 0 og 1, se tabell 5.2}} \text{ for forsøkspasient nr. } i \qquad (5.16)$$

Tabell 5.2 oppsummerer målingene fra forsøkene: ($\overbrace{\text{kvinner}}^{48}$ er rød , $\overbrace{\text{menn}}^{52}$ er blå)

0.91	0.81	0.93	0.83	0.92	0.81	0.88	0.93	0.87	0.89
0.90	0.84	0.92	0.90	0.91	0.90	0.84	0.89	0.90	0.94
0.79	0.89	0.88	0.91	0.88	0.87	0.86	0.88	0.92	0.88
0.86	0.90	0.86	0.83	0.87	0.89	0.86	0.89	0.94	0.93
0.92	0.87	0.84	0.78	0.94	0.80	0.84	0.88	0.87	0.88
0.83	0.88	0.90	0.99	0.95	0.94	0.89	0.91	0.90	0.89
0.92	0.95	0.90	0.91	0.86	0.93	0.88	0.94	0.93	0.89
0.90	0.94	0.89	0.91	0.94	0.91	0.98	0.88	0.99	0.90
0.93	0.91	0.90	0.91	0.89	0.95	0.98	0.90	0.90	0.95
0.90	0.92	0.85	0.92	0.96	0.90	0.86	0.92	0.93	0.92

Tabell 5.2: Y_i - effekten.

[9] Når man gjennomfører forsøkene, er det vesentlig at forholdene er så like som mulig rundt hvert forsøk. Hver enkelt pasient skal i prinsippet ha de samme vilkårene for å bli frisk som de andre. Dette bygger opp under antakelsen om at alle pasientene i populasjonen har lik sannsynlighet for å bli frisk. Doseringen må f.eks. være den samme, for hver pasient som testes.

b) **Resultat**: (frisk/ikke fris X_i - 0 eller 1)

Vi gir pasienten legemiddelet og venter i 2 uker. Deretter måles i hvilken grad pasienten ble frisk i form av et tall mellom 0 og 1, hvor 0 er ingen bedring, mens 1 er fullstendig frisk. Dersom en pasient måler mer enn 0.85, anses forsøket som en *suksess*.

Av de totalt $n = 100$ pasientene var det $\overbrace{88 \text{ pasienter}}^{\text{frisk}}$ som skåret 0.85 eller høyere, dvs. suksess.

$$\geq \ 0.85 \quad \Rightarrow \quad \text{frisk} \tag{5.17}$$
$$< \ 0.85 \quad \Rightarrow \quad \text{ikke frisk} \tag{5.18}$$

I lign.(5.9) på side 266 definerte vi X_i variabelen - frisk/ikke frisk:

$$X_i = \begin{cases} 1 & , \text{ dersom pasient nr. } i \text{ blir frisk, dvs. } y_i \geq 0.85 \\ 0 & , \text{ hvis ikke, dvs. } y_i < 0.85 \end{cases} \tag{5.19}$$

Konklusjon fra tabell 5.2: [10]

1	0	1	0	1	0	1	1	1	1
1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1
1	1	0	0	1	0	0	1	1	1
0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

Tabell 5.3: X_i - frisk eller ikke frisk.

[10] 88 pasienter blir friske (45 kvinner og 43 menn), 12 blir ikke friske.

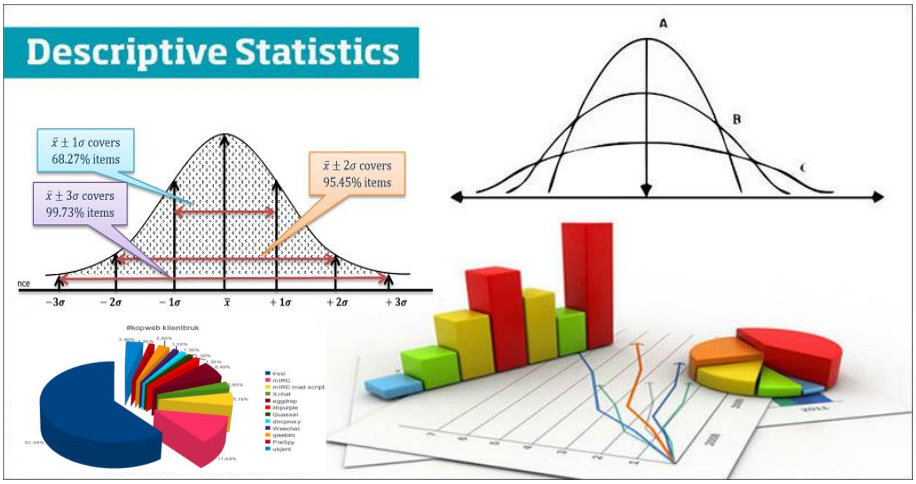
5.4 Steg 3 : Beskrivende statistikk

Definisjon: (beskrivende statistikk)

La $X_1, X_2, \ldots, X_n$ være en vilkårlig stokastisk variabel med tilhørende observasjoner $x_1, x_2, \ldots, x_n$.

beskrivende statistikk = størrelser som beskriver nøkkeltall for et sett med observasjoner $\underbrace{x_1, x_2, \ldots, x_n}_{\text{tall}}$.

■



Figur 5.11: Nøkkeltall.

Beskrivende statistikk måler ulike egenskaper ved observasjonene som f.eks.

- lokaliseringmål
- spredingsmål

Tabell 5.4 oppsummerer noen vanlige størrelser som beskriver nøkkeltall.

størrelser	**type**
median	lokaliseringsmål (midtpunkt)
gjennomsnitt	lokaliseringsmål (midtpunkt)
typetall	lokaliseringsmål (frekvens)
empirisk varians	spredningsmål
empirisk standardavvik	spredningsmål
variasjonsbredde	spredningsmål
kvartilavvik	spredningsmål

Tabell 5.4: Nøkkeltall.

Beskrivende statistikk sier noe om hvordan den (ukjente) sannsynlighetsfunksjonen $\overbrace{\overbrace{P(X=x)}^{\text{(diskret)}} \text{ eller } \overbrace{f_X(x)}^{\text{kontinuerlig}}}$ til X "ser ut". Tabell 5.4 viser noen størrelser som beskriver nøkkeltall for en vilkårlig sannsynlighetsfordeling.

5.4.1 Lokaliseringsmål

Definisjon: (median)

La n være en serie med tall/observasjoner i ordnet rekkefølge. Da er:

$$\text{median} = \begin{cases} \text{midtre observasjonen} & , \quad n = \text{odde} \\ \text{gjennomsnitt av to midterste observasjonene} & , \quad n = \text{like} \end{cases} \tag{5.20}$$

■

Definisjon: (gjennomsnitt)

La $\overbrace{x_1, x_2, x_3, ..., x_n}^{\text{tall}}$ være n antall observasjoner. Da er gjennomsnittet: [11]

$$\overline{x} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.21}$$

■

Definisjon: (typetall) [12]

La n være en serie med tall/observasjoner i ordnet rekkefølge. Da er:

$$\text{typetall} \quad = \quad \text{den verdien som forekommer hyppigst} \tag{5.22}$$

■

[11] Σ = den greske bokstaven "*sigma*". F.eks. $\Sigma_{i=1}^{3} x_i \;=\; x_1 + x_2 + x_3$.
[12] Kalles også modus eller modalverdi.

5.4.2 Spredingsmål

Definisjon: (empirisk varians) [13]

La $\overbrace{x_1, x_2, x_3, ..., x_n}^{\text{tall}}$ være observasjoner, og la $\overline{x}$ være gjennomsnittet.
Da er den empiriske variansen: [14]

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (5.23)$$

■

Definisjon: (empirisk standardavvik) [15]

Det empiriske standardavviket er:

$$s_x = \sqrt{s_x^2} \qquad (5.24)$$

■

[13] Kalles også **utvalgsvariansen**.

[14] Ulike estimater av variansen: I lign.(5.23) deler man på $n-1$, og ikke n. Om vi bruker det ene eller det andre er avhengig om $\overline{x}$ er gjennomsnittet for hele populajonen, eller bare et utvalg. I dette kurset skal vi imidlertid holde oss til definisjonen i lign.(5.23).

[15] Kalles også **utvalgstandardavviket**.

Definisjon: (variasjonsbredde)

La n være en serie med tall/observasjoner i ordnet rekkefølge. Da er:

$$\boxed{\text{variasjonsbredde} \;=\; \text{differansen mellom største og minste verdi}} \tag{5.25}$$

■

Definisjon: (kvartilavvik)

La n være en serie med tall/observasjoner i ordnet rekkefølge. Da er:

$$k_1 \;=\; \text{nedre kvartil , dvs.} \quad 25\% \text{ av observasjonene har verdi } \leq k_1 \tag{5.26}$$

$$k_2 \;=\; \text{medianen} \quad \text{, dvs.} \begin{cases} 50\% \text{ av observasjonene har verdi } \leq k_2 \\ 50\% \text{ av observasjonene har verdi } \geq k_2 \end{cases} \tag{5.27}$$

$$k_3 \;=\; \text{øvre kvartil} \quad \text{, dvs.} \quad 75\% \text{ av observasjonene har verdi } \leq k_3 \tag{5.28}$$

Da er

$$\boxed{\text{kvartilavvik} \;=\; k_3 \;-\; k_1} \tag{5.29}$$

■

Eksempel: (nøkkeltall for Y_i - beskrivende statistikk)

Tabell 5.5 viser nøkkeltall for målingene av effekten av legemiddelet tatt i forsøksrekken med $n = 100$ pasienter, dvs. for dataene i tabell 5.2.

størrelser	**nøkkeltall**
Median	0.90
Gjennomsnitt $\overline{y}$	0.8967
Typetall	0.90
Empirisk varians s_y^2	0.0017
Empirisk standardavvik s_y	0.0413
Variasjonsbredde	0.21
Kvartilavvik	0.04

Tabell 5.5: Nøkkeltall Y_i fra tabell 5.2.

a) Vis hvordan man regner ut nøkkeltallene i tabell 5.5.
Bruk dataene i tabell 5.2.

b) Lag et stolpediagram for dataene i tabell 5.2.

c) Marker nøkkeltallene fra tabell 5.5 inn i diagrammet.

Løsning:

a) Median: (midterste observasjon , $n = 100$ like antall observasjoner)

$$\underline{\underline{\text{median}}} \overset{\text{lign.(5.20)}}{=} \frac{0.90 + 0.90}{2} = \underline{\underline{0.90}} \tag{5.30}$$

Gjennomsnitt: (se tabell 5.2)

$$\underline{\underline{\overline{y}}} \overset{\text{lign.(5.21)}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{0.91 + 0.81 + \ \ldots\ 0.92}{100} = \underline{\underline{0.8967}} \tag{5.31}$$

Typetall: [16]

$$\underline{\underline{\text{typetall}}} \overset{\text{lign.(5.22)}}{=} \text{den verdien som forkommer hyppigst} = \underline{\underline{0.90}} \tag{5.32}$$

[16] Å regne ut typetall kan man se direkte fra tabell 5.2. Men man kan også bruke det data program, f.eks. Excel, til å regne ut typetallet.

Empirisk varians:

$$\underline{\underline{s_y^2}} \overset{\text{lign.(5.23)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \underline{\underline{0.0017}} \tag{5.33}$$

Empirisk standardavvik:

$$\underline{\underline{s_y}} \overset{\text{lign.(5.24)}}{=} \sqrt{s_y^2} = \underline{\underline{0.0413}} \tag{5.34}$$

Variasjonsbredde: [17]

$$\begin{aligned} \underline{\text{variansjonsbredde}} &\overset{\text{lign.(5.25)}}{=} \text{største verdi} - \text{minste verdi} \\ &= 0.99 - 0.78 = \underline{\underline{0.21}} \end{aligned} \tag{5.35}$$

Kvartilavvik: [18]

$$\begin{aligned} \underline{\text{kvartilavvik}} &\overset{\text{lign.(5.29)}}{=} k_3 - k_1 \\ &= 0.92 - 0.88 = \underline{\underline{0.04}} \end{aligned} \tag{5.36}$$

[17] Via f.eks. Excel er det lett å finne største og minste verdi av resultatet fra tabell 5.2 (bare ved å sortere etter stigende eller synkende verdi). Det er mer arbeid å finne største og minste verdi "for hånd".

[18] Via f.eks. Excel finner man kvartilavviket. Det er mye arbeid å gjøre det "for hånd".

b) Vi gjør samme tilnærming som vi gjorde i eksemplet med Hustadmarmor på side 191
La oss dele inn i et hensiktsmessig antall intervall, f.eks.:

$$\begin{aligned} &0.750 - 0.775 \quad , \quad 0.775 - 0.800 \quad , \quad 0.800 - 0.825 \quad , \quad 0.825 - 0.850 \\ &0.850 - 0.875 \quad , \quad 0.875 - 0.900 \quad , \quad 0.900 - 0.925 \quad , \quad 0.825 - 0.950 \\ &0.950 - 1.000 \end{aligned} \tag{5.37}$$

Relativ frekvensene for disse intervallene er: ($n = 100$)

$$\underline{f_r(n_1)} = \frac{n_1}{n} = \frac{0}{100} = \underline{0} \tag{5.38}$$

$$\underline{f_r(n_2)} = \frac{n_2}{n} = \frac{0}{100} = \underline{0} \tag{5.39}$$

$$\underline{f_r(n_3)} = \frac{n_3}{n} = \frac{3}{100} = \underline{0.03} \tag{5.40}$$

$$\underline{f_r(n_4)} = \frac{n_4}{n} = \frac{2}{100} = \underline{0.02} \tag{5.41}$$

$$\underline{f_r(n_5)} = \frac{n_5}{n} = \frac{2}{100} = \underline{0.07} \tag{5.42}$$

$$\underline{f_r(n_6)} = \frac{n_6}{n} = \frac{12}{100} = \underline{0.12} \tag{5.43}$$

$$\underline{f_r(n_7)} = \frac{n_7}{n} = \frac{28}{100} = \underline{0.28} \tag{5.44}$$

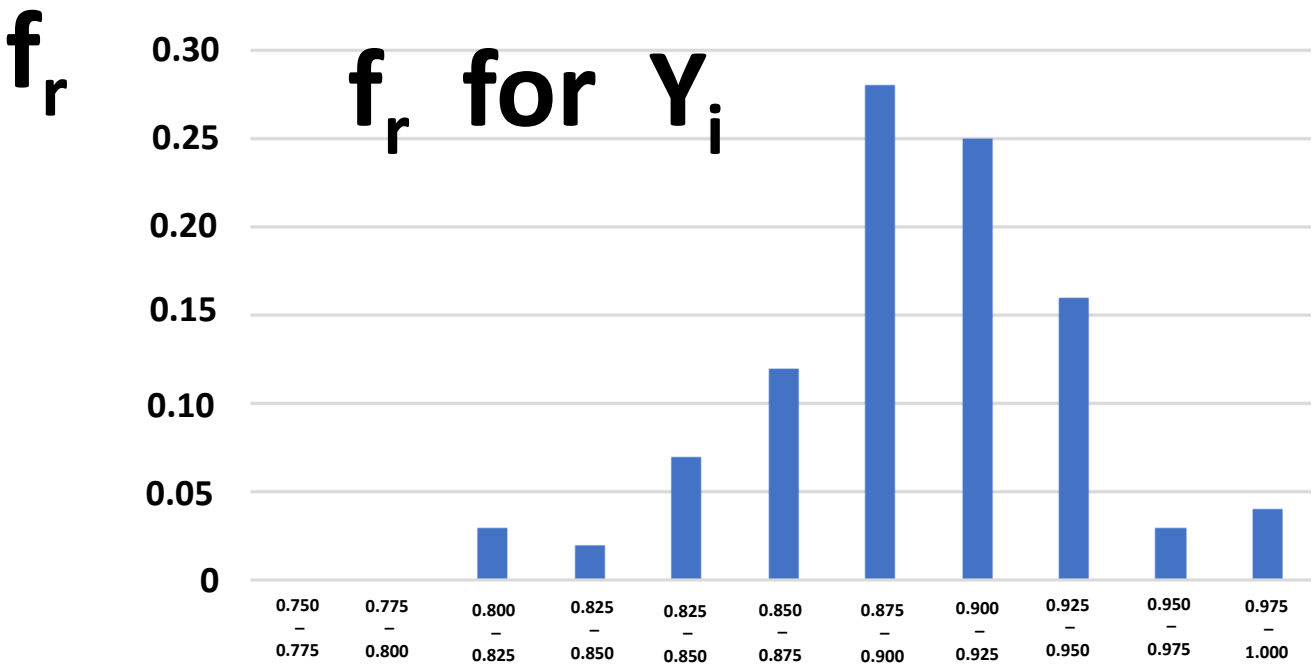
$$\underline{f_r(n_8)} = \frac{n_8}{n} = \frac{25}{100} = \underline{0.25} \tag{5.45}$$

$$\underline{f_r(n_9)} = \frac{n_9}{n} = \frac{16}{100} = \underline{0.16} \tag{5.46}$$

$$\underline{f_r(n_{10})} = \frac{n_{10}}{n} = \frac{3}{100} = \underline{0.03} \tag{5.47}$$

$$\underline{f_r(n_{11})} = \frac{n_{11}}{n} = \frac{4}{100} = \underline{0.04} \tag{5.48}$$

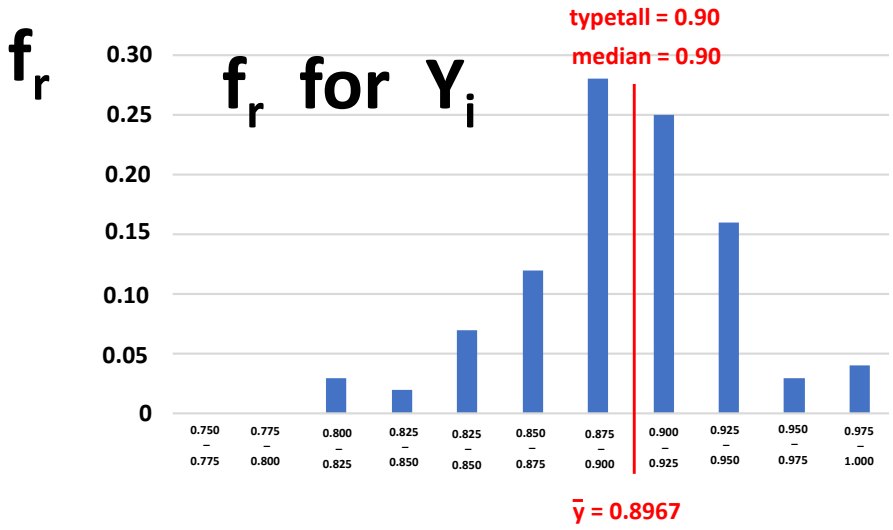
Stolpediagram:



Figur 5.12: Relativ frekvenser f_r for Y_i, se lign.(5.38)-(5.48).

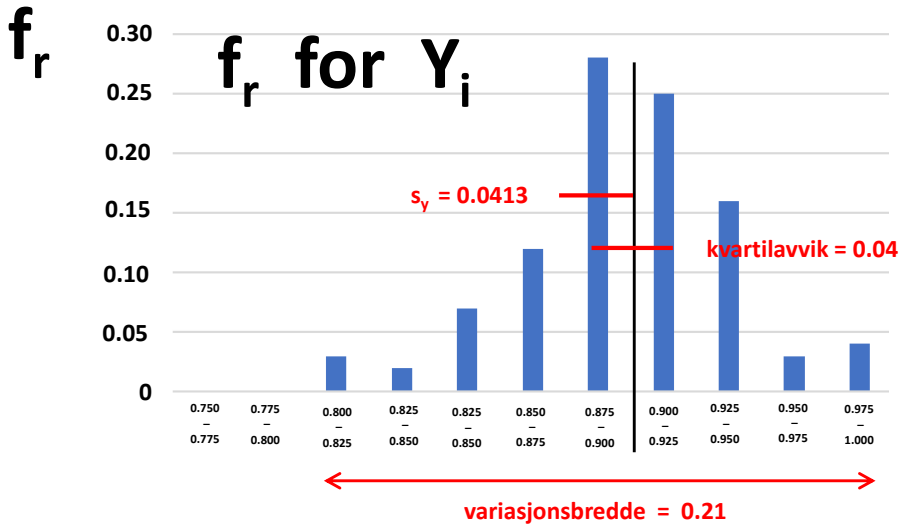
Ser du hvilken fordeling dette ligner på?

c) Stolpediagram med lokaliseringsmål:



Figur 5.13: Relativ frekvenser f_r, se lign.(5.38)-(5.48).

Stolpediagram med spredingsmål: [19]



Figur 5.14: Relativ frekvenser f_r, se lign.(5.38)-(5.48).

■

[19] Empirisk varians s_y^2 har ingen direkte geometrisk tolkning slik som empirisk standardavvik s_y har. Derfor er ikke s_y^2 markert i stolpediagrammet.

Eksempel: (nøkkeltall for X_i - beskrivende statistikk)

Tabell 5.6 viser nøkkeltall for målingene av $n = 100$ pasienter om de blir friske eller ikke, dvs. nøkkeltall for dataene i tabell 5.3.

størrelser	**nøkkeltall**
Median	1
Gjennomsnitt $\overline{x}$	0.88
Typetall	1
Empirisk varians s_x^2	0.1067
Empirisk standardavvik s_x	0.3266
Variasjonsbredde	1
Kvartilavvik	0

Tabell 5.6: Nøkkeltall X_i fra tabell 5.3.

a) Vis hvordan man regner ut nøkkeltallene i tabell 5.6.
Bruk dataene i tabell 5.2.

b) Lag et stolpediagram for dataene i tabell 5.3.

c) Marker nøkkeltallene fra tabell 5.6 inn i diagrammet.

Løsning:

a) Median: (midterste observasjon , $n = 100$ like antall observasjoner)

$$\underline{\underline{\text{median}}} \overset{\text{lign.(5.20)}}{=} \frac{1+1}{2} = \underline{\underline{1}} \tag{5.49}$$

Gjennomsnitt: (se tabell 5.3)

$$\underline{\underline{\overline{x}}} \overset{\text{lign.(5.21)}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\overbrace{1+1+1+...+1}^{88} + \overbrace{0+0+..+0}^{12}}{100} = \underline{\underline{0.88}} \tag{5.50}$$

Typetall: [20]

$$\underline{\underline{\text{typetall}}} \overset{\text{lign.(5.22)}}{=} \overbrace{\text{den verdien som forkommer hyppigst}}^{\text{88 1'erene og 12 0'ere}} = \underline{\underline{1}} \tag{5.51}$$

[20] Å regne ut typetall kan man se direkte fra tabell 5.2. Men man kan også bruke det data program, f.eks. Excel, til å regne ut typetallet.

Empirisk varians:

$$\underline{\underline{s_x^2}} \overset{\text{lign.(5.23)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \tag{5.52}$$

$$= \frac{(1-0.88)^2+\ldots+(1-0.88)^2}{100-1} = \underline{\underline{0.1067}} \tag{5.53}$$

Empirisk standardavvik:

$$\underline{\underline{s_x}} \overset{\text{lign.(5.24)}}{=} \sqrt{s_x^2} = \sqrt{0.1067} = \underline{\underline{0.3266}} \tag{5.54}$$

Variasjonsbredde:

$$\underline{\underline{\text{variansjonsbredde}}} \overset{\text{lign.(5.25)}}{=} \text{største verdi} - \text{minste verdi}$$

$$= 1-0 = \underline{\underline{1}} \tag{5.55}$$

Kvartilavvik: [21]

$$\underline{\underline{\text{kvartilavvik}}} \overset{\text{lign.(5.29)}}{=} k_3 - k_1$$

$$= 1-1 = \underline{\underline{0}} \tag{5.56}$$

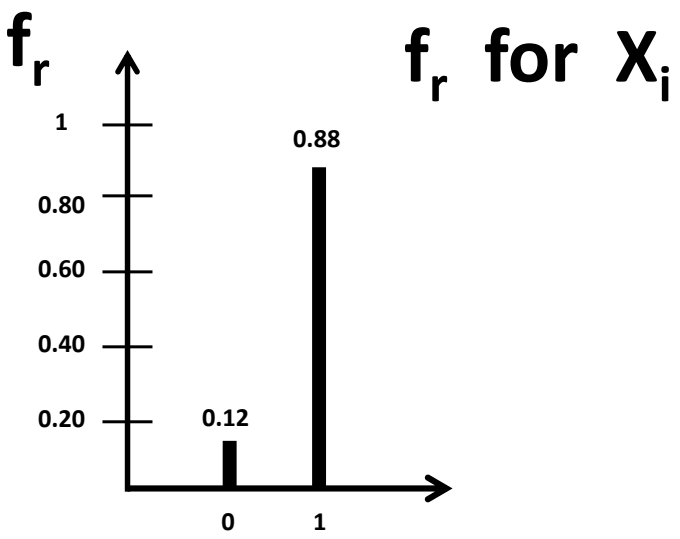
[21] Via f.eks. Excel finner man kvartilavviket. Det er mye arbeid å gjøre det "for hånd".

b) Variabelen X_i har kun to verdier, 0 og 1.
Relativ frekvensen for disse verdiene er: ($n = 100$)

$$\underline{f_r(n_0)} = \frac{n_0}{n} = \frac{12}{100} = \underline{0.12} \tag{5.57}$$

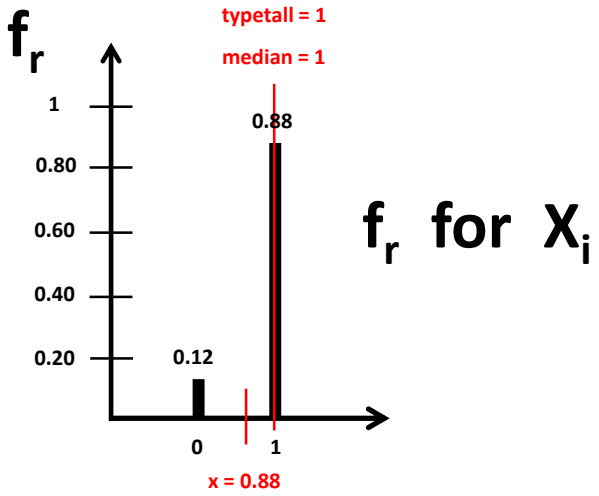
$$\underline{f_r(n_1)} = \frac{n_1}{n} = \frac{88}{100} = \underline{0.88} \tag{5.58}$$

Stolpediagram:



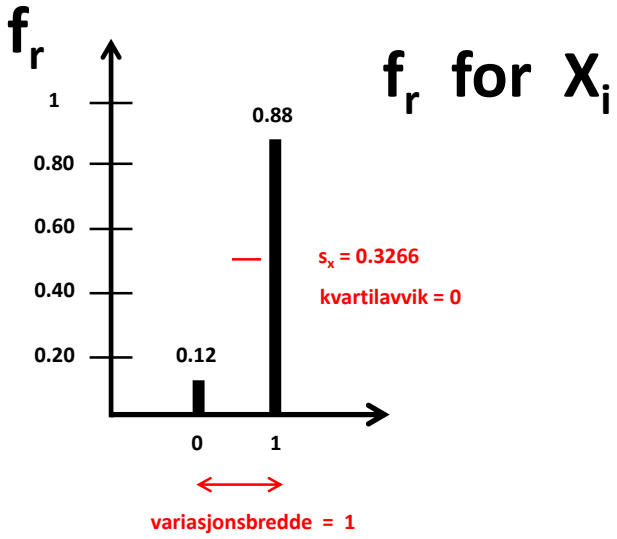
Figur 5.15: Relativ frekvenser $f_r(n_j)$ for X_i, se lign.(5.57)-(5.58).

c) Stolpediagram med lokaliseringsmål:



Figur 5.16: Relativ frekvenser f_r, se lign.(5.38)-(5.48).

Stolpediagram med spredingsmål: [22]



Figur 5.17: Relativ frekvenser f_r, se lign.(5.38)-(5.48).

■

[22] Empirisk varians s_x^2 har ingen direkte geometrisk tolkning slik som empirisk standardavvik s_x har. Derfor er ikke s_x^2 markert i stolpediagrammet.

5.5 Statistisk modell

Siste steg før vi er klare for å $\overbrace{\text{trekke konklusjoner om populasjonen på bakgrunn av forsøksrekken}}^{\text{statistisk inferens}}$, er å formulere en [23]

modell

om hvilken *sannsynlighetsfordeling* som beskriver det stokastiske aspektet ved forsøksrekken (og dermed også populasjonen).

En slik hypotese kalles en *statistisk modell* til forsøksrekken:

Figur 5.18: Statistisk modell.

[23] Hypotese er en gjetning, antagelse eller forklaring som synes rimelig ut fra foreliggende kunnskap, og som man forsøker å avkrefte eller bekrefte.

Definisjon: (statistisk modell til en forsøksrekke)

Vi har gjennomført en statistisk forsøksrekke med utvalg n.
Anta at $X_1, X_2, \ldots, X_n$ være i.i.d. [24] stokastiske forsøksvariabler for utfallet av forsøkene og X er tilhørende populasjonsvariabel.

En *statistisk modell* for forsøksrekken er da:

$$\boxed{X_1,\ X_2,\ X_3,\ \ldots\ ,X_n \overset{\text{i.i.d.}}{\sim} X \ \sim\ P_\theta(X=x)} \qquad (5.59)$$

hvor $\theta \in \Theta$ og $X \in V$, med:

$$V = \text{verdimengden til de stokastiske forsøksvariablene} \qquad (5.60)$$

$$P_\theta(X=x) = \text{parametrisk familie av sannsynlighetsfordelinger for } V \qquad (5.61)$$

$$\Theta = \text{parametermengden, dvs. de mulige verdiene for parametrene} \qquad (5.62)$$

■

[24] Independent identical distributed, dvs. uavhengige identiske fordelte variabler.

Eksempel: (statistisk modell - X_i, dvs. frisk/ikke frisk)

For forsøksrekken med $n = 100$ pasienter som fikk et nytt legemiddel, observerte vi om pasientene ble friske eller ikke.

I lign.(5.9) definerte vi forsøksvariablene:

$$X_i = \begin{cases} 1 & , \overbrace{\text{dersom pasient nr. } i \text{ blir frisk}}^{=p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=1-p} \end{cases} \tag{5.63}$$

Hva er den statistiske modellen for forsøkene beskrevet av
de stokastiske forsøksvariablene $X_1, X_2, \ldots, X_{100}$ og tilhørende populasjonsvariabel X?

Figur 5.19: Forsøk.

Løsning:

Statistisk forsøksrekke:
100 tilfeldige pasienter som fikk et nytt legemiddel. Hvert forsøk ble enten karakterisert som en suksess eller fiasko med en (ukjent) suksess-sannsynlighet p.

Den statistiske modellen for de stokastiske forsøksvariablene $X_1, X_2, \ldots, X_{100}$
og populasjonsvariabelen X er da gitt ved:

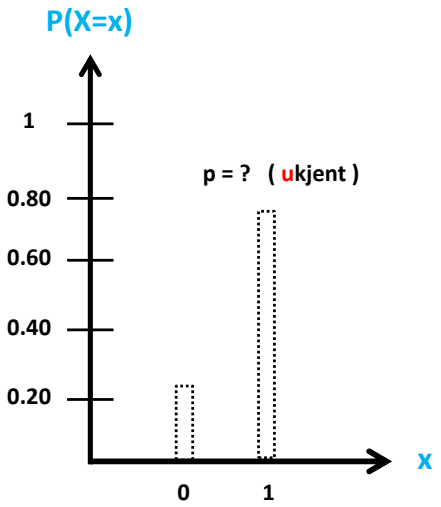
$$X_1,\ X_2,\ X_3,\ \ldots\ ,X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p]_{p\in[0,1]} \tag{5.64}$$

hvor

$$\{\,0,1\,\} = \text{verdimengden } V \text{ til de stokastiske variablene} \tag{5.65}$$

$$\text{Ber}[p]_{p\in[0,1]} = \text{familien av Bernouilli-fordelinger med parameter } \theta = p \tag{5.66}$$

$$[0,1] = \text{parametermengden } \Theta, \tag{5.67}$$
dvs. de mulige verdiene for parameteren $\theta = p$ som er $0 \le p \le 1$



Figur 5.20: $X \sim \text{Ber}[p]$, og p er ukjent.

■

Eksempel: (statistisk modell - Y_i , dvs. effekt)

For forsøksrekken med $n = 100$ pasienter som fikk et nytt legemiddel, ble effekten av legemiddelet målt som et tall mellom 0 og 1, se tabell 5.2.

I lign.(5.16) definerte vi:

$$Y_i = \underbrace{\text{effekten av legemiddelet}}_{\text{tall mellom 0 og 1, se tabell 5.2}} \text{ for forsøkspasient nr. } i \tag{5.68}$$

Hva er den statistiske modellen for de stokastiske forsøksvariablene $Y_1, Y_2, \ldots, Y_{100}$ og den tilhørende populasjonsvariabelen Y?

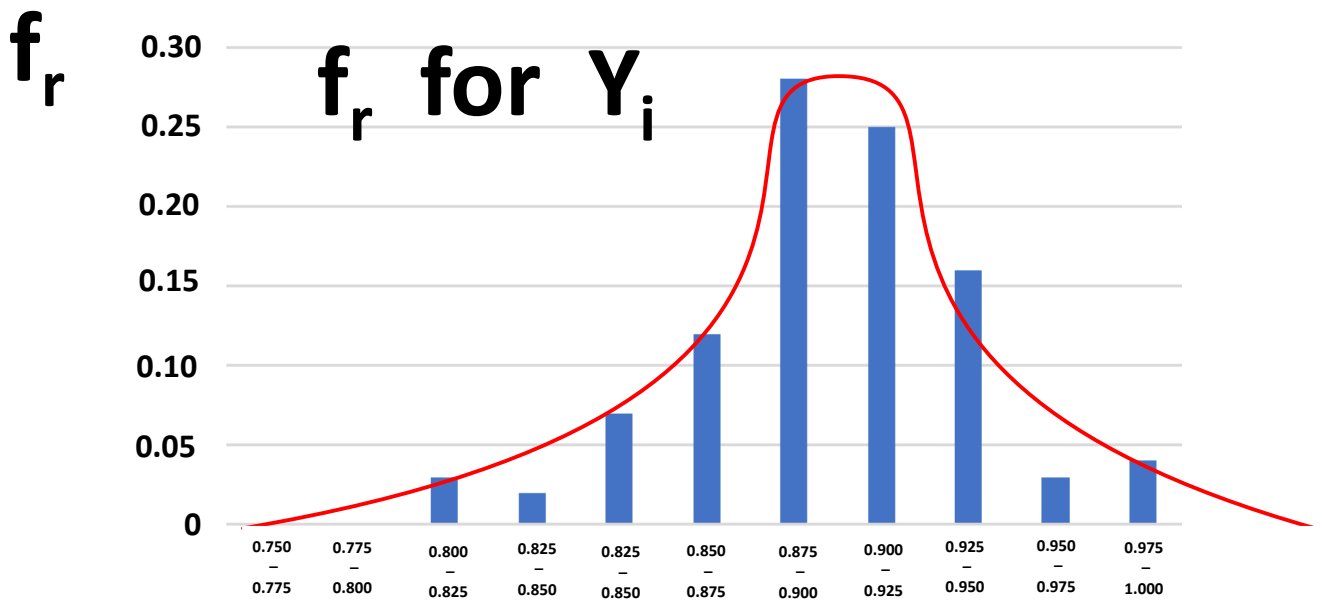
Figur 5.21: Forsøk.

Løsning:

Figur 5.22 viser frekvensdiagrammet for dataene fra tabell 5.2.

Formen på frekvensdiagrammet gir oss grunnlaget for å anta at disse stokastiske forsøksvariablene Y_i er tilnærmet *normalfordelte*, dvs.

$$Y_i \sim N[\mu, \sigma] \tag{5.69}$$



Figur 5.22: Relativ frekvenser f_r, se lign.(5.38)-(5.48).

Statistisk forsøksrekke:
100 tilfeldige pasienter som fikk et nytt legemiddel. Hvert forsøk ble enten karakterisert et tall mellom 0 og 1 som sier i hvilken grad en person blir frisk av legemiddelet.

Den statistiske modellen for de stokastiske forsøksvariablene $Y_1, Y_2, \ldots, Y_{100}$
og populasjonsvariabelen Y er da gitt ved:

$$\boxed{Y_1,\, Y_2,\, Y_3,\, \ldots\, ,Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu,\sigma]_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+}} \tag{5.70}$$

hvor [25]

$$\overbrace{\mathbb{R}}^{\text{reelle tall}} = \text{verdimengden } V \text{ til de stokastiske variablene} \tag{5.71}$$

$$N[\mu,\sigma]_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} = \text{familien av normalfordelinger med parametre } \mu \text{ og } \sigma \tag{5.72}$$

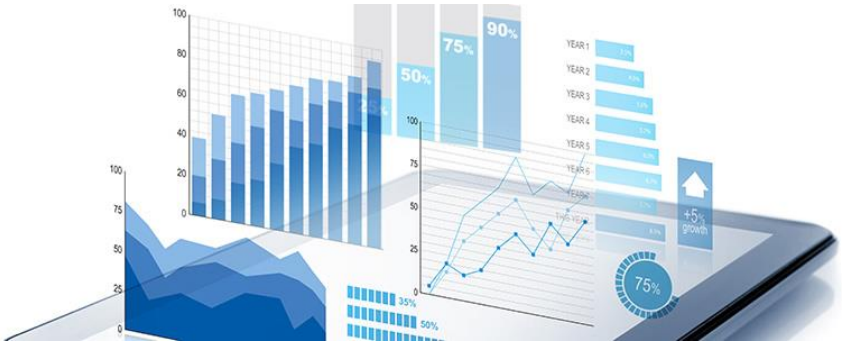
$$\mathbb{R}\times\mathbb{R}_+ = \text{parametermengden } \Theta, \text{ dvs. de mulige verdiene for parameteren } \mu \text{ og } \sigma \tag{5.73}$$

■

[25] $\mathbb{R}_+$ = alle positive reelle tall.
Elementene i $\mathbb{R}\times\mathbb{R}_+$ er alle par (μ,σ), hvor $-\infty < \mu < \infty$ og $\sigma > 0$.

Kapittel 6

Estimering og konfidensintervaller



Figur 6.1: Estimering og konfidensintervaller.

6.1 Motivasjon - statistisk inferens

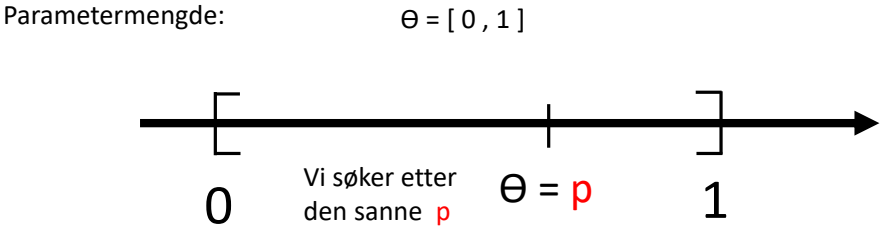
Vi har gjennomført en forsøksrekke på et tilfeldig utvalg av $n = 100$ pasienter som har fått et nytt legemiddel. Utfallet av et forsøk var enten suksess dersom effekten ble målt til 0.85 eller høyere og fiasko hvis ikke.

Statistisk modell:

Den statistiske modellen for de stokastiske forsøksvariablene $X_1, X_2, \ldots, X_{100}$ og populasjonsvariabelen X er gitt ved: (se lign.(5.64))

$$X_1,\ X_2,\ X_3,\ \ldots\ ,X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p]_{p\in[0,1]} \tag{6.1}$$

hvor $p \in [0, 1]$ er parametermengden og $V = \{0, 1\}$ verdimengden for modellen.



Figur 6.2: Mengden $\Theta = [0, 1]$.

To viktige størrelser:

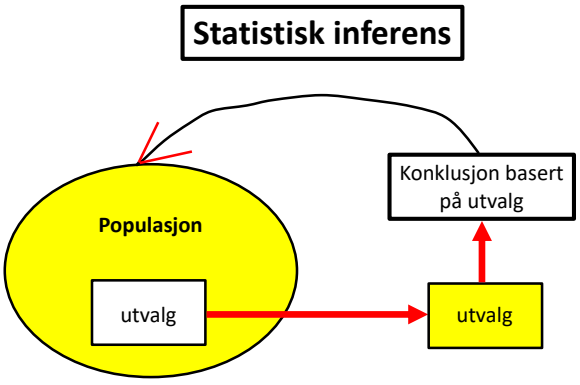
$$p \quad = \quad \text{den sanne (og ukjente) suksess-sannsynligheten for at legemiddel gir effekt} \geq 0.85 \tag{6.2}$$

$$\hat{p} \quad = \quad \text{et estimat for den sanne suksess-sannsynligheten } p \tag{6.3}$$

Vi er nå klare for å gjennomføre

statistisk inferens

dvs. vi er klare for å trekke konklusjoner (=inferere) basert på forsøksrekken om legemiddelets for hele populasjonen - altså for *enhver* pasient med den aktuelle lidelsen.



Figur 6.3: Statistisk inferens.

1) Første tilnærming - **inferens**:

Det viktigste målet er å finne et *estimat* $\hat{p}$ for den sanne suksess-sannsynligheten p.

Fra tabell 5.3 på side 270, dvs. resultatet om de ble friske eller ikke med $\overbrace{\text{0'ere og 1'ere}}^{\text{tabell 5.3}}$, finner man at det er 88 av 100 pasienter som ble friske. *Relativfrekvensen* er dermed:

$$\hat{p} = f_r = \frac{\text{gunstige}}{\text{mulige}} = \frac{\text{antall som ble friske}}{\text{antall som fikk medisin}} = \frac{88}{100} = 0.88 \qquad (6.4)$$

siden 88 av de 100 testede $\underbrace{\text{pasientene ble } \mathbf{friske}}_{\text{effekt } \geq 0.85}$.

Basert på $\hat{p}$, kan vi nå *inferere* (= konkludere) at suksess-sannsynligheten er 0.88 for *alle* med den aktuelle sykdommen - ikke bare for de som faktisk ble testet.

Spørsmål:

1. Er $\hat{p}$ en *korrekt* estimator for p?
2. Er $\hat{p}$ en *god* estimator for p?
3. Hva er *øvre og nedre grenser* for intervallet p tilhører med tilnærmet sannsynlighet 95 %?

For å forstå dybden i disse spørsmålene, må vi først innse at $\hat{p}$ er en *stokastisk variabel*.
Tallet 0.88 er en realisering av estimatoren basert på akkurat de 100 tilfeldig utvalgte pasientene vi valgte. Dersom vi gjentok forsøksrekken med at *nytt* utvalg på 100 pasienter, hadde vi ganske sikkert fått et nytt estimat.

2) Andre tilnærming - $\hat{p}$ er en stokastisk variabel:

Estimatoren $\hat{p}$ er forbundet med *usikkerhet* og er således en stokastisk variabel, dvs. $\hat{p}$ er en funksjon av de stokastiske forsøksvariablene $\underbrace{X_1, X_2 \ldots X_n}_{0 \text{ eller } 1}$: ($n = 100$)

$$\overbrace{\hat{p}}^{\text{stok. var.}} = \hat{p}(\overbrace{X_1, X_2 \ldots X_n}^{\text{stok. var.}}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \qquad (6.5)$$

Fra observasjonene $x_1, x_2 \ldots x_n$ (tallverdier 0 eller 1) fra forsøksrekken får vi en *realisering* av estimatoren:

$$\hat{p}(\overbrace{x_1, x_2 \ldots x_n}^{0 \text{ eller } 1}) = \underbrace{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{tall mellom 0 og 1}} = \frac{88}{100} = 0.88 \qquad (6.6)$$

som er samme svar som i den første tilnærmelsen i lign.(6.4).

Legg merke til at:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ er et tall mellom 0 og 1 siden $\overline{x}$ bare er gjennomsnittet mellom 0'ere og 1'ere
- X_i = den stokastiske forsøksvariabelen (som beskriver *sannsynlighetsmodellen* for forsøk nr. i)
- x_i = $\underbrace{\text{observasjonen}}_{\text{tall}}$ tilhørende forsøk nr. i

Svar på spørsmål:

1. Er $\hat{p}$ en *korrekt* estimator for p?

   Svar:

   En måte å si at en estimator er *korrekt* er dersom den forventede verdien til estimatoren er lik den sanne verdien:

$$\underline{\underline{E[\hat{p}]}} \;=\; E[\,\overline{X}\,] \;=\; E\Big[\,\frac{1}{n}\sum_{i=1}^{n} X_i\,\Big] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{n}{n}\,p \;=\; \underline{\underline{p}} \tag{6.7}$$

   siden, for en Bernoulli-fordeling, så er: [1]

$$E[X_i] \;=\; p \tag{6.11}$$

   En slik estimator $\hat{p}$ kalles en *forventningsrett* estimator.

■

[1] Fra lign.(5.15) på side 268 vet vi at både forsøksvariablene X_i og populasjonsvariablene X er Bernoulli-fordelt:

$$X_1,\, X_2,\, X_3,\, \ldots\,,\, X_n \;\overset{\text{i.i.d.}}{\sim}\; X \;\sim\; \text{Ber}[\,p\,] \tag{6.8}$$

For populasjonsvariablene er forventningen:

$$\underline{E[X]} \;=\; \sum_{i=0}^{1} x_i P(X = x_i) \;=\; 0\cdot(1-p) + 1\cdot p \;=\; \underline{p} \tag{6.9}$$

og tilsvarende for forsøksvariablene:

$$E[X_i] \;=\; p \tag{6.10}$$

2. Er $\hat{p}$ en *god* estimator for p?

   Svar:

   En estimator $\hat{p}$ er *god* dersom [2]

$$\boxed{\left| P_{\hat{p}}(A) - P_p(A) \right|} \qquad (6.12)$$

   er så liten som mulig for alle begivenheter $A \subset V$. Lign. (6.12) sier at sannsynlighetsloven $P_{\hat{p}}$ vi får fra estimatet $\hat{p}$ er så "nær" den virkelige sannsynlighetsloven P_p som mulig.

   Dette målet gir grunnlaget for å kunne si om en estimator er *bedre* enn en annen. Vi kommer ikke inn på dette området her.

■

[2] Streken | betyr absoluttverditegn, f.eks. $|0.4 - 0.6| = 0.2$.

3. Hva er *øvre og nedre grenser* for intervallet som p tilhører med tilnærmet sannsynlighet 95 %?

Svar:

Intervallet ($n = 100$)

$$\boxed{\left[\ LB_n^p\ ,\ UB_n^p\ \right]} \qquad (6.13)$$

inneholder den sanne sannsynligheten p med tilnærmet sannsynlighet minst 95 %, hvor: [3]

$$LB_n^p = 0.8264 \qquad (6.14)$$

$$UB_n^p = 0.9336 \qquad (6.15)$$

Intervallet i lign.(6.13) kalles et *asymptotisk* 95 %-*konfidensintervall*. [4]

■

I neste avsnitt formaliseres disse begrepene.

[3] At sannsynligheten er *tilnærmet* 95 % kommer av sentralgrenseteoremet, som sier at $\hat{p} = \overline{X}$ er *tilnærmet* normalfordelt.

[4] Vi skal senere vise hvordan man finner dette intervallet.

6.2 Estimatorer

Felles antagelser: (for definisjoner og setninger i avsnitt 6.2)

Anta at:

$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \tag{6.16}$$

for $\theta \in \Theta$.

Definisjon: (estimator)

En estimator $\hat{\theta}$ for den sanne parameteren $\theta \in \Theta$ er en funksjon av forsøksvariablene:

$$\boxed{\hat{\theta} \;=\; \hat{\theta}(X_1, X_2 \ldots X_n)} \tag{6.17}$$

■

Definisjon: (forventningsrett estimator)

En estimator $\hat{\theta}$ sies å være forventnings**rett** dersom

$$\boxed{E[\,\hat{\theta}\,] \;=\; \theta} \tag{6.18}$$

hvor θ den sanne parameteren $\theta \in \Theta$.

I motsatt fall er den forventningskjev.

■

Kommentarer:

- At en estimator er forventningsrett betyr at dersom forsøket gjentas mange ganger vil estimatoren i gjennomsnitt, i det lange løp, gi **rett** verdi.
- At en estimator er forventningskjev betyr at dersom forsøket gjentas mange ganger vil estimatoren i gjennomsnitt, i det lange løp, gi **gal** verdi, (dvs. en systematisk feil ved å bruke en estimator som ikke er forventningsrett).

Eksempel: (estimator for suksess-sannsynligheten p - legemiddel)

Vi viste i lign.(6.7) at estimatoren:

$$\boxed{\hat{p}(X_1, \ldots, X_{100}) \;=\; \overline{X}} \qquad (6.19)$$

er en forventningsrett estimator for p siden

$$E[\,\hat{p}(X_1, \ldots, X_{100})\,] \overset{\text{lign.(6.7)}}{=} p \qquad (6.20)$$

■

Eksempel: (estimator - Y_i, dvs. effekt)

Anta at:

$$Y_1,\, Y_2,\, Y_3,\, \ldots\, , Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\,\mu\,,\,\sigma\,] \tag{6.21}$$

hvor de stokastiske forsøksvariablene $\overbrace{Y_i \text{ for pasient nr. } i}^{i=1,2,3\ldots,100}$ er: (se lign.(5.16) side 269)

$$Y_i = \underbrace{\textbf{effekten} \text{ av legemiddelet}}_{\text{tall mellom 0 og 1, se tabell 5.2}} \text{ for forsøkspasient nr. } i \tag{6.22}$$

og de stokastisk populasjonsvariabelene er:

$$Y = \textbf{effekten} \text{ av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.23}$$

Lign.(6.21) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.

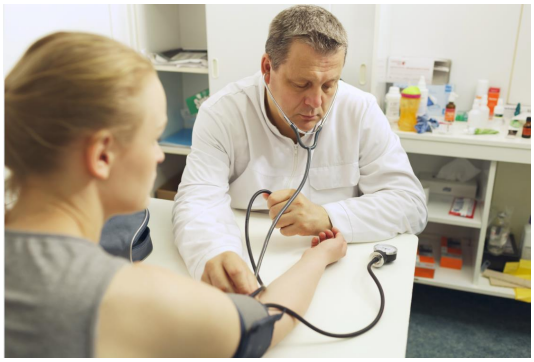


Foto: Colourbox

Figur 6.4: Forsøk.

Vi ønsker å finne *forventningsrette* estimatorer $\hat{\mu}$ og $\hat{\sigma}^2$.
Det er naturlig å foreslå følgende kandidater:

$$\hat{\mu} \;=\; \overline{Y} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (6.24)$$

$$\hat{\sigma}^2 \;=\; S_{y,n}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad (6.25)$$

hvor $n = 100$. [5]

a) Vis at $\hat{\mu}$ er forventningsrett.

b) Vis at $\hat{\sigma}^2$ ikke er forventningsrett.

c) Bestem på bakgrunn av oppgave **b** en estimator som er forventningsrett.
Kjenner du igjen denne fra kapittel 5?

[5] Legg merke til at det står $\frac{1}{n}$ i lign.(6.25). Det er fordi det er naturlig å foreslå et gjennomsnitt. Lign.(5.23) på side 274, derimot, har prefaktoren $\frac{1}{n-1}$ for den empiriske variansen s_x^2.

Løsning:

a) Forventningen av estimatoren $\hat{\mu}$ i lign.(6.24): [6]

$$E[\,\hat{\mu}\,] = E[\,\overline{Y}\,] \tag{6.26}$$

$$= E\Big[\,\frac{1}{n}\sum_{i=1}^{n} Y_i\,\Big] \tag{6.27}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\big[\,Y_i\,\big] = \frac{\cancel{n}\overbrace{E[Y_i]}^{=\mu}}{\cancel{n}} = \mu \tag{6.28}$$

siden vi har antatt at $Y_i \sim N[\mu, \sigma]$, dvs.

$$E[Y_i] = \mu \tag{6.29}$$

Lign.(6.28) viser at forventningen av estimatoren $\hat{\mu}$ er den sanne forventningen μ.

Estimatoren $\hat{\mu}$ er altså forventningsrett.

[6] Husk at $\hat{\mu}$ er en stokastisk variabel. Derfor gir det mening å ta forventningsverdien av den.

b) Forventningen av den estimatoren $\hat{\sigma}^2$i lign.(6.25): [7]

$$\underline{E[\,\hat{\sigma}^2\,]} = E[\,S^2_{y,n}\,] \tag{6.30}$$

$$= E\Big[\,\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\,\Big] \tag{6.31}$$

$$= E\Big[\,\frac{1}{n}\sum_{i=1}^{n}\Big(\,Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2\,\Big)\,\Big] \tag{6.32}$$

. . . and then a miracle occurs

$$= \underline{\frac{n-1}{n}\,\sigma^2} \tag{6.33}$$

som viser at forventningen av estimatoren $\hat{\sigma}^2$ ikke er den sanne variansen σ^2.

Estimatoren $S^2_{y,n}$ er altså ikke forventningsrett, dvs. den er forventningskjev.

[7] Husk at $\hat{\sigma}^2$ er en stokastisk variabel. Derfor gir det mening å ta forventningsverdien av den.

c) I lys av lign.(5.23) på side 274 så ser vi på følgende estimator:

$$\boxed{\hat{\sigma}^2 \;=\; S_y^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \qquad (6.34)$$

som har prefaktoren $\frac{1}{n-1}$. Fra lign.(6.25) ser vi at

$$\underline{S_y^2} \;=\; \frac{n}{n-1}\,\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \;\overset{\text{lign.(6.25)}}{=}\; \underline{\frac{n}{n-1}\,S_{y,n}^2} \qquad (6.35)$$

Forventningen av den stokastiske variabelen $\hat{\sigma}^2$ i lign.(6.25):

$$\begin{aligned}
\underline{E[\,\hat{\sigma}^2\,]} \;&=\; E[\,S_y^2\,] && (6.36)\\
&=\; E\Big[\,\frac{n}{n-1}\,S_{y,n}^2\,\Big] && (6.37)\\
&=\; \frac{n}{n-1}\,\overbrace{E\Big[\,S_{y,n}^2\,\Big]}^{=\,\frac{n-1}{n}\,\sigma^2} && (6.38)\\
&=\; \frac{\cancel{n}}{\cancel{n-1}}\,\frac{\cancel{n-1}}{\cancel{n}}\,\sigma^2 && (6.39)\\
&=\; \underline{\sigma^2} && (6.40)
\end{aligned}$$

som viser at forventningen av estimatoren $\hat{\sigma}^2$ i lign.(6.34) er den sanne forventningen σ^2. Estimatoren $\hat{\sigma}^2$ er altså forventningsrett.

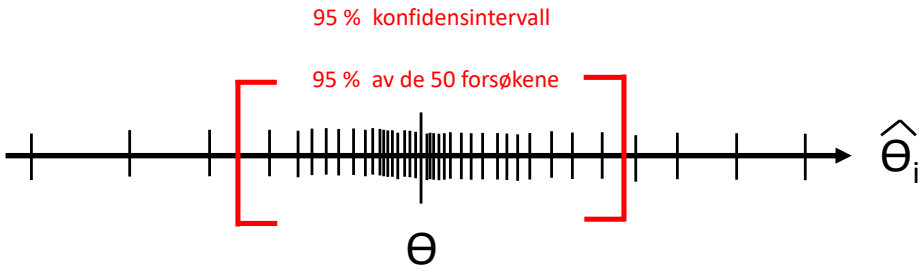
Ja, vi ser at realiseringen av estimatoren $\hat{\sigma}^2$ i lign.(6.34) er den empiriske variansen s_x^2 som definert i lign.(5.23) på side 274.

■

6.3 Konfidensintervaller

I forrige avsnitt definerte vi begrepet $\overbrace{\textit{estimator}\ \hat{\theta}}^{\text{stok. var.}}$ for $\overbrace{\theta}^{\textit{tall}}$.

En slik estimator er et *punktestimat*, dvs. det sier ikke noe om hvor mye $\hat{\theta}$ *varierer* rundt θ.
Hvis vi gjentar forsøksrekken 50 ganger, får vi 50 forskjellige punktestimat for θ.
Hvor stor er spredningen til disse 50 realiseringene $\hat{\theta}$?



Figur 6.5: Spredning rundt θ.

Konfidensintervallet gir oss svaret på spørsmålet ovenfor.
Vi antar, som i for forrige avsnitt om estimatorer, at vi har gjennomført en statistisk forsøksrekke.

Felles antagelser: (for definisjoner og setninger i avsnitt 6.3)

Anta at:

$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \tag{6.41}$$

for $\theta \in \Theta$ med verdimengde V.

Definisjon: ($(1-\alpha)\,100\,\%$-konfidensintervall for θ)

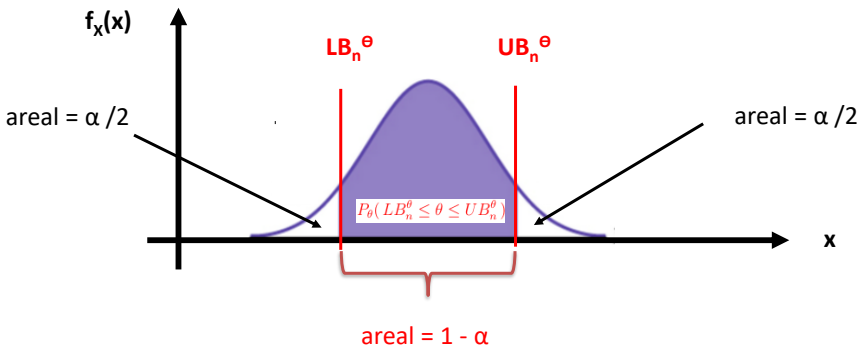
Et $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\,LB_n^\theta\,,\, UB_n^\theta\,]} \tag{6.42}$$

hvor LB_n^θ og UB_n^θ er fastsatt slik at sannsynligheten for at θ er inneholdt i dette intervallet er $(1-\alpha)\ 100\,\%$, dvs.:

$$\boxed{P_\theta(\, LB_n^\theta \leq \theta \leq UB_n^\theta\,) \;=\; 1-\alpha} \tag{6.43}$$

■



Figur 6.6: Konfidensintervall.

Kommentarer:

- Tallet α kalles

$$\boxed{\alpha \;=\; \text{signifikansnivået}}$$

og er en størrelse som er valgt av de som gjennomfører analysen. Som regel velges en *lav* verdi for α, dvs. typisk er $\alpha = 0.05$ eller lavere. Med $\alpha = 0.05$ fås et 95 % konfidensintervall, dvs. at intervallet inneholder θ med 95 % sannsynlighet.

- LB_n^θ står for "Lower Bound" eller nedre grense på norsk, og er en funksjon av forsøksvariablene:

$$\underbrace{LB_n^\theta \;=\; LB_n^\theta(X_1, \ldots X_n)}_{\text{stokastisk variabel}}$$

LB_n^θ er derfor en stokastisk variabel. Legg også merke til at LB_n^θ er en funksjon av størrelsen på utvalget, dsv. n.

- UB_n^θ står for "Upper Bound" eller øvre grense på norsk, og er en funksjon av forsøksvariablene:

$$\underbrace{UB_n^\theta \;=\; UB_n^\theta(X_1, \ldots X_n)}_{\text{stokastisk variabel}}$$

UB_n^θ er derfor en stokastisk variabel. Legg også merke til at UB_n^θ er en funksjon av størrelsen på utvalget, dvs. n.

Sannsynligheten i lign.(6.43) er ofte vanskelig å beregne nøyaktig.
Typisk vil både LB_n^θ og UB_n^θ være funksjoner av gjennomsnittet $\overline{X} = \frac{1}{n}\Big(X_1 + X_2 + X_3 + ... + X_n \Big)$.
Siden forsøksvariablene X_i er i.i.d. kan vi benytte *sentralgrenseteoremet* fra kapittel 4:
(typisk $n \gtrsim 30$, se lign.(4.172) på side 246)

$$\lim_{n\to\infty} \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \;\sim\; N[0,1] \tag{6.44}$$

Ved hjelp av lign.(6.44) kan tilnærmede uttrykk for LB_n^θ og UB_n^θ bestemmes slik at sannsynligheten i lign.(6.43) blir mer korrekt desto større n blir (utvalget).

Vi får det som kalles et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for θ.

Definisjon: (asymptotisk $(1-\alpha)\,100\,\%$-konfidensintervall for θ)

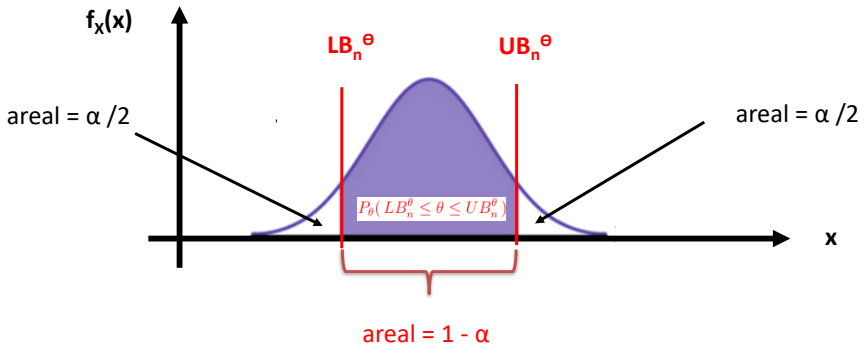
Et asymptotisk $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\,LB_n^\theta\,,\,UB_n^\theta\,]} \qquad (6.45)$$

hvor LB_n^θ og UB_n^θ er fastsatt slik at sannsynligheten for at θ er inneholdt i dette intervallet er $(1-\alpha)\,100\,\%$ når $n \to \infty$, dvs.:

$$\boxed{\lim_{n\to\infty} \; P_\theta(\; LB_n^\theta \;\leq\; \theta \leq UB_n^\theta\;) \;\;=\;\; 1-\alpha} \qquad (6.46)$$

■



Figur 6.7: Konfidensintervall.

Eksempel: (asymptotisk 95 %-konfidensintervall - sannsynlighet p)

Anta at:

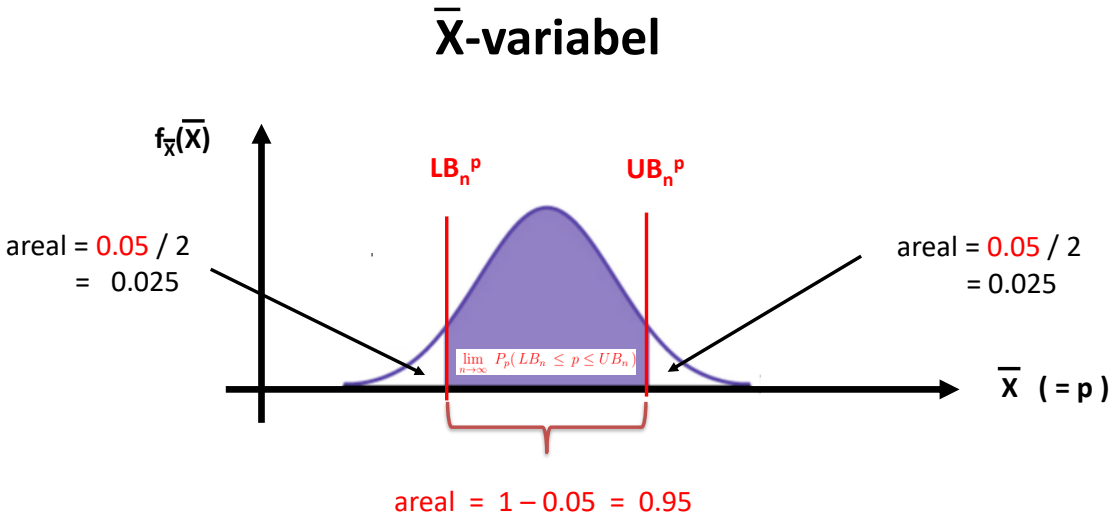
$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Ber}[p] \tag{6.47}$$

hvor de stokastiske forsøksvariablene X_i beskriver resultatene fra forsøkene med legemiddelet, dvs.:

$$X_i = \begin{cases} 1 & , \overbrace{\text{dersom pasient nr. } i \text{ blir frisk}}^{=\,p} \\ 0 & , \underbrace{\text{hvis ikke}}_{=\,1-p} \end{cases} \tag{6.48}$$

hvor

$$p = \text{suksess-sannsynlighet (ukjent } p \text{)} \tag{6.49}$$



Figur 6.8: 95 %-konfidensintervall.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^p \,,\, UB_n^p \,] \tag{6.50}$$

for den forventningsrette estimatoren $\hat{p}$, dvs.:

$$\hat{p}(X_1, \ldots, X_{100}) \overset{\text{lign.(6.19)}}{=} \overline{X} \tag{6.51}$$

med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = p$:

$$\boxed{\lim_{n\to\infty} P_p(\, LB_n^p \leq p \leq UB_n^p \,) \;=\; 1-\alpha} \qquad (6.52)$$

Fra lign.(6.19) vet vi da at: [8]

$$\hat{p} \;=\; \overline{X} \qquad (6.54)$$

er en forventningsrett estimator for den ukjente p.
Alle X_i i lign.(6.54) oppfyller i.i.d. (med $X_i \sim \text{Ber}[p]$). Fra sentralgrenseteoremet vet vi da at:

$$Z \;\; \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \;\; \overset{n\to\infty}{\sim} \;\; N\big[\,0\,,\,1\,\big] \qquad (6.55)$$

Med $n = 100$ så er n stor nok til at Z i lign.(6.55) med god tilnærmelse er en normalfordeling.
Vi må finne $E[\overline{X}]$ og $\sigma[\overline{X}]$ for å standardisere.

[8] Husk at:

$$\overline{X} \;=\; \frac{X_1 + X_2 + ...X_n}{n} \qquad (6.53)$$

Fra lign.(6.7) vet vi at

$$E[\hat{p}] \quad = \quad E[\overline{X}] \overset{\text{lign.(6.7)}}{=} p \tag{6.56}$$

altså estimatoren $\hat{p}$ er forvetningsrett. På tilsvarende måte kan man vise at: [9]

$$Var[\hat{p}] \quad = \quad Var[\overline{X}] \; = \; \frac{p(1-p)}{n} \tag{6.64}$$

Generelt er $\sigma[X] = \sqrt{Var[X]}$, dermed:

$$\sigma[\hat{p}] \quad = \quad \sigma[\overline{X}] \; = \; \sqrt{\frac{p(1-p)}{n}} \tag{6.65}$$

[9] Fra lign.(5.15) på side 268 vet vi at både forsøksvariablene X_i og populasjonsvariablene X er Bernoulli-fordelt:

$$X_1,\, X_2,\, X_3,\, \ldots\, , X_n \overset{\text{i.i.d.}}{\sim} X \; \sim \; \text{Ber}[p] \tag{6.57}$$

Variansen av $\hat{p} \; = \; \overline{X}$:

$$\underline{Var[\hat{p}]} \quad = \quad Var[\overline{X}] \tag{6.58}$$

$$= \quad Var\Big[\, \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \,\Big] \tag{6.59}$$

$$\overset{\text{i.i.d.}}{=} \quad \frac{1}{n^2}\Big(\, Var[X_1] + Var[X_2] + Var[X_3] + ... + Var[X_n] \, \Big) \tag{6.60}$$

$$= \quad \frac{\cancel{n} Var[X_i]}{n^{\cancel{2}}} \quad = \quad \underline{\frac{p(1-p)}{n}} \tag{6.61}$$

hvor vi har brukt at $Var[X_i] = Var[X]$ med

$$\underline{Var[X]} \quad = \quad \sum_{i=0}^{1} \big(\, x_i - E[X] \, \big) P(X = (x_i)) \tag{6.62}$$

$$= \quad (0-p)^2(1-p) \; + \; (1-p)^2 p \; = \; \underline{p(1-p)} \tag{6.63}$$

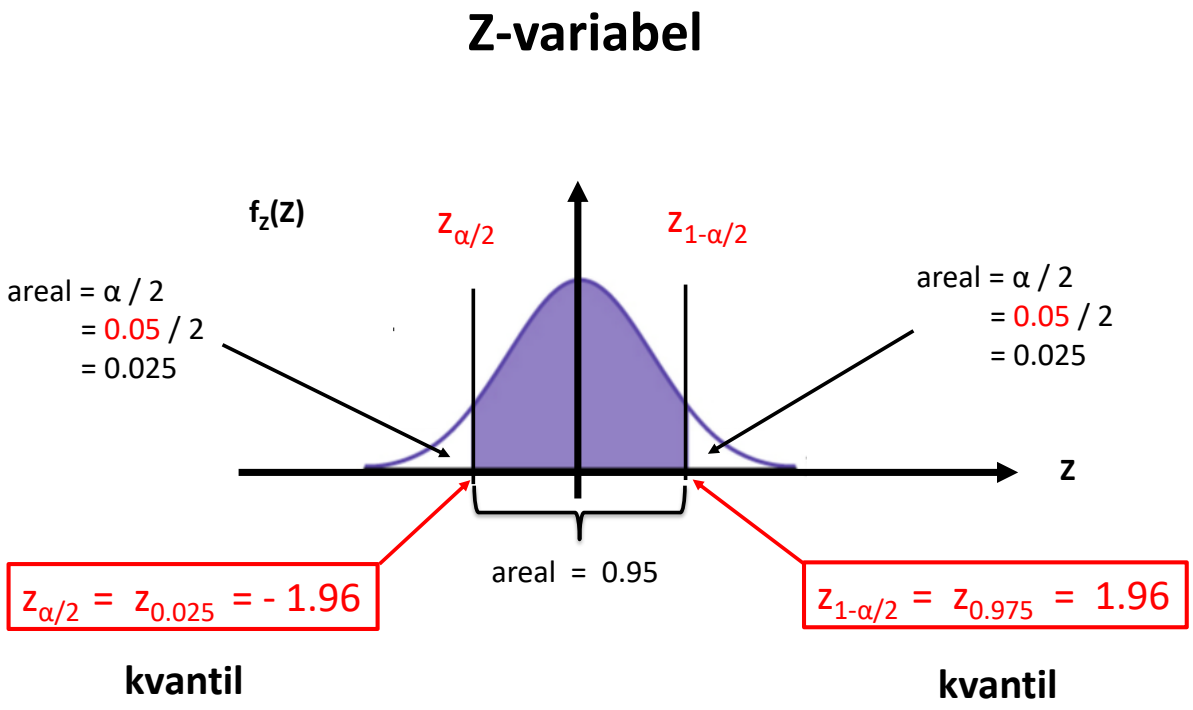
Variabelen Z i lign.(6.55) er da:

$$\underline{Z} \;=\; \frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]} \;=\; \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \;\overset{n\to\infty}{\sim}\; N[0,1] \tag{6.66}$$

Vi finner kvantilene i figur 6.9: [10]

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \quad , \quad z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.67}$$

via omvendt tabelloppslag fra side 206.



Figur 6.9: Kvantil.

[10] Absoluttverdiene til $z_{\alpha/2}$ og $z_{1-\alpha/2}$ er like fordi N-fordelingen er symmetrisk.

Fra definisjonen i lign.(6.52):

$$\lim_{n\to\infty} \; P_p(\, LB_n^p \;\leq\; p \leq UB_n^p \,) \;\; = \;\; 1-\alpha \tag{6.68}$$

Siden estimatoren $\hat{p}$ er normalfordelt for $n \to \infty$, se lign.(6.54) og (6.55), så standardiserer vi:

$$\lim_{n\to\infty} \; P\big(\, z_{\alpha/2} \;\leq\; Z \leq z_{1-\alpha/2} \,\big) \;\; = \;\; 1-\alpha \tag{6.69}$$

hvor Z er gitt ved lign.(6.66):

$$\lim_{n\to\infty} \; P\left(\; z_{\alpha/2} \;\leq\; \frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \;\leq\; z_{1-\alpha/2} \; \right) \;=\; 1-\alpha \tag{6.70}$$

$$\Updownarrow \qquad \text{(algebra)}$$

$$\lim_{n\to\infty} \; P\left(\; \underbrace{\overline{X} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)}}_{=\; LB_n^p} \;\leq\; p \;\leq\; \underbrace{\overline{X} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)}}_{=\; UB_n^p} \; \right) \;=\; 1-\alpha \tag{6.71}$$

og hvor vi definerer nedre og øvre grenser:

$$LB_n^p \;\; = \;\; \overline{X} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)} \tag{6.72}$$

$$UB_n^p \;\; = \;\; \overline{X} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)} \tag{6.73}$$

Problem:

Vi kan ikke bestemme en realisering av LB_n^p og UB_n^p siden de er avhengige av den ukjente suksess-sannsynligheten p.

Løsning:

Vi kan bytte ut p i lign.(6.72) og (6.73) med estimatoren $\hat{p} = \overline{X}$.
Det viser seg at denne tilnærmingen også oppfyller sentralgrenseteoremet. [11]

Med $\hat{p} = \overline{X}$ innsatt for p i lign.(6.72) og (6.73):

$$LB_n^p = \overline{X} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})} \tag{6.74}$$

$$UB_n^p = \overline{X} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})} \tag{6.75}$$

hvor

$$z_{\alpha/2} = \text{nedre kvantil til } N[0,1]\text{-fordelingen} \tag{6.76}$$

$$z_{1-\alpha/2} = \text{øvre kvantil til } N[0,1]\text{-fordelingen} \tag{6.77}$$

[11] Dette kommer som en følge av den såkalte store talls lov, som vi ikke skal komme inn på her.

Intervallet

$$\boxed{[\, LB_n^p \,,\, UB_n^p \,] \;=\; \left[\; \overline{X} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{X}(1-\overline{X})} \;,\; \overline{X} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\overline{X}(1-\overline{X})} \;\right]} \tag{6.78}$$

er dermed et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for suksess-sannsynligheten p.

Med $\alpha = 0.05$ så finner vi ved tabelloppslag:

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \tag{6.79}$$

$$z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.80}$$

Fra dataene $x_1, x_2 \ldots x_{100}$ fra forsøksrekken får vi en *realisering* (små $x_1, x_2, \ldots, x_{100}$) av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^p(x_1, x_2, \ldots, x_n)} = \overline{x} - \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{x}(1-\overline{x})} \tag{6.81}$$

$$= 0.88 - \frac{1.96}{\sqrt{100}} \sqrt{0.88(1-0.88)} \tag{6.82}$$

$$= \underline{0.8163} \tag{6.83}$$

$$\underline{UB_n^p(x_1, x_2, \ldots, x_n)} = \overline{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\overline{x}(1-\overline{x})} \tag{6.84}$$

$$= 0.88 - \frac{(-1.96)}{\sqrt{100}} \sqrt{0.88(1-0.88)} \tag{6.85}$$

$$= \underline{0.9437} \tag{6.86}$$

som gir realiseringen

$$\underline{\underline{[\, LB_n^p \,,\, UB_n^p \,]}} = [\, 0.8163 \,,\, 0.9437 \,] \tag{6.87}$$

av det asymptotiske 95 %-konfidensintervallet for sukess-sannsynligheten p.

■

Eksempel: (asymptotisk 95 %-konfidensintervall - effekten $\hat{\mu} = \overline{Y}$)

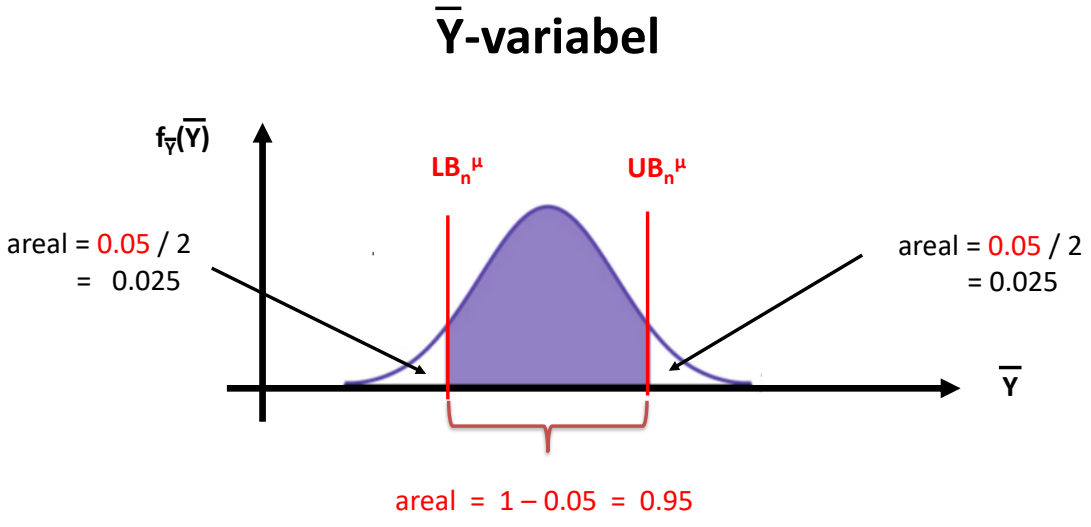
Anta at:

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma] \tag{6.88}$$

hvor de stokastiske forsøksvariablene Y_i beskriver effekten for pasient nr. i:

$$Y_i \overset{\text{lign.(5.16)}}{=} \text{effekten for forsøkspasient nr. } i \tag{6.89}$$

Lign.(6.88) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.



Figur 6.10: 95 %-konfidensintervall.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^{\mu}\,,\, UB_n^{\mu}\,] \tag{6.90}$$

for den sanne forventningsverdien μ med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = \mu$:

$$\boxed{\lim_{n\to\infty} P_\mu(LB_n^\mu \leq \mu \leq UB_n^\mu) = 1-\alpha} \qquad (6.91)$$

Alle Y_i i lign.(6.88) oppfyller i.i.d. (med $Y_i \sim \mathrm{N}[\mu, \sigma]$).
Sum normalfordelinger er forsatt normalfordelt: (se lign.(4.180) side 251)

$$\overline{Y} \sim N\big[\, E[\overline{Y}]\,,\, \sigma[\overline{Y}]\,\big] \qquad (6.92)$$

Vi må standardisere. [12]

[12] Estimatoren $\hat{\mu}$ er forventningsrett, dvs.:

$$E[\,\hat{\mu}\,] = E[\,\overline{Y}\,] = \mu \qquad (6.93)$$

altså estimatoren $\hat{\mu}$ er forvetningsrett. Variansen til $\hat{\mu}$ er:

$$Var[\,\hat{\mu}\,] = Var[\,\overline{Y}\,] = \frac{\sigma^2}{n} \qquad (6.94)$$

Generelt er $\sigma[Y] = \sqrt{Var[Y]}$, dermed:

$$\sigma[\,\hat{\mu}\,] = \sigma[\overline{Y}] = \frac{\sigma}{\sqrt{n}} \qquad (6.95)$$

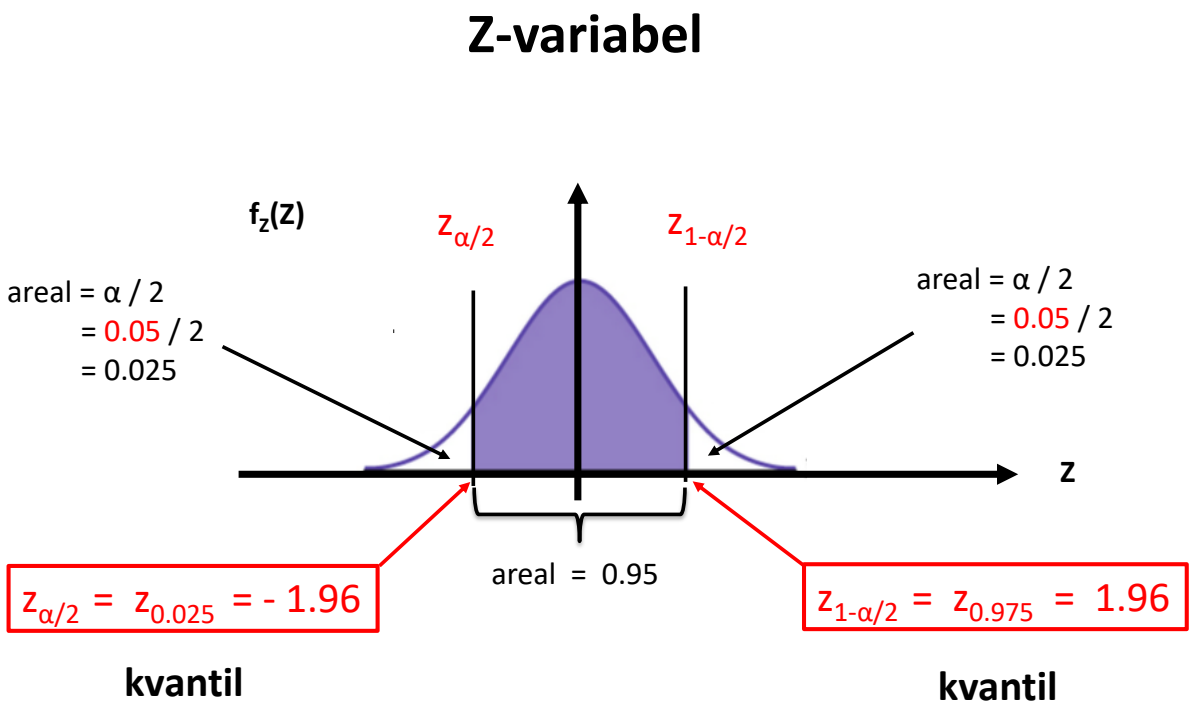
Standardiserer:

$$\underline{Z} = \frac{\overline{X} - E[\overline{Y}]}{\sigma[\overline{Y}]} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{n\to\infty}{\sim} N[0,1] \tag{6.96}$$

Vi finner kvantilene figur 6.11: [13]

$$z_{\alpha/2} = z_{0.025} = -1.96 \quad , \quad z_{1-\alpha/2} = z_{0.975} = 1.96 \tag{6.97}$$

via omvendt tabelloppslag fra side 206.



Figur 6.11: Kvantil.

[13] Absoluttverdiene til $z_{\alpha/2}$ og $z_{1-\alpha/2}$ er like fordi N-fordelingen er symmetrisk.

Fra definisjonen i lign.(6.91):

$$\lim_{n\to\infty} \; P_\mu(\, LB_n^\mu \;\leq\; \mu \leq UB_n^\mu \,) \quad = \quad 1-\alpha \tag{6.98}$$

Siden estimatoren $\hat{\mu}$ er normalfordelt, se lign.(6.92), så standardiserer vi:

$$\lim_{n\to\infty} \; P\big(\, z_{\alpha/2} \;\leq\; Z \leq z_{1-\alpha/2} \,\big) \quad = \quad 1-\alpha \tag{6.99}$$

hvor Z er gitt ved lign.(6.96):

$$\lim_{n\to\infty} \; P\left(\; z_{\alpha/2} \;\leq\; \frac{\overline{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \;\leq\; z_{1-\alpha/2} \; \right) \;=\; 1-\alpha \tag{6.100}$$

$$\Updownarrow \qquad \text{(algebra)}$$

$$\lim_{n\to\infty} \; P\left(\; \underbrace{\overline{Y} - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma}_{=\; LB_n^\mu} \;\leq\; \mu \;\leq\; \underbrace{\overline{Y} - \frac{z_{\alpha/2}}{\sqrt{n}}\,\sigma}_{=\; UB_n^\mu} \; \right) \;=\; 1-\alpha \tag{6.101}$$

og hvor vi definerer nedre og øvre grenser:

$$LB_n^\mu \quad = \quad \overline{Y} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\sigma \tag{6.102}$$

$$UB_n^\mu \quad = \quad \overline{Y} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\,\sigma \tag{6.103}$$

Problem:

Vi kan ikke bestemme en realisering av LU_n^μ og UB_n^μ siden de er avhengige av den ukjente standardavviket σ.

Løsning:

Vi kan bytte ut σ i lign.(6.102) og (6.103) med estimatoren $\hat{\sigma} = S_y$, jfr. lign.(6.34) på side 309:

$$\hat{\sigma}^2 = S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{6.104}$$

Det viser seg at denne tilnærmingen også oppfyller sentralgrenseteoremet. [14]

Med $\hat{\sigma} = S_y$ innsatt for σ i lign.(6.102) og (6.103):

$$LB_n^\mu = \overline{Y} - \frac{z_{1-\alpha/2}}{\sqrt{n}} S_y \tag{6.105}$$

$$UB_n^\mu = \overline{Y} - \frac{z_{\alpha/2}}{\sqrt{n}} S_y \tag{6.106}$$

[14] Dette kommer som en følge av den såkalte store talls lov, som vi ikke skal komme inn på her.

Intervallet

$$\boxed{[\, LB_n^{\mu}\, ,\, UB_n^{\mu}\,] \;=\; \left[\; \overline{Y} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\, S_y \;,\; \overline{Y} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\, S_y \;\right]} \tag{6.107}$$

er dermed et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for σ.

Med $\alpha = 0.05$ så finner vi ved tabelloppslag:

$$z_{\alpha/2} \;=\; z_{0.025} \;=\; -1.96 \tag{6.108}$$

$$z_{1-\alpha/2} \;=\; z_{0.975} \;=\; 1.96 \tag{6.109}$$

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 269) samt $s_y = 0.0413$ fra tabell 5.5 side 276 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$) av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{z_{1-\alpha/2}}{\sqrt{n}} s_y \tag{6.110}$$

$$= 0.8967 - \frac{1.96}{\sqrt{100}} \cdot 0.0413 \tag{6.111}$$

$$= \underline{0.8886} \tag{6.112}$$

$$\underline{UB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{z_{\alpha/2}}{\sqrt{n}} s_y \tag{6.113}$$

$$= 0.8967 - \frac{(-1.96)}{\sqrt{100}} \cdot 0.0413 \tag{6.114}$$

$$= \underline{0.9048} \tag{6.115}$$

som gir realiseringen

$$\underline{\underline{[\ LB_n^{\mu}\ ,\ UB_n^{\mu}\ ]}} = [\ 0.8886\ ,\ 0.9048\ ] \tag{6.116}$$

av det asymptotiske 95 %-konfidensintervallet for μ.

■

Eksempel: (asymptotisk 95 %-konfidensintervall - standardavvik $\hat{\sigma} = S_y$ til effekten Y)

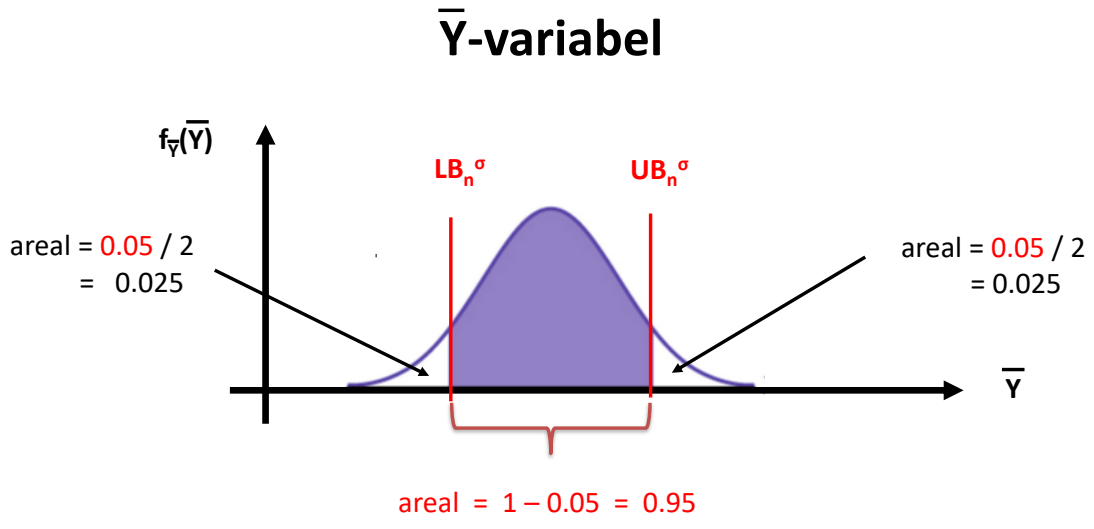
Anta at:

$$Y_1,\, Y_2,\, Y_3,\, \ldots\, , Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\,\mu\,,\,\sigma\,] \tag{6.117}$$

hvor de stokastiske populasjonsvariablene Y_i beskriver effekten for pasient nr. i:

$$Y_i \overset{\text{lign.(5.16)}}{=} \text{effekten for forsøkspasient nr. } i \tag{6.118}$$

Lign.(6.117) betyr blant annet at $E[Y] = E[Y_i] = \mu$ og at $\sigma^2[Y] = \sigma^2[Y_i] = \sigma^2$.



Figur 6.12: 95 %-konfidensintervall for σ^2.

Finn et asymptotisk konfidensintervall

$$[\, LB_n^{\sigma},\, ,\, UB_n^{\sigma},\,] \tag{6.119}$$

for den forventningsrette estimatoren $\hat{\sigma}$, hvor:

$$\hat{\sigma}^2 \;=\; S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{6.120}$$

med signifikansnivå $\alpha = 0.05$.

Løsning:

Definisjonen av et asymptotisk konfidensintervall er gitt ved lign.(6.46) med $\theta = \sigma$:

$$\lim_{n\to\infty} P_\mu(LB_n^\sigma \leq \sigma \leq UB_n^\sigma) = 1-\alpha \qquad (6.121)$$

På samme måte som i forrige eksempel så finner man at intervallet

$$[\ LB_n^\sigma\ ,\ UB_n^\sigma,\ ] = \left[\sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\, S_y\ ,\ \sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\, S_y \right] \qquad (6.122)$$

er et *asymptotisk* $(1-\alpha)\,100\,\%$-konfidensintervall for σ.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 269) samt
$s_y^2 = 0.0017$ fra tabell 5.5 side 276 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$)
av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{(n-1) + z_{\alpha/2}\sqrt{2(n-1)}}}\, s_y \tag{6.123}$$

$$= \sqrt{\frac{100-1}{(100-1) + 1.96\sqrt{2(100-1)}}} \cdot 0.0413 \tag{6.124}$$

$$= \underline{0.0365} \tag{6.125}$$

$$\underline{UB_n^\sigma(y_1, y_2 \ldots, y_n)} = \sqrt{\frac{n-1}{(n-1) + z_{\alpha/2}\sqrt{2(n-1)}}}\, s_y \tag{6.126}$$

$$= \sqrt{\frac{100-1}{(100-1) - 1.96\sqrt{2(100-1)}}} \cdot 0.0413 \tag{6.127}$$

$$= \underline{0.0486} \tag{6.128}$$

som gir realiseringen

$$\underline{\underline{[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ]}} = [\ 0.0365\ ,\ 0.0486\ ] \tag{6.129}$$

av det asymptotiske 95 %-konfidensintervallet for standardavviket σ til effekten Y.

■

6.4 Student's t-fordeling og χ^2_k-fordeling

I forrige avsnitt konstruerte vi asymptotiske 95%-konfidensintervaller for (μ, σ) for effekten av legemiddelet, se lign.(6.107) og (6.122):

$$[\, LB_n^{\mu}\,,\, UB_n^{\mu}\,] \overset{\text{lign.(6.107)}}{=} \left[\, \overline{X} \;-\; \frac{z_{1-\alpha/2}}{\sqrt{n}}\, S_y \;,\; \overline{X} \;-\; \frac{z_{\alpha/2}}{\sqrt{n}}\, S_y \,\right] \tag{6.130}$$

$$[\, LB_n^{\sigma},\; UB_n^{\sigma}\,] \overset{\text{lign.(6.122)}}{=} \left[\, \sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\, S_y \;,\; \sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\, S_y \,\right] \tag{6.131}$$

hvor

$$X = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.132}$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \tag{6.133}$$

$$z_{\alpha/2} = \text{nedre kvantil til } N[0,1]\text{-fordelingen} \tag{6.134}$$

$$z_{1-\alpha/2} = \text{øvre kvantil til } N[0,1]\text{-fordelingen} \tag{6.135}$$

$$n = \text{antall forsøk (antall pasienter)} \tag{6.136}$$

$$\alpha = \text{signifikansnivået} \tag{6.137}$$

Spørsmål:

Kan vi konstruere 95 %-konfidensintervaller for (μ, σ) for effekten av legemiddelet selv når antall forsøk er lavt, f.eks. $n \gtrsim 30$? [15]

Sagt med andre ord:
Kan vi finne eksakte 95 %-konfidensintervaller for (μ, σ)? [16]

Svar:

Svaret er ja, men da må vi introdusere to nye sannsynlighetsfordelinger:

student's t-fordelingen for eksakt konfidensintervall for μ

χ^2_k-fordelingen for eksakt konfidensintervall for σ

[15] Grensen på 30 forsøk kommer fra gyldigheten til sentralgrenseteoremet, typisk $n \gtrsim 30$, se lign.(4.172) på side 246.

[16] Dvs. ikke asymptotiske.

6.4.1 χ^2_k-fordeling ("kji"-fordeling)

Definisjon: (χ^2_k-fordeling)

La $Z_1, Z_2, \ldots, Z_k$ være i.i.d. standard normalfordelte stokastiske variabler, dvs. $Z_i \sim N[0,1]$. Da er:

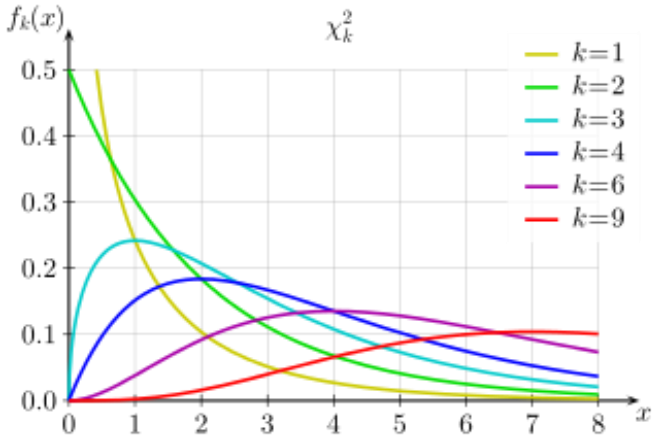
$$Q = \sum_{i=1}^{k} Z_i^2 = Z_1^2 + \cdots + Z_k^2 \qquad (6.138)$$

kji-fordelt med k frihetsgrader. Vi skriver:

$$Q \sim \chi^2_k \qquad (6.139)$$

■

Figur 6.13 viser χ^2_k-fordelingen for ulike ulike frihetsgrader k.



Figur 6.13: χ^2_k-fordelingen.

Setning: (S_y^2 er χ^2_{n-1}-fordelt)

La $Y_1, Y_2, \ldots, Y_n$ være i.i.d. normalfordelte stokastiske variabler, dvs. $Y_i \sim N[\mu, \sigma]$.
Da er:

$$\boxed{\frac{(n-1)S_y^2}{\sigma^2} \sim \chi^2_{n-1}} \qquad (6.140)$$

χ^2_{n-1}-fordelt med $n-1$ frihetsgrader, hvor

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \qquad (6.141)$$

■

Kommentar

- Dette resultatet er svært viktig siden det gir oss fordelingen til estimatoren $\hat{\sigma}^2 = S_y^2$ for variansen σ^2. Vi kan dermed bruke dette teoremet for å konstruere et eksakt 95 %-konfidensintervall for σ.

Men først introduserer vi Student's t-fordelingen som bygger på kji-fordelingen og som gir oss muligheten for å konstruere et eksakt 95 %-konfidensintervall for μ.

6.4.2 Student's t-fordeling

Definisjon: (Student's t-fordeling)

La $Z \sim N[0,1]$ og $Q \sim \chi^2_n$ være stokastiske variabler.
Da er:

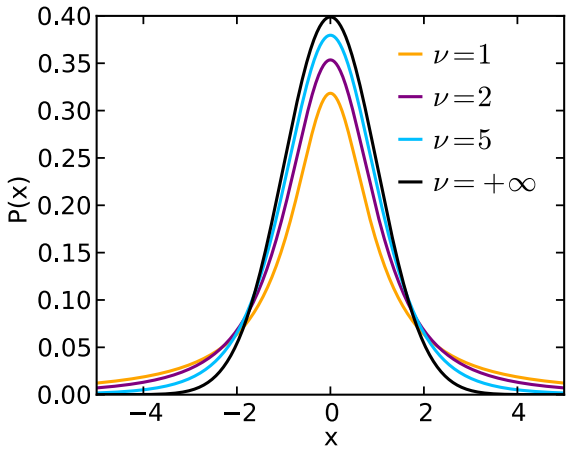
$$T = \frac{\sqrt{k}Z}{\sqrt{Q}} \tag{6.142}$$

Student's t-fordelt med $n-1$ frihetsgrader. Vi skriver

$$T \sim t_{n-1} \tag{6.143}$$

■

Figur 6.14 viser Student's t-fordelingen for ulike ulike frihetsgrader.



Figur 6.14: Student's t-fordelingen.

Setning: (Student's t-fordeling)

La $Y_1, Y_2, \ldots, Y_n$ være i.i.d. normalfordelte stokastiske variabler, dvs. $Y_i \sim N[\mu, \sigma]$.
Da er den stokastiske variabelen

$$\boxed{T = \frac{\overline{Y} - \mu}{S_y/\sqrt{n}} \sim t_{n-1}} \qquad (6.144)$$

Student's t-fordelt med $n-1$ frihetsgrader.

■

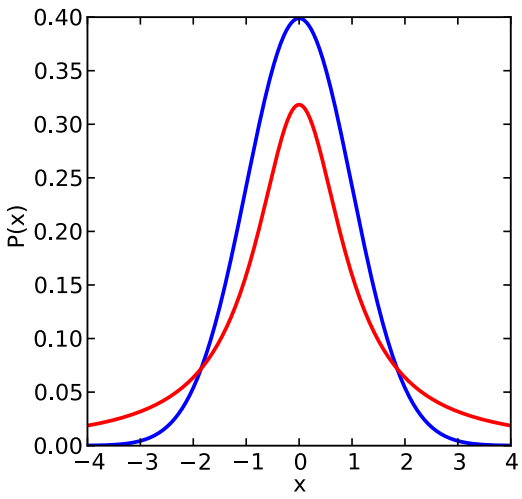
Bevis:

Vi hopper over beviset i dette kompendiet.

■

Kommentarer

- Student's t-fordelingen er svært lik normalfordelingen, men for liten n er halene mye "tykkere" enn halene til normalfordelingen, se figur 6.15.
- Når n vokser, blir Student's t-fordelingen mer og mer lik normalfordelingen, og i grensen når $n \to \infty$, er de helt like.
- Antall frihetsgrader uttrykker graden av usikkerhet som skyldes at σ er estimert.
  Stor frihetsgrad betyr liten usikkerhet.



Figur 6.15: Student's t-fordelingen (rød) med $n = 1$ frihetsgrad sammenlignet med normalfordelingen (blå).

6.4.3 Eksakte $(1-\alpha)$-konfidensintervaller for μ og σ

Eksempel: (eksakte 95%-konfidensintervall for μ av *effekten* av legemiddel)

La oss se på eksemplet med legemiddel:

$$Y \overset{\text{lign.(6.23)}}{=} \text{effekten av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.145}$$

hvor $Y \sim N[\mu, \sigma]$ med:

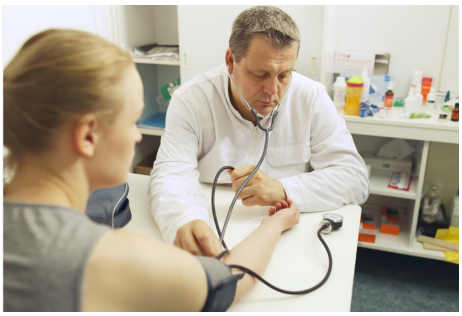
$$E[Y] = \mu \tag{6.146}$$

$$Var[Y] = \sigma^2 \tag{6.147}$$

Bruker resultatet i lign. (6.144) for å konstruere et eksakt $(1-\alpha)$-konfidensintervall for μ i eksempelet med legemiddel:

$$T = \frac{\sqrt{n}(\overline{Y}-\mu)}{S_y} \sim t_{n-1} \tag{6.148}$$

Foto: Colourbox



Figur 6.16: Forsøk.

Løsning:

Vi bruker Student's t-fordelingens kvantiler, $q_{\alpha/2}$ og $q_{1-\alpha/2}$, for å konstruere et eksakt 95 %-konfidensintervall for μ:

$$P\left(q_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{Y}-\mu\right)}{S_y} \leq q_{1-\alpha/2} \right) = 1-\alpha \tag{6.149}$$

$$\Updownarrow \quad \text{(algebra)}$$

$$P\left(\underbrace{\overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}} S_y}_{= LB_n^\mu} \leq \mu \leq \underbrace{\overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}} S_y}_{= UB_n^\mu} \right) = 1-\alpha \tag{6.150}$$

hvor vi definerer dermed nedre og øvre grenser:

$$LB_n^\mu = \overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}} S_y \tag{6.151}$$

$$UB_n^\mu = \overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}} S_y \tag{6.152}$$

hvor

$$q_{\alpha/2} = \text{nedre kvantil til Student's } t\text{-fordelingen} \tag{6.153}$$

$$q_{1-\alpha/2} = \text{øvre kvantil til Student's } t\text{-fordelingen} \tag{6.154}$$

Intervallet

$$[\ LB_n^{\mu}\ ,\ UB_n^{\mu}\ ]\ =\ \left[\ \overline{Y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\,S_y\ ,\ \overline{Y} - \frac{q_{\alpha/2}}{\sqrt{n}}\,S_y\ \right] \tag{6.155}$$

er dermed et eksakt $(1-\alpha)\,100\,\%$-konfidensintervall for μ for effekten Y.

Med $\alpha = 0.05$ og $n = 100$ for effekten Y så finner vi ved tabelloppslag: [17]

$$q_{\alpha/2} \quad = \quad q_{0.025} \ = -1.984 \tag{6.158}$$

$$q_{1-\alpha/2} \quad = \quad q_{0.975} \ = \ \ 1.984 \tag{6.159}$$

[17] Til sammenligning har vi

$$z_{\alpha/2} \quad = \quad z_{0.025} \ = -1.96 \tag{6.156}$$

$$z_{1-\alpha/2} \quad = \quad z_{0.975} \ = \ \ 1.96 \tag{6.157}$$

for N-fordeling, altså tykkere "hale" i Student's t-fordeling.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 269) samt $s_y = 0.0413$ fra tabell 5.5 side 276 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$ av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{q_{1-\alpha/2}}{\sqrt{n}}\, s_y \tag{6.160}$$

$$= 0.8967 - \frac{1.984}{\sqrt{100}} \cdot 0.0413 \tag{6.161}$$

$$= \underline{0.8155} \tag{6.162}$$

$$\underline{UB_n^{\mu}(y_1, y_2, \ldots, y_n)} = \overline{y} - \frac{q_{\alpha/2}}{\sqrt{n}}\, s_y \tag{6.163}$$

$$= 0.8967 - \frac{(-1.984)}{\sqrt{100}} \cdot 0.0413 \tag{6.164}$$

$$= \underline{0.9444} \tag{6.165}$$

som gir realiseringen [18]

$$\underline{\underline{[\, LB_n^{\mu}\ ,\ UB_n^{\mu}\,]}} = [\, 0.8155\ ,\ 0.9444\,] \tag{6.167}$$

av det eksakte 95 %-konfidensintervallet for μ.

■

[18] Det tilsvarende asymptotiske 95 %-konfidensintervallet for μ fant i lign.(6.116) på side 331:

$$[\, LB_n^{\mu}\ ,\ UB_n^{\mu}\,] = [\, 0.8886\ ,\ 0.9048\,] \tag{6.166}$$

altså det eksakte intevallet er større enn det asymptotiske.

Eksempel: (eksakte 95%-konfidensintervall for σ av *effekten* av legemiddel)

La oss se på eksemplet med legemiddel:

$$Y \overset{\text{lign.(6.23)}}{=} \text{effekten av legemiddelet for en } \underbrace{\text{tilfeldig valgt pasient i populasjonen}}_{\text{alle som har sykdommen, ikke bare utvalget } n=100} \tag{6.168}$$

hvor $Y \sim N[\mu, \sigma]$ med:

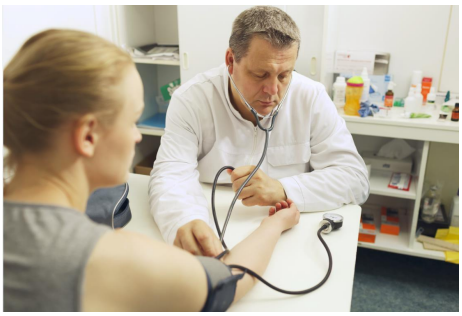
$$E[Y] = \mu \tag{6.169}$$

$$Var[Y] = \sigma^2 \tag{6.170}$$

Bruker resultatet i lign. (6.144) for å konstruere et eksakt $(1-\alpha)$-konfidensintervall for σ i eksempelet med legemiddel:

$$\frac{(n-1)S_y^2}{\sigma^2} \sim \chi^2_{n-1} \tag{6.171}$$

Foto: Colourbox



Figur 6.17: Forsøk.

Løsning:

Vi bruker χ^2_{n-1}-fordelingens α kvantiler, $\chi_{\alpha/2}$ og $\chi_{1-\alpha/2}$, for å konstruere et eksakt 95 %-konfidensintervall for σ:

$$P\left(\chi_{\alpha/2} \leq \frac{(n-1)S_y^2}{\sigma^2} \leq \chi_{1-\alpha/2} \right) = 1-\alpha \tag{6.172}$$

$$\Updownarrow \quad \text{(algebra)}$$

$$P\left(\underbrace{\sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y}_{=\, LB_n^\sigma} \leq \sigma \leq \underbrace{\sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y}_{=\, UB_n^\sigma} \right) = 1-\alpha \tag{6.173}$$

hvor vi definerer dermed nedre og øvre grenser:

$$LB_n^\sigma = \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y \tag{6.174}$$

$$UB_n^\sigma = \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y \tag{6.175}$$

med

$$\chi_{\alpha/2} = \text{nedre kvantil til } \chi^2_{n-1}\text{-fordelingen} \tag{6.176}$$

$$\chi_{1-\alpha/2} = \text{øvre kvantil til } \chi^2_{n-1}\text{-fordelingen} \tag{6.177}$$

Intervallet

$$[\, LB_n^{\sigma}\,,\, UB_n^{\sigma}\,] \;=\; \left[\, \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, S_y \;,\; \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, S_y \,\right] \tag{6.178}$$

er dermed et eksakt $(1-\alpha)\,100\,\%$-konfidensintervall for σ for effekten Y.

Med $\alpha = 0.05$ og $n = 100$ for effekten Y så finner vi ved tabelloppslag: [19]

$$\chi_{\alpha/2} \;=\; \chi_{0.025} \;=\; 73.4 \tag{6.179}$$

$$\chi_{1-\alpha/2} \;=\; \chi_{0.975} \;=\; 128.4 \tag{6.180}$$

[19] Kvantilene $\chi_{\alpha/2}$ og $\chi_{1-\alpha/2}$ er forskjellige fordi χ-fordelingen ikke er symmetrisk, se figur (6.13) side 338.

Fra dataene $y_1, y_2 \ldots y_{100}$ fra forsøksrekken (se tabell 5.2 side 269) samt $S_y^2 = 0.0017$ fra tabell 5.5 side 276 får vi en *realisering* (små $y_1, y_2, \ldots, y_{100}$ av den nedre og øvre grensen: ($n = 100$)

$$\underline{LB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\, s_y \tag{6.181}$$

$$= \sqrt{\frac{100-1}{128.4}} \cdot 0.0413 \tag{6.182}$$

$$= \underline{0.0362} \tag{6.183}$$

$$\underline{UB_n^\sigma(y_1, y_2, \ldots, y_n)} = \sqrt{\frac{n-1}{\chi_{\alpha/2}}}\, s_y \tag{6.184}$$

$$= \sqrt{\frac{100-1}{73.4}} \cdot 0.0413 \tag{6.185}$$

$$= \underline{0.0479} \tag{6.186}$$

som gir realiseringen [20]

$$\underline{\underline{[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ]}} = \underline{\underline{[\ 0.0362\ ,\ 0.0479\ ]}} \tag{6.188}$$

av det eksakte 95 %-konfidensintervallet for standardavviket σ til effekten Y.

■

[20] Det tilsvarende asymptotiske 95 %-konfidensintervallet for σ fant i lign.(6.129) på side 335:

$$[\ LB_n^\sigma\ ,\ UB_n^\sigma\ ] = [\ 0.0365\ ,\ 0.0486\ ] \tag{6.187}$$

altså det er liten forskjell mellom det asymptotiske og det eksakte konfidensintervallet. Dette skyldes at $n = 100$ er stor. I grensen når $n \to \infty$ så blir de like.

Konfidensintervall for p

Asymptotisk:

$$[\,LB_n^p\,,\,UB_n^p\,] \;=\; \left[\;\overline{X}\;-\;\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\;,\;\overline{X}\;-\;\frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X})}\;\right]$$

= [0.8163 , 0.9437] (α = 0.05)

Konfidensintervall for μ

Asymptotisk:

$$[\,LB_n^\mu\,,\,UB_n^\mu\,] \;=\; \left[\;\overline{Y}\;-\;\frac{z_{1-\alpha/2}}{\sqrt{n}}\,S_y\;,\;\overline{Y}\;-\;\frac{z_{\alpha/2}}{\sqrt{n}}\,S_y\;\right]$$

= [0.8886 , 0.9048] (α = 0.05)

Eksakt:

$$[\;LB_n^\mu\;,\;UB_n^\mu\;] \;=\; \left[\;\overline{Y}-\frac{q_{1-\alpha/2}}{\sqrt{n}}\,S_y\;,\;\overline{Y}-\frac{q_{\alpha/2}}{\sqrt{n}}\,S_y\;\right]$$

= [0.8155 , 0.9444] (α = 0.05)

Konfidensintervall for σ

Asymptotisk:

$$[\;LB_n^\sigma\;,\;UB_n^\sigma,\;] \;=\; \left[\;\sqrt{\frac{n-1}{(n-1)+z_{1-\alpha/2}\sqrt{2(n-1)}}}\,S_y\;,\;\sqrt{\frac{n-1}{(n-1)+z_{\alpha/2}\sqrt{2(n-1)}}}\,S_y\;\right]$$

= [0.0365 , 0.0486] (α = 0.05)

Eksakt:

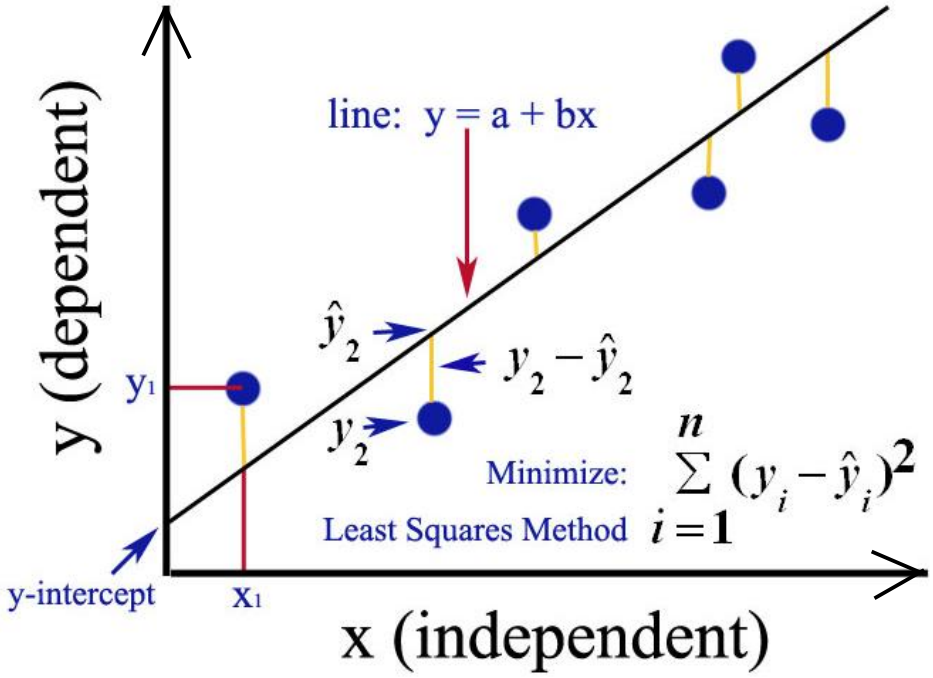
$$[\;LB_n^\sigma\;,\;UB_n^\sigma\;] \;=\; \left[\;\sqrt{\frac{n-1}{\chi_{1-\alpha/2}}}\,S_y\;,\;\sqrt{\frac{n-1}{\chi_{\alpha/2}}}\,S_y\;\right]$$

= [0.0362 , 0.0479] (α = 0.05)

Figur 6.18: Estimering og konfidensintervaller.

Kapittel 8

Regresjonsanalyse

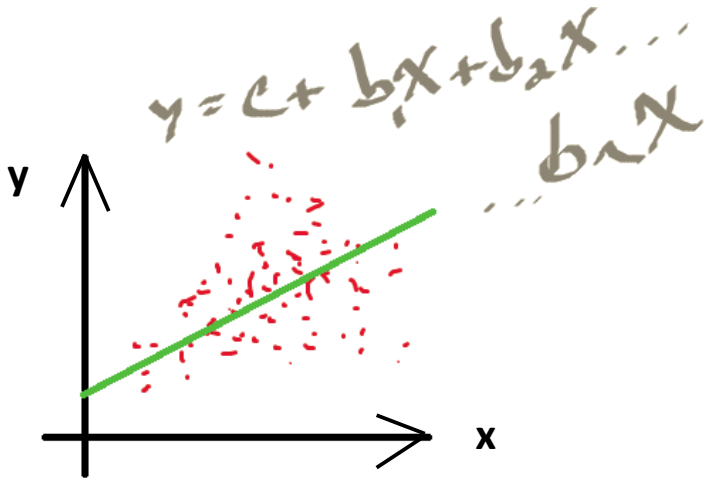


Figur 8.1: Regresjon.

8.1 Introduksjon

Regresjonsanalyse:

- Teori og metoder for å analysere og utnytte samvariasjon mellom variable.
- Formål:
  konstruere modeller som kan brukes til å anslå verdien ("prediksjon/forutsi") av en variabel Y ved hjelp av informasjon om en annen variabel X.
- terminologi:
  - variabel X: $\overbrace{\text{uavhengig variabel eller forklaringsvariabel}}^{\text{har info om dette/kjenner denne}}$
  - variabel Y: $\underbrace{\text{avhengig variabel eller responsvariabel}}_{\text{ønsker å anslå denne}}$
- Man skiller ofte mellom lineær regresjon og ikke-lineær regresjon.
- I dette kurset skal vi kun se på:
  - lineær regresjon
  - samspill mellom bare to variabler



Figur 8.2: Lineær regresjon.

8.2 Statistiske mål (to variabler)

I noen situasjoner ønsker vi å undersøke **samvariasjonen** mellom to utvalg.

Definisjon: (empirisk varians) [1]

La $x_1, x_2, x_3, ..., x_n$ være observasjoner, og la $\overline{x}$ være gjennomsnittet.
Da er den empiriske kovariansen: [2]

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{8.1}$$

■

Den empiriske kovariansen s_{xy} er et mål på graden av samvariasjon mellom størrelskene x_i og tilhørende y_i.
Men den kan være vanskelig å tolke fordi:

- vi må sammenligne med andre tall som er naturlig å sammenligne med for å kunne forstå s_{xy} bedre
- s_{xy} er enhetsavhengig og gir dermed ulikt resultat dersom vi f.eks. regner med timer, minutter eller sekunder

For å gi en mer presis tolkning av graden av **SAMVARIASJON** så går vi derfor et skritt videre og definerer korrelasjonskoeffisienten r_{xy}:

[1] Kalles også **utvalgsvariansen**.
[2] Den empiriske variansen er analog til estimatoren i lign.(6.34) på side 309.

Definisjon: (korrelasjonskoeffisient)

La $x_1, x_2, x_3, ..., x_n$ og $y_1, y_2, y_3, ..., y_n$ være observasjoner. **Korrelasjonskoeffisienten** r_{xy} er da:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} \tag{8.2}$$

■

Noen kommentarer:

- Ved å dele på s_x og s_y så får man en *normalisert* [3] versjon av s_{xy}, dvs.

$$-1 \leq r_{xy} \leq 1 \tag{8.3}$$

- r_{xy} er enhetsuavhengig
- $r_{xy} = -1$:
  - perfekt negativ korrelasjon, dvs. store x hører sammen med små y.
  - lineær[4] sammenheng mellom x og y
- $r_{xy} = 1$:
  - perfekt positiv korrelasjon, dvs. store x hører sammen med store y.
  - lineær sammenheng mellom x og y
- $r_{xy} = 0$:
  - ingen korrelasjon
  - ukorrelert
- r_{xy} er et mål på lineær korrelasjon

[3] Legg merke til begrepet *normalisert*. Det skal vi komme tilbake til ved flere anledninger.

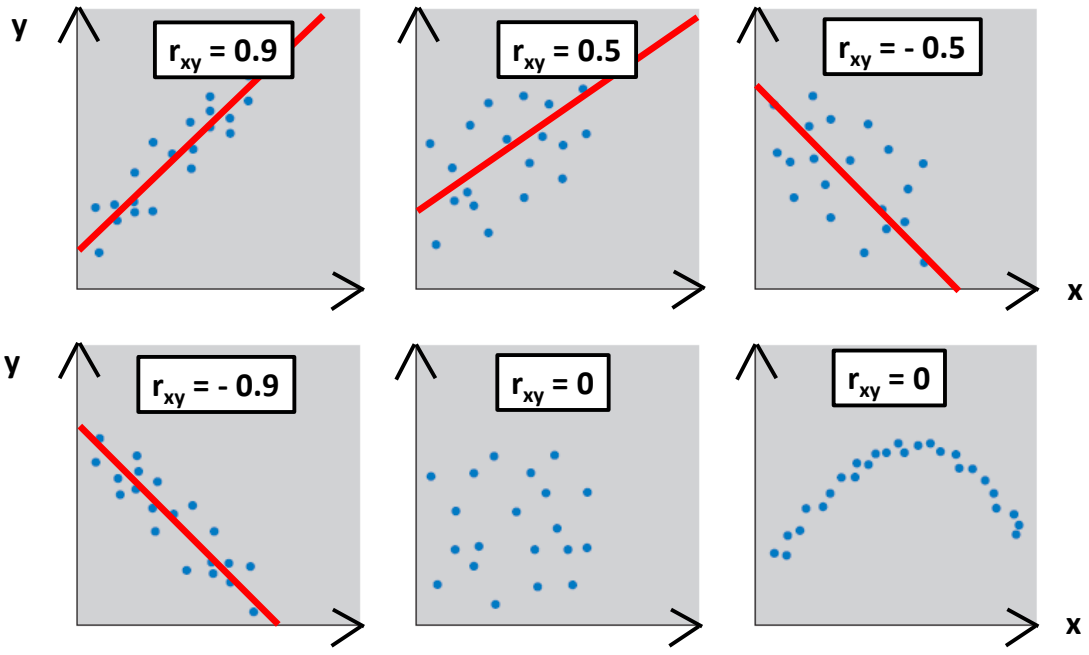
[4] Lineær sammenhenger mellom x og y betyr at de kan skrives på formen: $y = ax + b$, (a og b er konstanter). Lineær er altså det samme som en rett linje.

Eksempel: (lineær sammenheng)

La oss se nærmere på to størrelser x og y. Disse størrelsene kan være hva som helst, f.eks. pris på aksje x og pris på aksje y. Anta at disse størrelsene varierer med tiden. Anta videre at man måler x og y over en periode på 20 dager. For dag 1 er verdiene x_1 og y_1. For dag 2 har verdiene endret seg til x_2 og y_2 osv. Helt frem til dag 20 hvor størrelsene har verdiene x_{20} og y_{20}. Vi har altså samhørende **observasjoner** av par (x_i, y_i):

$$(x_1, y_1) \quad , \quad (x_2, y_2) \quad , \quad \ldots \quad , \quad (x_{19}, y_{19}) \quad , \quad (x_{20}, y_{20}) \tag{8.4}$$

La oss se på 6 forskjellige datasett som vist i figur 8.3:



Figur 8.3: Sammenheng mellom x og y samt tilhørende korrelasjonskoeffisienten r_{xy}.

Husk at r_{xy} er et mål på **lineær** sammenheng mellom x og y. For de tilfellene hvor r_{xy} er "nære" $+1$ eller -1 så er det "nære" en lineær sammenheng mellom x og y. Lineær regresjonsanalyse går ut på å:

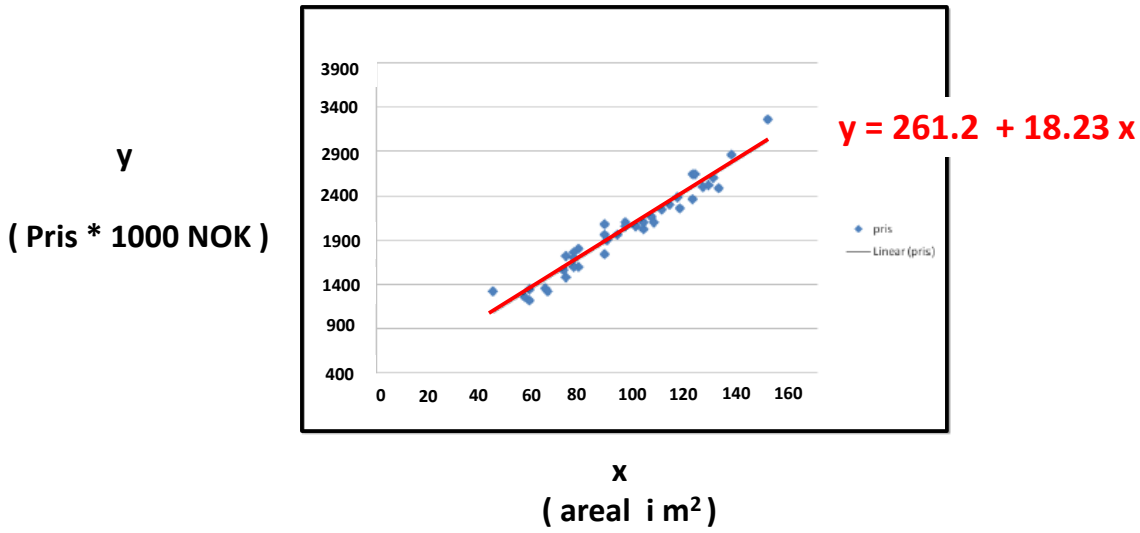
finne estimat for den rette linjen som "passer best" med datasettet

Hvorfor gir det ikke mening å finne linjen som "passer best" i de to grafene nederst til høyre i figur (8.3)?

■

Eksempel: (lineær regresjon)

La se på samvariasjon mellom utsalgspris og boareal for leiligheter i Molde. Et datasett med priser y og areal x gir spredningsplottet i figur 8.4. Vi ser en **tydelig lineær** sammenheng mellom variablene.



Figur 8.4: Sammenheng mellom y (pris) og x (areal), en regresjonslinje.

Den linjen som "passer best" kalles regresjonslinjen. For dataene i figur (8.4) er det:

$$\boxed{y = 261.2 \; + \; 18.23\, x} \qquad (8.5)$$

hvor y er pris på leilighet oppgitt i antall 1000 NOK. For en leilighet med boareal $x = 100$ m^2 vil den estimerte modellen predikere prisen:

$$\underline{\underline{y(100)}} \;=\; 261.2 \;+\; 18.23 \cdot 100 \;=\; \underline{\underline{20842}} \qquad (8.6)$$

dvs. litt under 2.1 mill.

Tallet 2.1 mill. er prediksjonen fra modellen for prisen på en 100 m^2 leilighet.

■

To spørsmål:

1) Hvordan finner man regresjonslinjen?

2) Siden modellen bare bruker arealet og ingen annen informasjon må vi vente noe feilmarginer. I hvor stor grad forklarer arealet x prisen y?

Disse spørsmålene skal vi besvare.

8.3 Teoretisk modell vs estimert modell

Det er viktig å skille mellom teoretisk modell og estimert modell (regresjonslinje).

Teoretisk modell:

En teoretisk modell beskriver hvordan vi tenker oss den virkelige sammenhengen mellom variablene, typisk:

$$\underbrace{y = a + bx + e}_{\text{eksakt}} \tag{8.7}$$

hvor variabelen e beskriver "avviket"[5] fra den eksakte lineære funksjonen.

Estimert modell: (med "hatt")$^{\text{passer best}}$

Parametrene a og b i lign.(8.7) er ukjente.
Men de kan estimeres ut fra et datasett med samhørende observasjoner av par:

$$\underbrace{(x_1, y_2)\ ,\ (x_2, y_2)\ ,\ \ldots\ ,\ (x_n, y_n)}_{\text{observasjoner}}\ . \tag{8.8}$$

Vi ønsker å bestemme estimater for a og b.
Disse estimatene har en "hatt" på seg, $\hat{a}$ og $\hat{b}$.
Linjen med estimatene $\hat{a}$ og $\hat{b}$ kalles **regresjonslinjen**:

$$\underbrace{\hat{y} = \hat{a} + \hat{b}x}_{\text{regresjonslinje (passer best)}} \tag{8.9}$$

Estimatene $\hat{a}$ og $\hat{b}$ bestemmes ved å finne den linje som "passer best" med datasettet.

[5] "e" står for "error".

8.4 Residual og *sse*

1) Observasjoner:

De røde punktene i figur 8.5 viser de fem observasjonspunktene (x_1, y_1), (x_2, y_2), ... , (x_5, y_5).

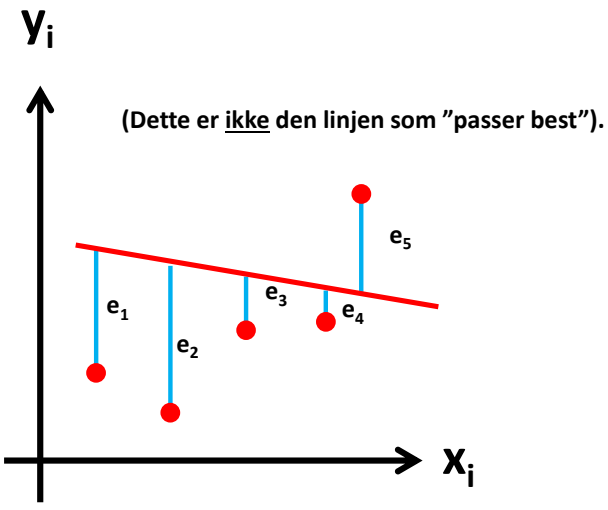
2) Rett linje:

Den røde linjen i figur 8.5 er en linje som ikke passer best til observasjonene x_i
og tilhørende verdi y_i.

Forskjellen mellom de observerte verdiene y_i og de tilsvarende verdiene til den rette linjen er de loddrette avstandene (blå linjer) som vist i figur 8.5. Forskjellen/avviket mellom observert verdi og prediksjonen som den rette linjen foreslår for datapunktet er:

$$\underbrace{e_i \;=\; y_i - \hat{y}_i}_{\text{residual}} \qquad (8.10)$$

og kalles residual eller estimat for eksperimentfeilen.
Residualen e_i måler dermed feilen vi gjør ved å bruke verdien på den rette linjen istedet for de observerte verdiene. Residualen e_i kan være positiv, negativ eller 0. [6]



Figur 8.5: Residual.

[6] Ingen residual i figur 8.5 er null. Alle er negative bortsett fra e_5.

Residualen $e_i = y_i - \hat{y}_i$:

- e_i kan være positiv, negativ eller 0
- e_i = avvik mellom estimert linje og observerte y_i-verdier.
- e_i = residual nr. i

Definisjon: (sse) [7]

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Størrelsen *sse*, "*sum squared error*", er da definert ved: [8]

$$\boxed{sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (8.12)$$

hvor [9]

$$\hat{y}_i = \text{estimat/prediksjon for } y_i \qquad (8.14)$$

$$y_i = \text{faktiske observasjonene/dataene nr. } i \qquad (8.15)$$

■

[7] *sse* står for sum square error.

[8] Siden $e_i \overset{\text{lign.}(8.10)}{=} y_i - \hat{y}_i$ så kan *sse* alternativt skrives:

$$\boxed{sse = \sum_{i=1}^{n} e_i^2} \qquad (8.11)$$

[9] I vårt kurs dreier prediksjonene gitt som linjen

$$\underbrace{\hat{y}_i = \hat{a} + \hat{b}x_i}_{\text{prediksjoner}} \qquad (8.13)$$

hvor $\hat{a}$ og $\hat{b}$ er det estimatene som gir en linje som "passer best" med observasjonene. Vi skal finne uttrykk for disse optimale $\hat{a}$ og $\hat{b}$ i neste avsnitt.

8.5 Minste kvadraters regresjonslinje

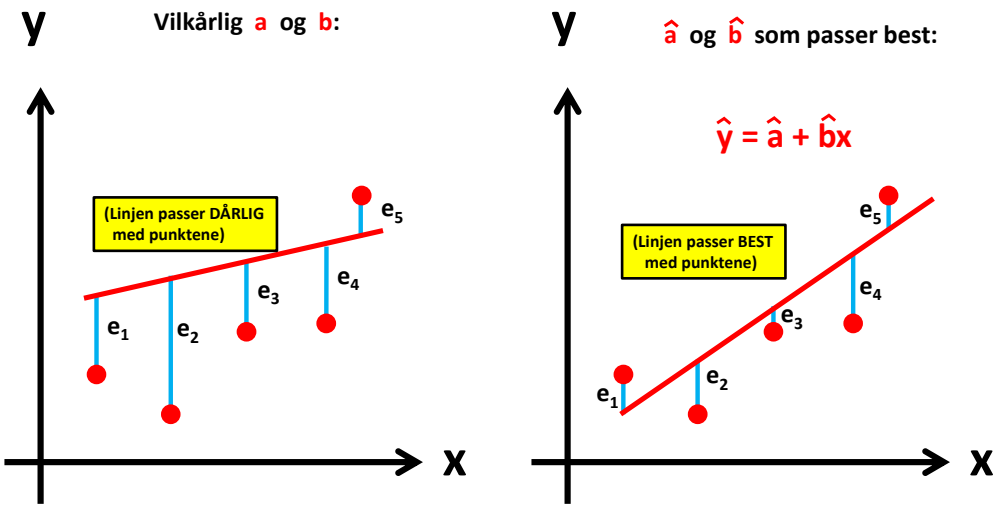
Linjen på venstre side i figur 8.6 er åpenbart ikke den som "passer best".
Linjen på høyre, derimot, passer bedre. Vi ønsker nå å finne den linjen som "passer best". Med det mener vi:

Passer best:	*sse* er minst mulig.

Å finne den linjen som passer best med datasettet er det samme som å finne den a og b som gir minst SSE, dvs. minst "*sum squared error*" i forhold til datasettet. SSE er minst der hvor stigningen er null, dvs. den deriverte med hensyn på de respektive parametrene, er lik null: [10] [11]

$$\frac{\partial sse}{\partial a} = \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}\Big(y_i-(a+bx_i)\Big)^2\right) = 2(-1)\sum_{i=1}^{n}\Big(y_i-a-bx_i\Big) = 0 \qquad (8.16)$$

$$\frac{\partial sse}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}\Big(y_i-(a+bx_i)\Big)^2\right) = 2(-1)\sum_{i=1}^{n}\Big(y_i-a-bx_i\Big)x_i = 0 \qquad (8.17)$$



Figur 8.6: Linje som ikke passer best, og linje som passer best.

[10] Bruker kjerneregelen som vi lærte om i "*MAT100 Matematikk*".

[11] At lign.(8.16) er et minimum, og ikke et maksimun, innser man siden $\frac{\partial^2 SSE}{\partial a^2} > 0$ og $\frac{\partial^2 SSE}{\partial a^2} > 0$.

De spesielle verdiene for a og b som minimerer SSE har fått egen notasjon, $\hat{a}$ og $\hat{b}$.
Disse er definert ved lign.(8.16). Eksplisitte uttrykk for disse finnes ved å løse nevnte ligning:

Siden gjennomsnittet $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gir $n\overline{x} = \sum_{i=1}^{n} x_i$, og tilsvarende for y, så fås:

$$\overbrace{\sum_{i=1}^{n} y_i}^{=n\overline{y}} - \hat{a}\overbrace{\sum_{i=1}^{n}}^{=n} - \hat{b}\overbrace{\sum_{i=1}^{n} x_i}^{=n\overline{x}} = 0 \tag{8.18}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\underbrace{\sum_{i=1}^{n} x_i}_{=n\overline{x}} - \hat{b}\sum_{i=1}^{n} x_i^2 = 0 \tag{8.19}$$

og

$$\cancel{n}\overline{y} - \hat{a}\cancel{n} - \hat{b}\cancel{n}\overline{x} = 0 \tag{8.20}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\,n\,\overline{x} - \hat{b}\sum_{i=1}^{n} x_i^2 = 0 \tag{8.21}$$

Løser med hensyn på $\hat{a}$ og $\hat{b}$ alene gir:

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} \tag{8.22}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{8.23}$$

Disse ligningene kan skrives ved hjelp av den empiriske variansen, lign.(5.23), og den empiriske kovariansen, lign.(8.1) på følgende måte: [12]

$$\boxed{\begin{aligned} \hat{a} &= \overline{y} - \hat{b}\,\overline{x} && (8.24)\\ \hat{b} &= \frac{s_{xy}}{s_x^2} && (8.25)\end{aligned}}$$

hvor

$$s_x^2 \overset{\text{lign.(5.23)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{8.26}$$

$$s_{xy} \overset{\text{lign.(8.1)}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{8.27}$$

[12] Man må utføre et par linjer med algebra for å innse at lign.(8.24) gir lign.(8.26). Detaljene er ikke tatt med her. Men kanskje du greier seg selv?

Setning: (minste kvadraters sum - lineære regresjonslinje)

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Minste kvadraters sum gir den lineære regresjonslinjen: q[13]

$$\boxed{\hat{y} \;=\; \hat{a} \;+\; \hat{b}\, x\;,} \qquad (8.28)$$

hvor

$$\boxed{\begin{aligned} \hat{b} \;&=\; \frac{s_{xy}}{s_x^2} && (8.29)\\ \hat{a} \;&=\; \overline{y} \;-\; \hat{b}\,\overline{x} && (8.30) \end{aligned}}$$

og

$$s_x^2 \overset{\text{lign.}(5.23)}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (8.31)$$

$$s_{xy} \overset{\text{lign.}(8.1)}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \qquad (8.32)$$

og hvor $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ og $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

■

[13] Den linjen som passer best, dvs. minst *sse*, har altså fått navnet regresjonslinje.

Eksempel: (transport - s_{xy} , regresjonslinje)

Et transportfirma har et varemottak for vogntog med spesialgods. Det tar svært lang tid å laste av et vogntog med denne type last. Transportfirmaet gjør derfor én måling per dag i en periode på 10 dager: når et tilfeldig vogntog ankommer varemottaket så teller de antall vogntog x som står foran i kø. I tillegg så måler de ventetiden y for det nylig ankomne vogntoget.

La oss definere følgende variabler:

$$x = \text{antall vogntog foran i kø} \tag{8.33}$$
$$y = \text{antall timer i ventetid} \tag{8.34}$$

For enkelhetsskyld så måler de y kun i hele timer. Resultatet er:

x_i (antall vogntog foran i kø)	2	12	1	1	10	25	3	9	27	2
y_i (antall timer ventetid)	3	11	3	1	12	21	6	4	31	2

Figur 8.7: Observasjoner x_i og y_i, hvor $i = 1, 2, ..., 10$.

Figur 8.8: Vogntog og varemottak.

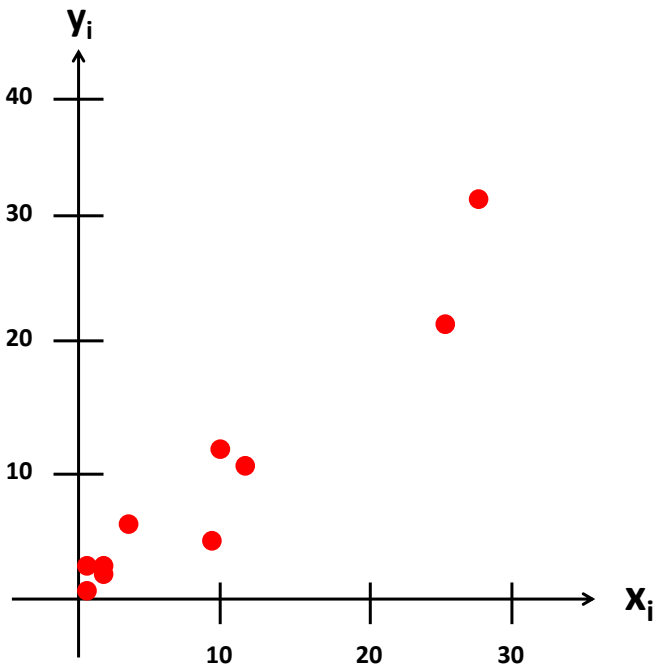
a) Gi en grafisk fremstilling av x_i og y_i. Kommenter svaret.

b) Finn gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$.

c) Finn kovariansen s_{xy}.

d) Finn minste kvadraters lineæregresjonslinje for x og y.

a) Grafisk fremstilling av x_i og y_i:



Figur 8.9: Grafisk fremstilling av x_i og y_i, hvor $i = 1, 2, ..., 10$.

Kommentar:

Fra denne grafen ser vi at:

- små x_i-verdier faller sammen med små y_i-verdier
- store x_i-verdier faller sammen med store y_i-verdier

Dermed innser vi at det er en viss grad av **samsvar** mellom x_i og y_i.

b) Gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$ er: ($n = 10$)

$$\underline{\underline{\overline{x}}} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\Big(2 + 12 + 1 + ... + 2\Big) = \underline{\underline{9.2}} \quad \text{(antall vogntog i kø)} \tag{8.35}$$

$$\underline{\underline{\overline{y}}} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}\Big(3 + 11 + 3 + ... + 2\Big) = \underline{\underline{9.4}} \quad \text{(ventetid, i timer)} \tag{8.36}$$

c) Vi bruker gjennomsnittsverdiene $\overline{x}$ og $\overline{y}$ når vi skal finne **kovariansen** s_{xy}: ($n = 10$)

$$\underline{\underline{s_{xy}}} \overset{\text{lign.(8.1)}}{\underset{\text{n=10}}{=}} \frac{1}{10-1}\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) \tag{8.37}$$

$$= \frac{1}{10-1}\Big[(2-9.2)(3-9.4) + (12-9.2)(11-9.4) + ... + (2-9.2)(2-9.4)\Big]$$

$$= \underline{\underline{90.8}} \tag{8.38}$$

d) Den empiriske variansen s_x^2 kan regnes ut via definisjonen i lign.(5.23):

$$s_x^2 \overset{\text{lign.(5.23)}}{\approx} 94.6 \tag{8.39}$$

I oppgave **c** regnet vi ut kovariansen, se lign.(8.38):

$$s_{xy} \overset{\text{lign.(8.38)}}{=} 90.8 \tag{8.40}$$

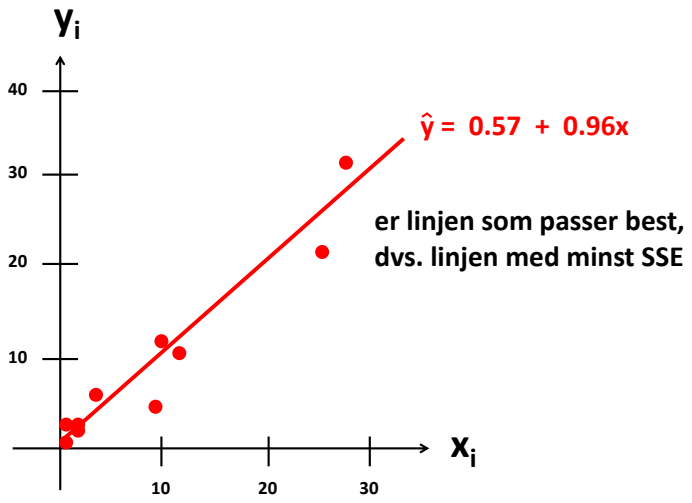
Parametrene $\hat{b}$ og $\hat{a}$ er da:

$$\underline{\hat{b}} = \frac{s_{xy}}{s_x^2} = \frac{90.8}{94.6} \approx \underline{0.96} \tag{8.41}$$

$$\underline{\hat{a}} = \overline{y} - b\overline{x} = 9.4 - 0.96 \cdot 9.2 \approx \underline{0.57} \tag{8.42}$$

Minste kvadraters lineære regresjonslinje blir dermed ifølge lign.(8.28):

$$\underline{\underline{\hat{y} = 0.57 + 0.96\,x}} \tag{8.43}$$



Figur 8.10: Grafisk fremstilling av x_i og y_i.

8.6 Forklaringsfraft og *sst*

På side 384, lærte vi at "problemet" med kovariansen s_{xy} er at den kan være vanskelig å tolke. Dette bl.a. fordi:

- størrelsen s_{xy} kan gi "store" eller "små" tall som vi må sammenligne med andre tall for å kunne forstå bedre
- s_{xy} er enhetsavhengig og gir dermed ulikt resultat dersom vi f.eks. regner med timer, minutter eller sekunder

Dette problemet ble løst ved å introdusere korrelasjonskoeffisienten r_{xy}. Denne koeffisienten har egenskaper som oppsummert på side 384.

Samme "problem" har *sse* i lign. (8.12). Hva betyr det at *sse* er "liten"? Eller "stor"? I forhold til hva? Vi løser dette problemet ved å introdusere "forklaringskraften" r^2. Før vi definerer r^2 må vi først definere en ny størrelse, nemlig *sst*:

Definisjon: (sst) [14]

La $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ være observasjonspar/datasett.
Størrelsen sst, "*sum squared total*", er da definert ved:

$$\boxed{sst \;=\; \sum_{i=1}^{n} (y_i - \overline{y})^2} \qquad (8.44)$$

$$y_i \qquad = \qquad \text{faktiske observasjonene/dataene nr. } i \qquad (8.45)$$

$$\overline{y} \overset{\text{lign.(5.21)}}{=} \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (8.46)$$

■

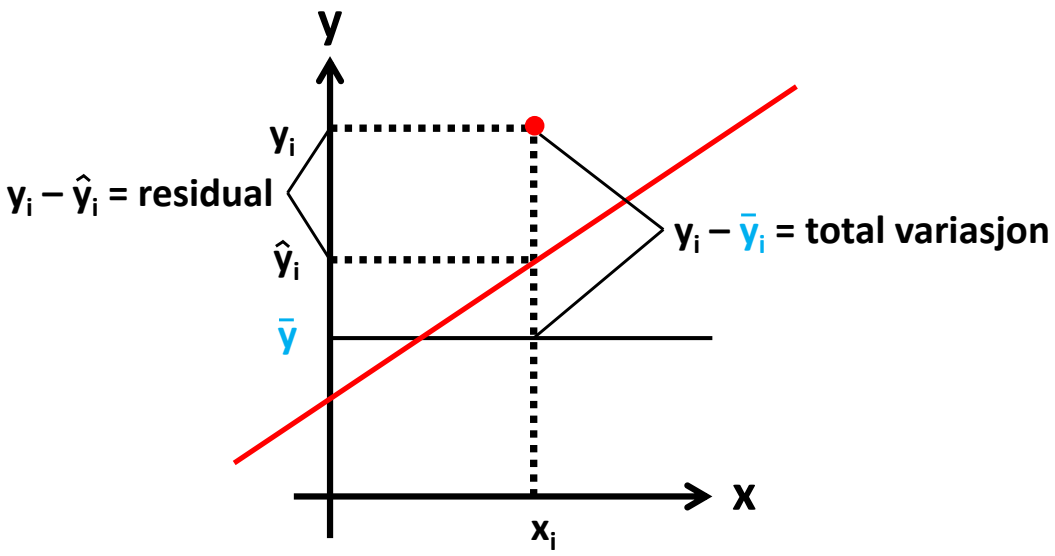
Sammenlign definisjonen ovenfor med sse fra lign.(8.12):

$$sse \;=\; \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (8.47)$$

Ser du forskjellen på lign.(8.47) og (8.44)?

[14] sst står for sum square total.

Visualisering av residual og total variasjon:



Figur 8.11: Residual og total variasjon.

Definisjon: (forklaringskraft)[15]

La $x_1, x_2, x_3, ..., x_n$ og $y_1, y_2, y_3, ..., y_n$ være observasjoner.
Forklaringskraft r^2 er da:

$$\boxed{r^2 = 1 - \frac{sse}{sst}} \tag{8.48}$$

hvor

$$sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.49}$$

$$sst = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{8.50}$$

og

$$y_i = \text{faktiske observasjonene/dataene nr. } i \tag{8.51}$$

$$\hat{y}_i = \text{estimat/prediksjon for } y_i \tag{8.52}$$

$$\overline{y} \overset{\text{lign.(5.21)}}{=} \frac{1}{n} \sum_{i=1}^{n} y_i \tag{8.53}$$

■

[15] Kallse også forklarings*grad* eller forklarings*styrke*.

Noen kommentarer: [16]

- Forklaringskraften r^2 er *normalisert*:

$$\boxed{0 \leq r^2 \leq 1} \qquad (8.54)$$

- r^2 er enhetsuavhengig
- $r^2 = 1$:
  - alle observerte punkter ligger på regresjonslinjen
- $r^2 = 0$:
  - verdien for x har ingen betydning for verdien av y
- r^2 sier noe om:
  - hvor stor andel av den totale vasiasjonen som forklares av regresjonslinjen

Forklaringskraft r^2 og korrelasjonskoeffisienten r_{xy} som vi lærte om på side 384 har analoge egenskaper.

[16] r_{xy} og r^2 har ikke noe med hverandre å gjøre. Likevel kan det være hensiktsmessig å sammenligne egenskapene til r^2 som oppsummert her, med egenskapene til r_{xy} på side 384.

Eksempel: (logistikk - forklaringskraft)

Finn forklaringskraften r^2 for eksemplet fra side 396.

Løsning: (forklaringskraft , logistikk)

La oss se atter en gang se på eksemplet fra side 396.

x_i (antall vogntog foran i kø)	2	12	1	1	10	25	3	9	27	2
y_i (antall timer ventetid)	3	11	3	1	12	21	6	4	31	2

Figur 8.12: Samsvarende verdier av antall vogntog foran i kø x og antall timer ventetid y.

Fra lign.(8.43) vet vi:

$$\hat{y}_i \;=\; 0.57 \;+\; 0.96\, x_i \tag{8.55}$$

De faktiske verdiene y_i er vet tabellen i figur (8.12). Dermed kan regne ut sse fra lign. (8.12):

$$\begin{aligned}
\underline{sse} \quad &= \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \quad \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \quad (3 - 0.57 - 0.96 \cdot 2)^2 \quad + (11 - 0.57 - 0.96 \cdot 12)^2 \\
&+ \quad (3 - 0.57 - 0.96 \cdot 1)^2 \quad + (1 - 0.57 - 0.96 \cdot 1)^2 \\
&+ \quad (12 - 0.57 - 0.96 \cdot 10)^2 + (21 - 0.57 - 0.96 \cdot 25)^2 \\
&+ \quad (6 - 0.57 - 0.96 \cdot 3)^2 \quad + (4 - 0.57 - 0.96 \cdot 9)^2 \\
&+ \quad (31 - 0.57 - 0.96 \cdot 27)^2 + (2 - 0.57 - 0.96 \cdot 2)^2 \\
&\approx \quad \underline{74.2}
\end{aligned} \tag{8.56}$$

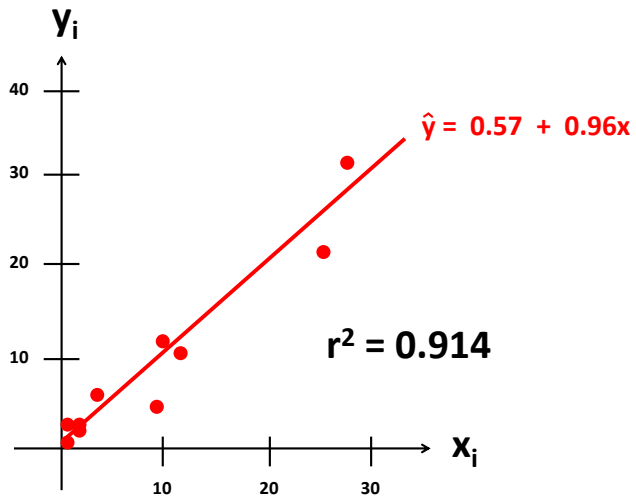
Fra lign.(8.36) vet vi gjennomsnittet $\overline{y} = 9.4$.
Dermed kan regne ut sse fra lign. (8.44):

$$\begin{aligned} \underline{sst} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ &= (3-9.4)^2 + (11-9.4)^2 \\ &+ (3-9.4)^2 + (1-9.4)^2 \\ &+ (12-9.4)^2 + (21-9.4)^2 \\ &+ (6-9.4)^2 + (4-9.4)^2 \\ &+ (31-9.4)^2 + (2-9.4)^2 \\ &= \underline{858.4} \end{aligned} \tag{8.57}$$

Forklaringskraft R^2 er da: (se lign.(8.48)

$$\underline{\underline{r^2}} = 1 - \frac{sse}{sst} = 1 - \frac{74.2}{858.4} = \underline{\underline{0.914}} \tag{8.58}$$

dvs. regresjonslinjen forklarer hele 91.4 % av den totale variasjonen.
Vi sier at modellen har <u>stor forklaringskraft</u>.

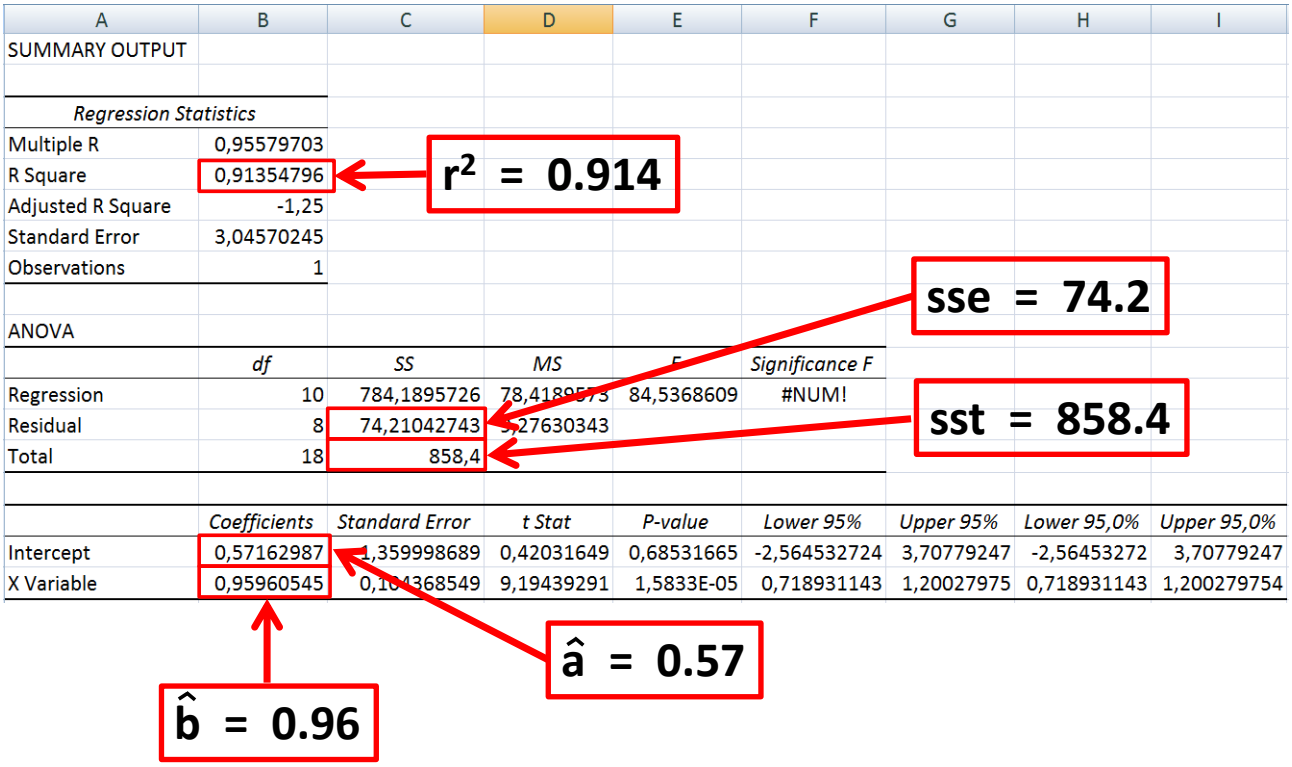


Figur 8.13: Grafisk fremstilling av x og y.

Kommentar:

Som regel gjør man ikke all denne regningen "for hånd".
Mange dataprogrammer kan hjelpe oss å regne ut statistiske størrelser, f.eks. Excel.

Dersom vi bruker Excel på eksemplet vårt så får vi en utskrift som vist i figuren nedenfor. Da kan vi lese av de størrelsene vi regnet ut "for hånd" direkte fra Excel-utskriften.



Figur 8.14: Utskrift fra Excel.

■