Ny og bedre versjon

MAT110

Statistikk 1

Formelsamling 2019

Per Kristian Rekdal

Bård-Inge Pettersen

Høgskolen i Molde
Vitenskapelig høgskole i logistikk

Innhold

Forord

Dette er formelsamlingen i emnet "*MAT110 Statistikk 1*" for 2019 ved Høgskolen i Molde.

Hjelpemidler eksamen:

Godkjent kalkulator og formelsamling

- Kun originalversjonen av formelsamlingen utgitt av "*SiMolde Bok*" er lov å ha med på eksamen. (Dette fordi det skal være lett å se at dere har med den riktige og lovlige formelsamlingen på eksamen).

- Det er lov å skrive egne notater i formelsamlingen som dere kan ta med på eksamen. Men: Ikke skriv av hele eksempler og hele oppgaver. (Dersom dette blir praktisert i stor grad må vi revurdere denne ordningen i forhold til neste års studenter).

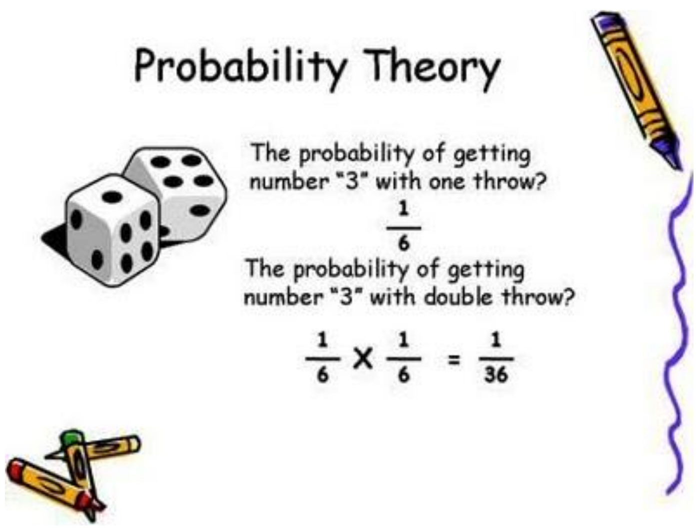
Per Kr. Rekdal

Per Kristian Rekdal

Bård-Inge Pettersen

Kapittel 1

Sannsynlighetsteori



Figur 1.1: Sannsynlighetsteori.

1.1 Sannsynlighetsmodell

1.1.1 Utfallsrommet

Definisjon: (utfallsrom)

Resultatet av et gitt eksperiment E har et visst antall *mulige utfall*.
Denne mengden av mulige utfall kalles utfallsrommet til E: [1]

$$\boxed{\Omega_E = \left\{ \text{ mengden av } \underline{\text{alle}} \text{ mulige utfall ved eksperiment E } \right\} \qquad (1.1)}$$

Dersom det er underforstått hvilket eksperiment det er snakk om, brukes bare Ω.

■

[1] Den greske bokstaven Ω kalles "omega".

1.1.2 Mengdelære

Siden en definerer de mulige utfallene av et eksperiment som en *mengde*, er det viktig med en liten introduksjon til mengdelære.

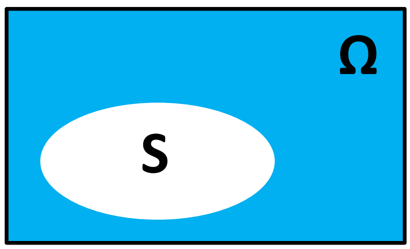
Se tillegg A.

1.1.3 Begivenheter - delmengder av utfallsrommet

Definisjon: (begivenhet) [2] [3]

$$\boxed{\text{begivenhet } S \quad = \quad \text{en } \textit{mulig} \text{ delmengde } S \text{ av utfallsrommet } \Omega, \text{ dvs. } S \subseteq \Omega} \tag{1.2}$$

■



Figur 1.2: Begivenhet S.

[2] Noen bruker også "hendelse" istedet for "begivenhet". I dette kurset brukes samme konvensjon som i læreboken, dvs. "begivenhet".

[3] Definisjonen av en begivenhet har ordet "mulig" i seg fordi det viser seg at ved når utfallsrommet f.eks. er alle tall større enn 0 (f.eks. temperatur), finnes det delmengder hvor sannsynligheten ikke er vel-definert. Begivenheter antas derfor alltid å ha en veldefinert sannsynlighet. I det diskrete tilfellet kan sannsynligheten til enhver delmengde evalueres.

1.1.4 Sannsynlighetsmodell

Definisjon: ($\overbrace{\text{sannsynlighetsmodell} \quad - \quad \text{3 aksiomer}}^{\text{fundamentale egenskaper}}$)

En *sannsynlighetsmodell* for et eksperiment E består av:

1) et utfallsrom Ω_E (1.3)

2) en sannsynlighetslov $P(-)$ (1.4)
som for hver begivenhet $S \subseteq \Omega$ tilordner en sannsynliget $P(S) \in [0, 1]$

hvor sannsynlighetsloven P har følgende 3 fundamentale egenskaper: (3 aksiomer)

$$\boxed{\begin{array}{rcll} P(S) & \geq & 0 & \text{(ikke-negativitet)} \quad (1.5) \\ P(S_1 \cup S_2 \cup ...) & = & P(S_1) + P(S_2) + ... & \text{(addisjon)} \quad (1.6) \\ P(\Omega) & = & 1 & \text{(normalisering)} \quad (1.7) \end{array}}$$

hvor, lign.(1.6), gjelder for alle parvis disjunkte begivenheter S_1, S_2, $S_3, \ldots$

■

1.1.5 Fundamentale setninger i sannsynlighetsteori

Setning 1: (sannsynligheten av den tomme begivenheten er null)

Sannsynligheten for den tomme begivenheten $\emptyset$ er lik null, dvs.:

$$\boxed{P(\emptyset) \;=\; 0 \qquad\qquad (1.8)}$$

■

Setning 2: (den spesielle addisjonssetningen) [4]

Dersom begivenhetene A og B er disjunkte, dvs. $A \overbrace{\cap}^{\text{og}} B = \varnothing$, så gjelder:

$$\boxed{P(A \overbrace{\cup}^{\text{eller}} B) \;=\; P(A) + P(B) \qquad\qquad (1.9)}$$

■

[4]Den spesielle addisjonssetningen kan åpenbart utvides til et endelig antall disjunkte begivenheter.

Setning 3: (differansen mellom to begivenheter)

Anta en har to begivenheter A og B slik at $B \subset A$, da gjelder

$$\boxed{P(A - B) \;=\; P(A) - P(B) \qquad (1.10)}$$

■

Setning 4: (den generelle addisjonssetningen)

For begivenhetene A og B gjelder:

$$\boxed{P(A \overbrace{\cup}^{\text{eller}} B) \;=\; P(A) + P(B) \;-\; \underbrace{P(A \overbrace{\cap}^{\text{og}} B)}_{\text{ekstra ledd}} \qquad (1.11)}$$

Setning 5: (komplementærsetningen)

La A være en vilkårlig begivenhet, da gjelder alltid:

$$\boxed{P(\overline{A}) \;=\; 1 - P(A) \qquad (1.12)}$$

■

Setning 6: (total sannsynlighet)

For begivenhetene A og B gjelder:

$$\boxed{P(A) \;=\; P(A \cap \overline{B}) \;+\; P(A \cap B) \qquad (1.13)}$$

■

Setning 7: (tvillingsetningen) [5]

For begivenhetene A og B gjelder:

$$\boxed{P(\overline{A} \overbrace{\cap}^{\text{og}} \overline{B}) \;=\; 1 \;-\; P(A \overbrace{\cup}^{\text{eller}} B) \qquad (1.15)}$$

■

[5]Hvorfor er "tvillingsetningen" et hensiktsmessig navn for lign.(1.15)? Dersom en bytter A med $\overline{A}$ og B med $\overline{B}$ i lign. (1.15), fås ligningen:

$$\boxed{P(A \overbrace{\cap}^{\text{og}} B) \;=\; 1 \;-\; P(\overline{A} \overbrace{\cup}^{\text{eller}} \overline{B}) \qquad (1.14)}$$

som kan anses som '"tvillingsetningen" til lign. (1.15).

1.1.6 Diskret sannsynlighetsmodell

Definisjon: (diskret sannsynlighetsmodell)

En *sannsynlighetsmodell* $\big(\Omega, P(-)\big)$ for et eksperiment hvor utfallsrommet er *endelig*, dvs.:

$$\Omega \ = \ \Big\{ \ u_1 \ , \ u_2 \ , \ \ldots \ , \ u_n \ \Big\} \tag{1.16}$$

kalles en diskret sannsynlighetsmodell.

$\blacksquare$

Anta sannsynlighetene til de n utfallene er gitt ved:

$$P(u_i) \;=\; p_i, \qquad\qquad i = 1, 2, \ldots, n \tag{1.17}$$

Ved aksiomene om ikke-negativitet og normalisering, dvs. lign. (1.5) og (1.7), følger umiddelbart at $p_i \geq 0$ for alle $i = 1, 2, \ldots, n$ og

$$\underline{\underline{P(\Omega)}} \;=\; P\big(\{\; u_1,\; u_2,\; \ldots,\; u_n\}\big) \;=\; \sum_{i=1}^{n} P(u_i) \;=\; \sum_{i=1}^{n} p_i \;=\; \underline{\underline{1}} \tag{1.18}$$

Siden Ω har n utfall, finnes det totalt 2^n begivenheter. La S være en vilkårlig begivenhet med totalt $k \leq n$ utfall:

$$S \;=\; \Big\{\; u_{j_1} \;,\; u_{j_2} \;,\; \ldots \;,\; u_{j_k} \;\Big\} \tag{1.19}$$

Fra den spesielle addisjonssetningen (1.9) følger da umiddelbart at:

$$\boxed{P(S) \;=\; \sum_{i=1}^{k} p_{j_i} \qquad\qquad (1.20)}$$

For en diskret sannsynlighetsmodell er dermed sannsynlighetsloven fullstendig bestemt fra utfallssannsynlighetene p_i.

1.1.7 Uniform sannsynlighetsmodell

Definisjon: (uniform sannsynlighetsmodell)

En diskret sannsynlighetsmodell $\big(\Omega, P(-)\big)$ hvor :

$$\Omega \;=\; \Big\{\; u_1 \;,\; u_2 \;,\; \ldots,\; u_n \;\Big\} \tag{1.21}$$

kalles **uniform** dersom alle utfallene er like sannsynlige, dvs:

$$\boxed{P(u_i) \;=\; p_i \;=\; \frac{1}{n} \qquad (1.22)}$$

hvor $i = 1, 2, \ldots, n$

■

La S være en vilkårlig begivenhet med totalt $k \leq n$ utfall:

$$S \;=\; \Big\{\; u_{j_1} \;,\; u_{j_2} \;,\; \ldots \;,\; u_{j_k} \Big\} \tag{1.23}$$

Lign. (1.20) for diskrete sannsynlighetsmodeller gir da at

$$\boxed{P(S) \;=\; \sum_{i=1}^{k} p_{j_i} \;=\; \sum_{i=1}^{k} \frac{1}{n} \;=\; \frac{k}{n} \qquad (1.24)}$$

1.1.8 Kombinatorikk

For $\overbrace{\text{uniforme}}^{\text{NB!}}$ sannsynlighetsmodeller handler beregning av sannsynligheter for en begivenhet egentlig om *telling*, dvs. fagfeltet *kombinatorikk*

Se tillegg B.

1.2 Betinget sannsynlighet

Definisjon: (betinget sannsynlighet)

Den betingede sannsynligheten er definert ved:

$$\boxed{P(A|B) \;=\; \frac{P(A\cap B)}{P(B)} \qquad (1.25)}$$

hvor

$$
\begin{aligned}
P(A|B) &= \text{sannsynligheten for } A \text{ gitt } B \text{ (} B \text{ allerede har inntruffet)} && (1.26)\\
&\quad \text{(\textit{betinget} sannsynlighet)} \\
P(A\cap B) &= \text{“og”-sannsynligheten for } A \text{ og } B \\
P(B) &= \text{sannsynligheten for } B \quad \neq\ 0 && (1.27)\\
&\quad \text{(\underline{u}\textit{betinget} sannsynlighet)}
\end{aligned}
$$

■

1.2.1 Multiplikasjonssetningen

Setning: (*multiplikasjons*setningen, generelle)

For begivenhetene A og B gjelder:

$$\boxed{P(A \overbrace{\cap}^{\text{og}} B) \;=\; P(A|B)\,P(B) \qquad (1.28)}$$

■

1.2.2 Uavhengighet

Definisjon: (uavhengighet)

To begivenheter A og B er uavhengige dersom

$$\boxed{P(A|B) \;=\; P(A) \qquad (1.29)}$$

■

Setning: (multiplikasjonssetningen, spesielle)

Dersom begivenhetene A og B er uavhengige, så gjelder:

$$\boxed{P(A \overbrace{\cap}^{\text{og}} B) \;\stackrel{\text{uavh.}}{=}\; P(A) \cdot P(B) \qquad (1.30)}$$

■

1.2.3 Bayes lov

Setning: (*Bayes* lov)

For begivenhetene A og B gjelder:

$$\boxed{P(A|B) \;=\; P(B|A) \cdot \frac{P(A)}{P(B)} \qquad (1.31)}$$

■

1.2.4 Oppsplitting av Ω

Setning: (oppsplitting av Ω i 2)

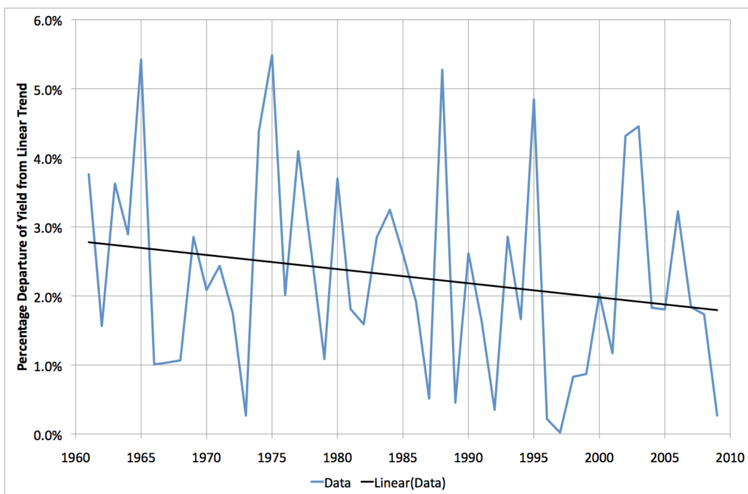
For betingelsene A og B gjelder at:

$$\boxed{P(A) \;=\; P(A|B_1) \cdot P(B_1) \;+\; P(A|B_2) \cdot P(B_2) \qquad (1.32)}$$

■

Kapittel 2

Stokastiske variabler, forventning og varians



Figur 2.1: Forventning og varians.

2.1 Stokastiske variabler

Definisjon: (stokastisk variabel) [1]

En $\underbrace{\text{stokastisk}}_{\text{tilfeldig}}$ variabel er en størrelse X som kan anta ulike verdier x med ulike sannsynligheter.

■

2.1.1 Diskret VS kontinuerlig stokastiske variabler

Definisjon: ($\overbrace{\text{diskret}}^{\text{"hakkete"}}$ stokastisk variabel)

En diskret stokastisk variabel er en størrelse X som kan ha et endelig antall eller et tellbart uendelig antall mulige ulike verdier x med ulike sannsynligheter.

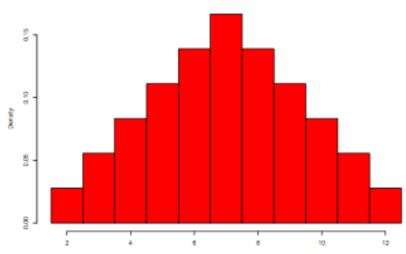
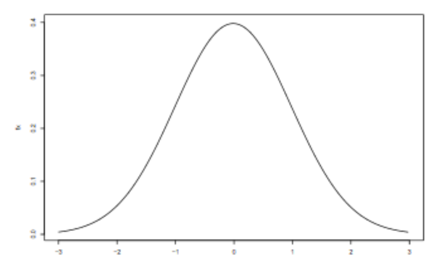
■

Definisjon: ($\overbrace{\text{kontinuerlig}}^{\text{"glatt"}}$ stokastisk variabel)

En kontinuerlig stokastisk variabel er en størrelse X som kan ha alle verdier i et intervall på tallinjen eller hele tallinjen.

■

[1] En mer teknisk (matematisk) versjon av definisjonen av en stokastisk variabel er: Med en tilfeldig variabel mener vi en funksjon X som til ethvert mulig utfall definerer et bestemt reelt tall.

Diskret stokastisk variabel	**Kontinuerlig stokastisk variabel**
X Mulige verdier x: Endelig eller tellbart mange	X Mulige verdier x: Intervall eller hele R
Eksempel: $\{0, 1, \ldots, n\}$	Eksempel: $[0, 1]$ eller $[0, \infty)$
Sannsynlighetsfordeling: $f(x) = P(X = x)$ for alle mulige x	Sannsynlighetsfordeling: $f(x)$ definert for alle reelle x ved $P(a < X < b) = \int_a^b f(x)dx$
	

Figur 2.2: Diskret VS kontinuerlig stokastisk variabel.

Definisjon: (sannsynlighetsfordeling, diskret)

En sannsynlighetsfordeling $P(x_i)$ til en diskret stokastisk variabel X er en funksjon definert ved

$$\boxed{\underbrace{P(x_i)}_{\text{liten } x \text{ for verdier}} = \underbrace{P(X = x_i)}_{\text{stor } X \text{ for selve variabelen}} \qquad (2.1)}$$

hvor

$$i = 1, 2, 3, ... \tag{2.2}$$

$$P(x_i) = \text{sannsynligheten for at } X \text{ har verdien } x_i \tag{2.3}$$

slik at $\sum_i P(x_i) = 1$.

■

Definisjon: (sannsynlighetsfordeling, kontinuerlig)

En sannsynlighetstetthet $f(x)$ til en kontinuerlig stokastisk variabel X er en funksjon

$$\boxed{\underbrace{f(x)}_{\text{liten } x \text{ for verdier}} \qquad (2.4)}$$

slik at $\int_{-\infty}^{\infty} f(x)dx = 1$.

■

2.2 Forventning og varians

2.2.1 Forventning

Definisjon: (forventningsverdi, diskret)

For en *diskret* tilfeldig variabel X med de mulige verdiene x_1, x_2, ... , x_n er forventningsverdien:

$$\boxed{E[X] \overset{\text{def.}}{=} \sum_{i=1}^{n} x_i \, P(X = x_i) \qquad (2.5)}$$

hvor [2]

$$n = \text{antall utfall i utfallsrommet} \qquad (2.6)$$
$$P(X = x_i) = \text{punktsannsynligheten} \qquad (2.7)$$

■

Definisjon: (forventningsverdi, kontinuerlig)

For en *kontinuerlig* tilfeldig variabel X har forventningsverdien:

$$\boxed{E[X] \overset{\text{def.}}{=} \int_{-\infty}^{\infty} x \, f(x) \, dx \qquad (2.8)}$$

hvor

$$f(x) = \text{sannsynlighetstetthet} \qquad (2.9)$$

■

[2] For en terning er $n = 6$.

2.2.2 Varians

Definisjon: (varians)

For en stokastisk variabel X er variansen

$$\boxed{Var[X] \overset{\text{def.}}{=} E\big[\,(X - E[X])^2\,\big] \qquad (2.10)}$$

altså

$$Var[X] = \text{forventet avvik} \tag{2.11}$$

■

Diskret:

$$\boxed{Var[X] \overset{\text{lign.(2.5)}}{=} \sum_{i=1}^{n} \big(x_i - E[X]\big)^2 P(X = x_i) \qquad (2.12)}$$

Kontinuerlig:

$$\boxed{Var[X] \overset{\text{lign.(2.8)}}{=} \int_{-\infty}^{\infty} \big(x_i - E[X]\big)^2 f(x) \qquad (2.13)}$$

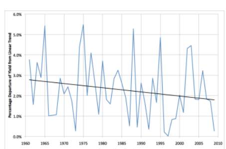
Setning: (varians , “varianssetningen”)

La X være en stokastisk variabel. For variansen gjelder da:

$$\boxed{Var[X] \ = \ E[X^2] - E[X]^2 \qquad (2.14)}$$

∎

2.2.3 Kovarians

Var[X]	Cov[X,Y]
Definisjon: $$Var[X] = E[\,(X - E[X])^2\,]$$	Definisjon: $$Cov[X,Y] = E[\,(X - E[X])\,(Y - E[Y])\,]$$
Mål på: • spredning • avvik mellom (X og E[X])	Mål på: • samvariasjon / korrelasjon • avvik mellom (X og E[X]) og (Y og E[Y])
Var[X] utskrevet: $$Var[X] = \sum_{i=1}^{n} (x_i - E[X])^2\, P(X=x_i)$$	Cov[X,Y] utskrevet: $$Cov[X,Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - E[X])\,(y_j - E[Y])\, p(x_i,y_j)$$
	

Figur 2.3: $Var[X]$ VS $Cov[X,Y]$.

Definisjon: (kovarians)

La X og Y være to stokastiske variabler. Med $\overbrace{\text{kovariansen}}^{\text{samvariasjon/korrelasjon}}$ mellom disse mener vi:

$$\overbrace{Cov[X,Y]}^{\text{samvariasjon/korrelasjon}} = E\big[\,(X - E[X])(Y - E[Y])\,\big] \tag{2.15}$$

■

Diskret:

$$Cov[X,Y] = \sum_{i=1}^{n}\sum_{j=1}^{m} \Big(x_i - E[X]\Big)\Big(y_j - E[Y]\Big)\, p(x_i, y_j) \tag{2.16}$$

Kontinuerlig:

$$Cov[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big(x - E[X]\Big)\Big(y - E[Y]\Big)\, p(x,y)\; dx\, dy \tag{2.17}$$

Definisjon: (*simultan*fordeling)

La X og Y være to stokastiske variabler. Med simultanfordeling menes:

$$\boxed{P(X = x\ ,\ Y = y)\ =\ p(x, y) \qquad (2.18)}$$

■

Definisjon: (*marginal*fordeling)

La X og Y være to stokastiske variabler. Med marginalfordeling menes:

$$P(X = x)\ =\ \sum_{y} p(x, y) \qquad (2.19)$$

$$P(Y = y)\ =\ \sum_{x} p(x, y) \qquad (2.20)$$

■

Setning: (“kovarianssetningen” , den generelle)

La X og Y være to stokastiske variabler. For $\overbrace{\text{kovariansen}}^{\text{samvariasjon/korrelasjon}}$ gjelder da:

$$\boxed{\overbrace{Cov[X,Y]}^{\text{samvariasjon/korrelasjon}} \quad = \quad E[X\,Y] \; - \; E[X]\,E[Y] \qquad (2.21)}$$

■

Definisjon: (uavhengighet) [3]

La X og Y være to stokastiske variabler. Disse er uavhengige dersom:

$$\boxed{P(X = x \;,\; Y = y) \quad = \quad P(X = x) \cdot P(Y = y) \qquad (2.22)}$$

for alle $X = x$ og $Y = y$.

■

Setning: (uavhengighet)

Dersom X og Y er to uavhengige stokastiske variabler så gjelder:

$$\boxed{E[X \cdot Y] \;=\; E[X]\,E[Y] \qquad (2.23)}$$

■

[3] Jamfør den *analoge* setn. for uavhengighet mellom begivenheter A og B i lign.(1.30): $P(A \cap B) = P(A) \cdot P(B)$.

Setning: ("kovarianssetningen" , den spesielle)

La X og Y være to stokastiske variabler. Dersom X og Y er uavhengige, så er [4]

$$\boxed{\underbrace{Cov[X,Y] \quad = \quad 0}_{\text{ukorrelerte}} \qquad\qquad (2.24)}$$

■

[4] Uavhengighet er altså et **sterkere** krav enn ukorrelert.
Dersom en skøyteløper har verdenrekorden på 10 000 meter så har han/hun også norgesrekorden.

2.2.4 Noen regneregler

La a og b være konstanter. La videre X og Y være to stokastiske variabler. Da gjelder:

Regneregler for forventning:

$$\boxed{\begin{aligned} E[a] &= a && (2.25)\\ E[a+X] &= a+E[X] && (2.26)\\ E[a\cdot X] &= a\,E[X] && (2.27)\\ E[aX+bY] &= aE[X]\;+\;bE[Y] && (2.28) \end{aligned}}$$

Regneregler for varians:

$$\boxed{\begin{aligned} Var[a] &= 0 && (2.29)\\ Var[a+X] &= Var[X] && (2.30)\\ Var[a\,X] &= a^2\,Var[X] && (2.31)\\ Var[X] &= Var[-X] && (2.32)\\ Var[aX+bY] &= a^2Var[X]\;+\;b^2Var[Y]\;+\;2ab\,Cov[X,Y] && (2.33) \end{aligned}}$$

Dessuten er alltid $Var[X] \geq 0$, dvs. en varians kan aldri være negativ.

Setning: (kovarians II)

La X og Y være to stokastiske variabler. Generelt gjelder da følgende sammenhengen mellom variansen og kovariansen:

$$\boxed{\overbrace{Var[aX+bY]}^{\text{variasjon/(spredning)}} \;=\; a^2\,Var[X] \;+\; b^2\,Var[Y] \;+\; 2ab\cdot\overbrace{Cov[X,Y]}^{\text{samvariasjon}} \qquad (2.34)}$$

hvor a og b er konstanter.

■

Setning (spesialtilfelle , kovarians II)

La X og Y være to stokastiske variabler. Dersom X og Y er ukorrelerte, dvs.:

$$\overbrace{Cov[X,Y] \;=\; 0}^{X \text{ og } Y \text{ ukorrelerte}} \tag{2.35}$$

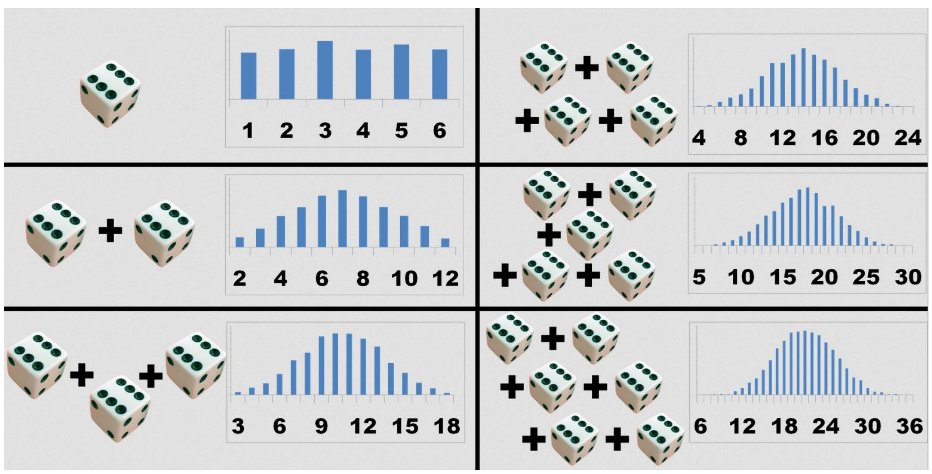
så er:

$$\boxed{Var[aX+bY] \;=\; a^2\,Var[X] \;+\; b^2\,Var[Y] \qquad (2.36)}$$

■

Kapittel 3

Diskrete stokastiske fordelinger



Figur 3.1: Sentralgrensesetningen ("CLT").

3.1 Bernoulli-fordelingen

Definisjon: ($\overbrace{\text{Bernoulli}}^{\text{diskret}}$-fordeling , $X \sim \overbrace{\text{Ber}[p]}^{\text{1 param.}}$)

Bernoulli-fordelingen er en diskret fordeling med to mulige utfall,
$X = 1$ (”suksess”) og $X = 0$ (”fiasko”)

$$\boxed{P(X = x) \overset{\text{def.}}{=} \begin{cases} p & , \quad X = 1 \\ \\ 1 - p & , \quad X = 0 \end{cases} \qquad (3.1)}$$

hvor

$$\begin{aligned} X &= \text{stokastisk variabel} && (3.2) \\ &= \text{antall “suksesser” i et } \textbf{Bernoulli}\text{-forsøk} \end{aligned}$$

$$p = \text{sannsynlighet for at } X = 1 \text{ (“suksess”)} \qquad (3.3)$$

$$1 - p = \text{sannsynlighet for at } X = 0 \text{ (“fiasko”)} \qquad (3.4)$$

■

3.1.1 Forventningsverdi

Setning: (forventning av $X \sim \overbrace{\text{Bin}[n,p]}^{\text{2 param.}}$)

La X være en **Bernoulli** variabel, dvs. $X \sim \text{Ber}[p]$. Da gjelder:

$$\boxed{E[X] \;=\; p \qquad\qquad (3.5)}$$

■

3.1.2 Varians

Setning: (varians av $X \sim \overbrace{\text{Ber}[p]}^{\text{1 param.}}$

La X være en **Bernoulli** variabel, dvs. $X \sim \text{Ber}[p]$. Da gjelder:

$$\boxed{Var[X] \;=\; p\,(1-p) \qquad\qquad (3.6)}$$

■

3.2 Den binomiske fordelingen

Definisjon: $\Big($ $\overbrace{\text{binomisk}}^{\text{diskret}}$ fordeling , $X \sim \overbrace{\text{Bin}[n,p]}^{\text{2 param.}}$ $\Big)$

Punktsannsynlighetene for en binomisk fordeling er:

$$\boxed{P(X=x) \overset{\text{def.}}{=} \binom{n}{x} p^x \, (1-p)^{n-x} \qquad (3.7)}$$

hvor

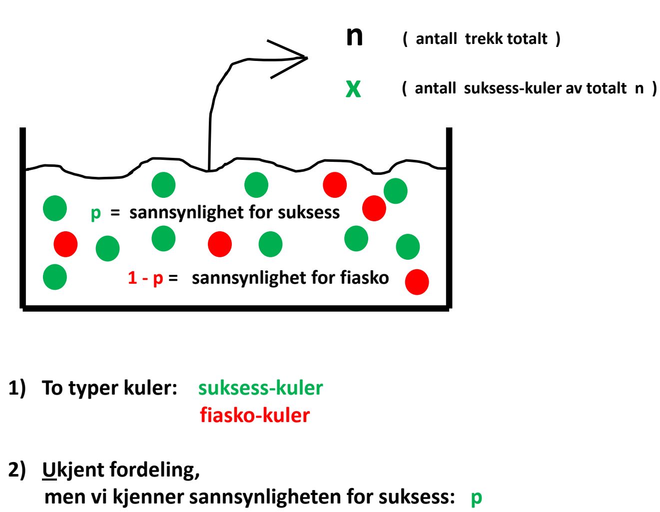
$$\begin{aligned} X &= \text{stokastisk variabel} && (3.8) \\ &= \text{antall "suksesser" i en } \textbf{binomisk} \text{ forsøksserie på totalt } n \text{ forsøk} \end{aligned}$$

$$p = \text{sannsynlighet for "suksess"} \qquad (3.9)$$

$$n = \text{totalt antall forsøk} \qquad (3.10)$$

$$\underbrace{\binom{n}{x}}_{\text{binomialkoeff.}} = \frac{n!}{(n-x)!\;x!} \qquad (\text{"}n \text{ over } x\text{"}) \qquad (3.11)$$

■



Figur 3.2: Binomisk forsøksserie.

3.2.1 Forventningsverdi

Setning: (forventning av $X \sim \overbrace{\text{Bin}[n,p]}^{\text{2 param.}}$)

La X være en **binomisk** variabel, dvs. $X \sim \text{Bin}[n,p]$. Da gjelder:

$$\boxed{E[X] \;=\; n \cdot p \qquad (3.12)}$$

■

3.2.2 Varians

Setning: (varians av $X \sim \overbrace{\text{Bin}[n,p]}^{\text{2 param.}}$

La X være en **binomisk** variabel, dvs. $X \sim \text{Bin}[n,p]$. Da gjelder:

$$\boxed{Var[X] \;=\; n\,p\,(1-p) \qquad (3.13)}$$

■

3.3 Den hypergeometriske fordelingen

Definisjon: ($\overbrace{\text{hypergeometrisk}}^{\text{diskret}}$ fordeling , $X \sim \overbrace{\text{Hyp}[N, M, n]}^{\text{3 param.}}$ [1])

Punktsannsynlighetene for en hypergeometrisk fordeling er

$$\boxed{P(X = x) \overset{\text{def.}}{=} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}} \tag{3.14}$$

hvor

$$X = \text{stokastisk variabel}$$
$$= \text{antall ``spesielle'' elementer i det tilfeldige utvalget på } n \text{ trukne elementer} \tag{3.15}$$

$$N = \text{antall elementer i grunnmengden} \tag{3.16}$$

$$M = \text{antall ``spesielle'' elementer} \tag{3.17}$$

$$N - M = \text{antall ``vanlige'' elementer} \tag{3.18}$$

$$n = \text{antall trukne elementer} \tag{3.19}$$

$$\underbrace{\binom{N}{n}}_{\text{binomialkoeff.}} = \frac{N!}{(N-n)!\, n!} \qquad (\text{``}N \text{ over } n\text{''}) \tag{3.20}$$

■

[1] Skrivemåten $X \sim \text{Hyp}[N, M, n]$ betyr at den stokastiske variabelen X har en sannsynlighetsfordeling $P(X = x)$ som er hypergeometrisk fordelt.

3.3.1 Forventning og varians

Setning: (forventning av $X \sim \overbrace{\text{Hyp}[N, M, n]}^{\text{3 param.}}$

La X være en **hypergeometrisk** variabel, dvs. $X \sim \text{Hyp}[N, M, n]$. Da gjelder:

$$\boxed{E[X] \quad = \quad n\,\frac{M}{N} \qquad\qquad (3.21)}$$

■

Setning: (varians av $X \sim \overbrace{\text{Hyp}[N, M, n]}^{\text{3 param.}}$)

La X være en **hypergeometrisk** variabel, dvs. $X \sim \text{Hyp}[N, M, n]$. Da gjelder:

$$\boxed{Var[X] \quad = \quad \frac{N-n}{N-1}\,n\,\frac{M}{N}\left(1 - \frac{M}{N}\right) \qquad\qquad (3.22)}$$

■

3.4 Sammenheng mellom Hyp$[N, M, n]$ og Bin$[n, p]$

Setning: $\quad \Big(\overbrace{\text{Hyp}[N, M, n]}^{\text{3 param.}} \longrightarrow \overbrace{\text{Bin}[n, p]}^{\text{2 param.}} \Big)$

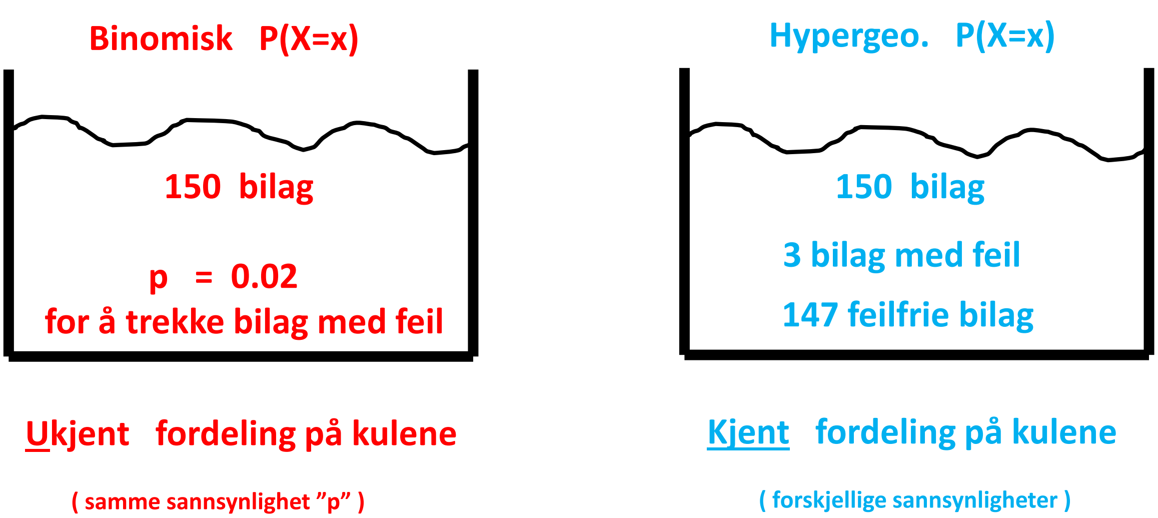
Dersom grunnmengden N i en hypergeometrisk forsøkserie er mye større enn utvalget n, typisk

$$\boxed{N \;\gtrsim\; 20 \cdot n} \tag{3.23}$$

så er

$$\boxed{X \sim \text{Hyp}[N, M, n] \quad \overset{N \gtrsim 20\cdot n}{\approx} \quad X \sim \text{Bin}[n, p]} \tag{3.24}$$

■



Figur 3.3: **<u>U</u>kjent** og **kjent** fordeling mellom kulene i urnen.

3.5 Poissonfordelingen

Definisjon: $\left(\overbrace{\text{Poisson}}^{\text{diskret}}\text{fordeling} \;,\; X \sim \overbrace{\text{Poi}[\lambda]}^{\text{1 param.}} \right)$

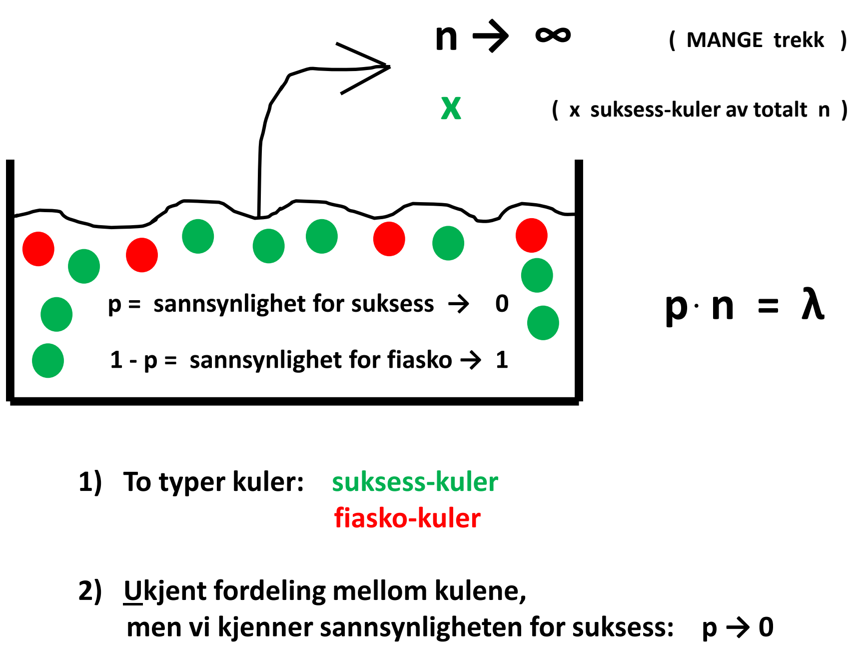
Punktsannsynlighetene for en Poissonfordeling er

$$\boxed{P(X = x) \stackrel{\text{def.}}{=} \frac{\lambda^x}{x!}\, e^{-\lambda}} \tag{3.25}$$

hvor

$$
\begin{aligned}
X &= \text{stokastisk variabel} \\
&= \text{antall begivenheter som inntreffer innenfor en gitt } \underline{\text{tid}} \text{ eller gitt } \underline{\text{rom}} \qquad (3.26) \\
\lambda &= \text{gjennsomsnittlig antall begivenheter innenfor en gitt } \underline{\text{tid}} \text{ eller gitt } \underline{\text{rom}} \qquad (3.27)
\end{aligned}
$$

■



Figur 3.4: Poisson fordeling = binomisk fordeling når $p \longrightarrow 0^+$ og $n \longrightarrow \infty$.

Setning: $\left(\overbrace{\text{Bin}[n,p]}^{\text{2 param.}} \longrightarrow \overbrace{\text{Poi}[\lambda]}^{\text{1 param.}} \right)$

I en binomisk sannsynlighetsfordeling, dersom sannsynligheten for "suksess" p er svært liten og antall forsøk n er svært stor, typisk

$$\boxed{n \gtrsim 50 \qquad \text{og} \qquad p \lesssim 0.05 \qquad (3.28)}$$

så er

$$\boxed{X \sim \text{Bin}[n,p] \quad \overset{n \gtrsim 50 \text{ og } p \lesssim 0.05}{\approx} \quad X \sim \text{Poi}[\lambda] \qquad (3.29)}$$

■

3.5.1 Forventning og varians

Setning: (forventning av $X \sim \text{Poi}[\lambda]$ [2])

La X være en **Poisson**fordelt variabel, dvs. $X \sim \text{Poi}[\lambda]$. Da gjelder:

$$\boxed{E[X] \;=\; \lambda \qquad\qquad (3.30)}$$

■

Setning: (varians av $X \sim \text{Poi}[\lambda]$)

La X være en **Poisson**fordelt variabel, dvs. $X \sim \text{Poi}[\lambda]$. Da gjelder:

$$\boxed{Var[X] \;=\; \lambda \qquad\qquad (3.31)}$$

■

[2] Skrivemåten $X \sim \text{Poi}[\lambda]$ betyr at den stokastiske variabelen X har en sannsynlighetsfordeling $P(X = x)$ som er Poissonfordelt.

Kapittel 4

Kontinuerlige stokastiske fordelinger og CLT

4.1 Normalfordelingen (kontinuerlig)

<u>**Definisjon**</u>: ($\overbrace{\text{normal}}^{\text{kont.}}$fordeling [1], $X \sim \overbrace{N[\mu, \sigma]}^{\text{2 param.}}$)

Tetthetsfunksjonen til en normalfordeling er: [2]

$$\boxed{f_X(x) \;=\; \frac{1}{\sqrt{2\pi\,\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \qquad (4.1)}$$

hvor [3]

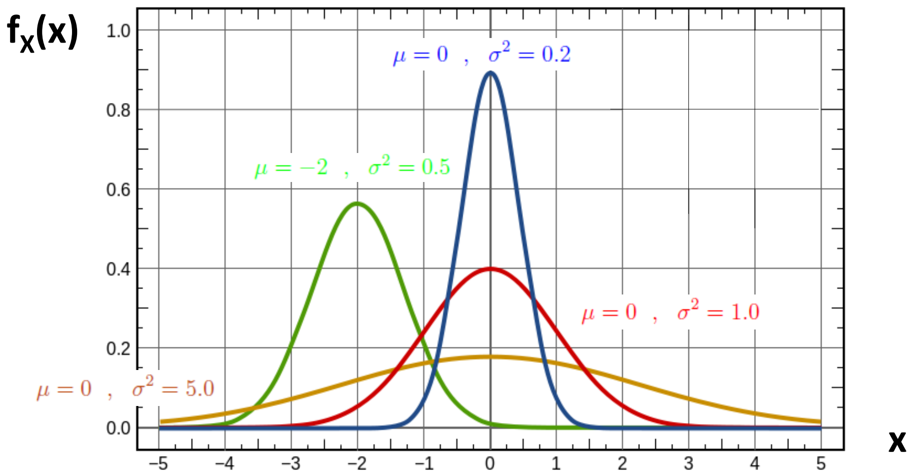
$$\mu \;=\; E[X] \;=\; \text{forventingen av } X \qquad (\text{ lokalisering av toppunkt }) \qquad (4.2)$$

$$\sigma^2 \;=\; Var[X] \;=\; \text{variansen av } X \qquad (4.3)$$

$$\sigma \;=\; \sqrt{\sigma^2} \;=\; \text{standardavviket av } X \qquad (\text{ halvbredden av kurven }) \qquad (4.4)$$

■

[1] Normalfordelingen kalles også en *Gauss*-fordeling.
[2] Lign.(4.1) er en tetthetsfunksjon.
[3] μ leses "my". σ leses "sigma".



Figur 4.1: **Normal**fordelinger for ulike verdier av μ og σ.

4.1.1 Standardisering

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

La X være en *kontinuerlig* stokastisk variabel. Dersom

$$E[X] \;=\; \mu \;=\; 0 \qquad , \qquad Var[X] \;=\; \sigma^2 \;=\; 1 \tag{4.5}$$

så vil den generelle normalfordelingen i lign.(4.1) redusere seg til

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (4.6)}$$

Dette kalles den **standardiserte** normalfordelingen.

■

4.1.2 Standardisering = omskalering

$$\boxed{Z \;=\; \frac{X - \mu}{\sigma} \qquad (4.7)}$$

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:

$$\boxed{Z = \frac{X - \mu}{\sigma} \qquad (4.8)}$$

Da er den *kumulative* sannsynlighetsfordelingen gitt ved:

$$\boxed{\underbrace{P(X \leq x)}_{\text{integral}} = \underbrace{P(Z \leq z) = G(z)}_{\text{tabelloppslag}} \qquad (4.9)}$$

hvor

$$\underbrace{G(z) = \int_{-\infty}^{z} f_Z(z)\, dz}_{\text{tabelloppslag for normalfordeling}} \tag{4.10}$$

og $f_Z(z)$ er gitt ved lign.(4.6).

■

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

Pga. symmetri innser vi at fra figur (4.2) at

$$G(z) \;+\; G(-z) \;=\; \text{arealet under hele kurven} \;=\; 1 \tag{4.11}$$

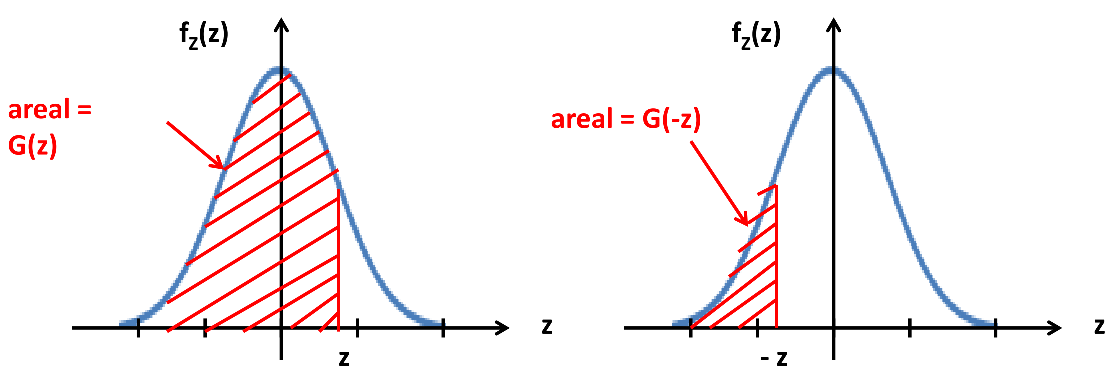
Altså vi kan skrive:

$$\boxed{G(-z) \;=\; 1 \;-\; G(z) \qquad (4.12)}$$

eller ekvivalent

$$P(Z \leq -z) \;=\; 1 \;-\; P(Z \leq z) \tag{4.13}$$

$\blacksquare$



Figur 4.2: Arealene representerer Gauss-integralene $G(z)$ og $G(-z)$.

4.1.3 Standardavvik σ og %-vis areal

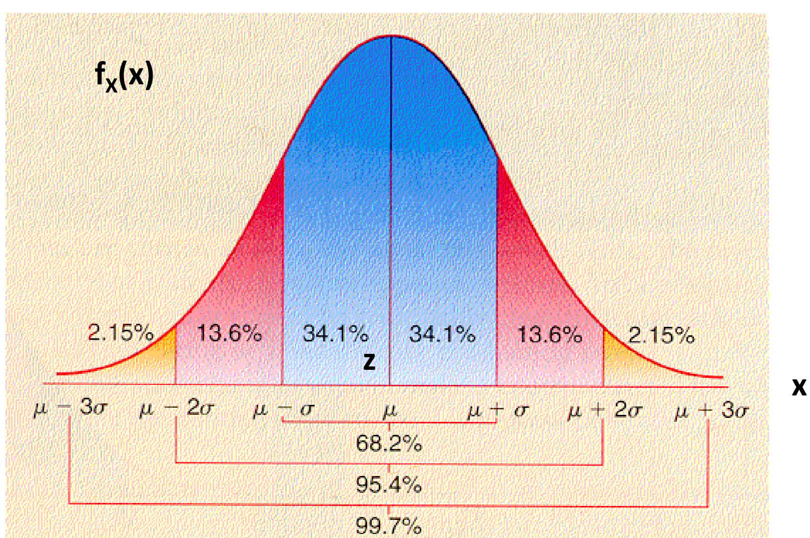
For en normalfordeling med et gitt standardavvik σ, så dekker intervallet

$$\mu - \sigma \;\leq\; X \;\leq\; \mu + \sigma \quad \text{hele } 68.2\ \% \text{ av arealet} \tag{4.14}$$

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

$$\mu - 2\sigma \;\leq\; X \;\leq\; \mu + 2\sigma \quad \text{hele } 95.4\ \% \text{ av arealet} \tag{4.15}$$

under sannsynlighetfordelingen $f_X(x)$.



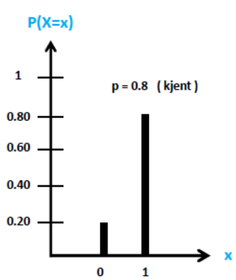
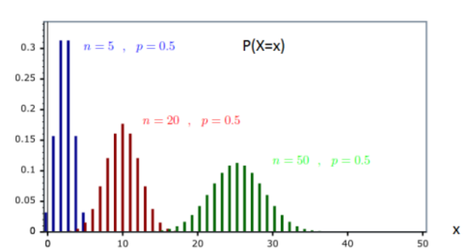
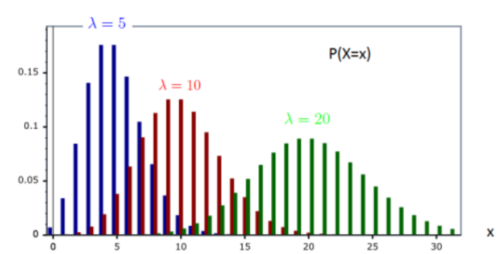
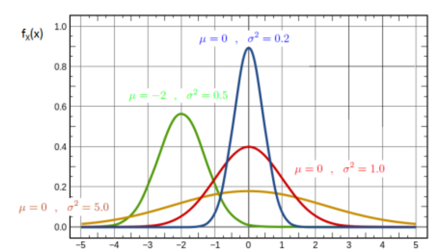
Figur 4.3: Standardavvik σ og %-vis areal for en normalfordeling.

4.2 Oversikt: Ber, Bin, Poi og N

Fire viktige sannsynlighetsfordelinger sannsynlighetsfordelinger:

1. Bernoulli fordeling
2. Binomisk fordeling
3. Poissonfordelingen
4. Normalfording

Ber[p]	Bin[n , p]	Poi[λ]	N[μ , σ]
1 param.	2 param.	1 param.	2 param.
diskret	diskret	diskret	kontinuerlig
$P(X=x) \overset{\text{def.}}{=} \begin{cases} p & , \ X=1 \\ 1-p & , \ X=0 \end{cases}$	$P(X=x) \overset{\text{def.}}{=} \binom{n}{x} p^x (1-p)^{n-x}$	$P(X=x) \overset{\text{def.}}{=} \frac{\lambda^x}{x!} e^{-\lambda}$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$E[X] = p$	$E[X] = n \cdot p$	$E[X] = \lambda$	$E[X] = \mu$
$Var[X] = p(1-p)$	$Var[X] = n \cdot p(1-p)$	$Var[X] = \lambda$	$Var[X] = \sigma^2$

Ber[p]	Bin[n , p]	Poi[λ]	N[μ , σ]
1) 2 mulige utfall 2) «p» sannsynlighet for suksess 3) «1-p» sanns. for fiasko 4) n=1 antall trukne elementer	1) 2 mulig utfall 2) samme ”p” for ”suksess” 3) uavhengige 4) n antall forsøk	1) x antall begivenheter innenfor en gitt tid 2) λ = rate	1) Tetthetsfunksjon $f_X(x)$ 2) Gausskurve
- utfall 1 eller 0 - <u>registerer</u> suksess (X=1) eller fiasko (X=0)	- <u>kjenner ikke</u> fordelingen i urnen - m / tilbakelegging - <u>teller opp</u> 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 <u>tilnæring</u> være normalfordelt (f.eks. ”CLT”)
			

4.3 Sentralgrensesetningen

Setning: ("*CLT*" [4] , sentralgrensesetningen) (versjon 1)

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

- 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.)

$$\boxed{\overline{X} = \frac{X_1 + X_2 + X_3 + \ ... \ + X_n}{n}} \tag{4.16}$$

er normalfordelt i grensen når antall forsøk n blir uendelig:

$$\boxed{\overline{X} \underset{n\to\infty}{\sim} \mathrm{N}\left[\mu \, , \, \frac{\sigma}{\sqrt{n}}\right]} \tag{4.17}$$

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

$$\boxed{E[\,\overline{X}\,] = \mu \qquad \text{og} \qquad Var[\,\overline{X}\,] = \frac{\sigma^2}{n}} \tag{4.18}$$

■

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

Setning: ("*CLT*" [5] , sentralgrensesetningen) (versjon 2)

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

- 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.)

$$\boxed{Y \;=\; X_1 \,+\, X_2 \,+\, X_3 \,+\, ... \,+\, X_n} \tag{4.19}$$

er tilnærmet normalfordelt når antall forsøk n **blir stor**: [6]

$$\boxed{Y \underset{n=\text{stor}}{\sim} \mathrm{N}\Big[n\mu \,,\, \sqrt{n}\sigma \Big]} \tag{4.20}$$

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

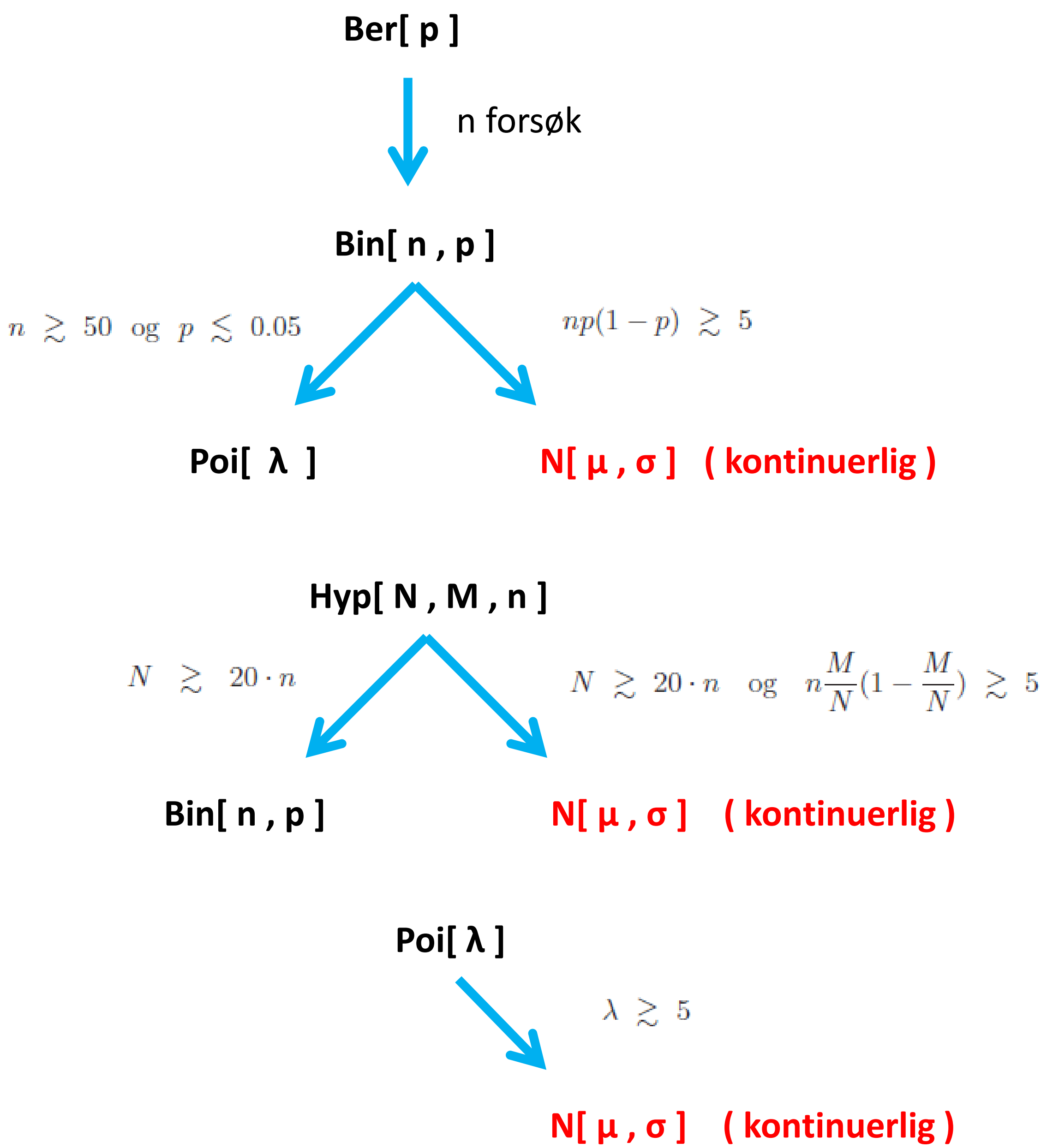
$$\boxed{E[\,Y\,] \;=\; n\mu \qquad \text{og} \qquad Var[\,Y\,] \;=\; n\,\sigma^2} \tag{4.21}$$

■

[5] På engelsk brukes ofte forkortelsen *CLT*, dvs. "*central limit theorem*".
[6] Selv om lign.(4.21) divergerer for $n \to \infty$ så er lign.(4.20) en god tilnæremelse for endelige, men store n.

4.4 Diskrete fordelinger $\longrightarrow$ normalfordeling

4.4.1 Sammenheng: Ber, Bin, Hyp, Poi og N



Figur 4.4: Sammenheng: Ber, Bin, Hyp, Poi og N.

4.5 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

$$Y \;=\; X_1 \;+\; X_2 \;+\; X_3 \tag{4.22}$$

binomisk fordelt:

$$\boxed{Y \;\sim\; \text{Bin}\big[\, n_Y \;,\; p \,\big] \qquad (4.23)}$$

hvor

$$n_Y \;=\; n_1 + n_2 + n_3 \tag{4.24}$$

■

<u>**Setning**</u>: (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

$$Y \quad = \quad X_1 \quad + \quad X_2 \quad + \quad X_3 \tag{4.25}$$

Poisson fordelt:

$$\boxed{Y \;\sim\; \text{Poi}\left[\,\lambda_Y\,\right] \qquad (4.26)}$$

hvor

$$\lambda_Y \;\;=\;\; \lambda_1 + \lambda_2 + \lambda_3 \tag{4.27}$$

■

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

$$Y \quad = \quad aX_1 \quad + \quad bX_2 \quad + \quad cX_3 \tag{4.28}$$

normalfordelt:

$$\boxed{Y \;\sim\; N\left[\,\mu_Y \;,\; \sigma_Y\,\right] \qquad (4.29)}$$

hvor

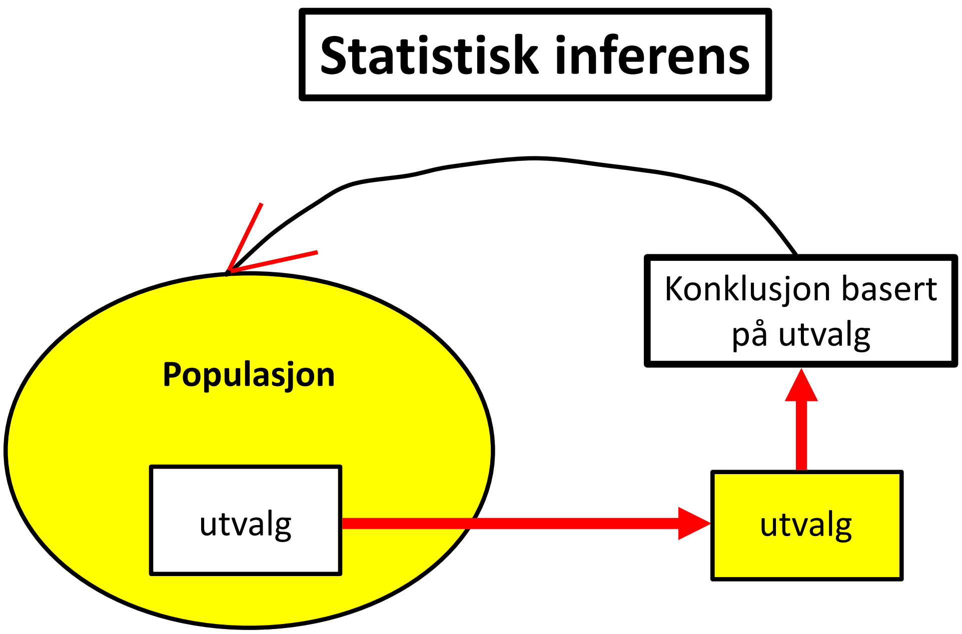
$$\mu_Y \quad = \quad a\mu_1 + b\mu_2 + c\mu_3 \tag{4.30}$$

$$\sigma_Y^2 \quad = \quad a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2 \tag{4.31}$$

■

Kapittel 5

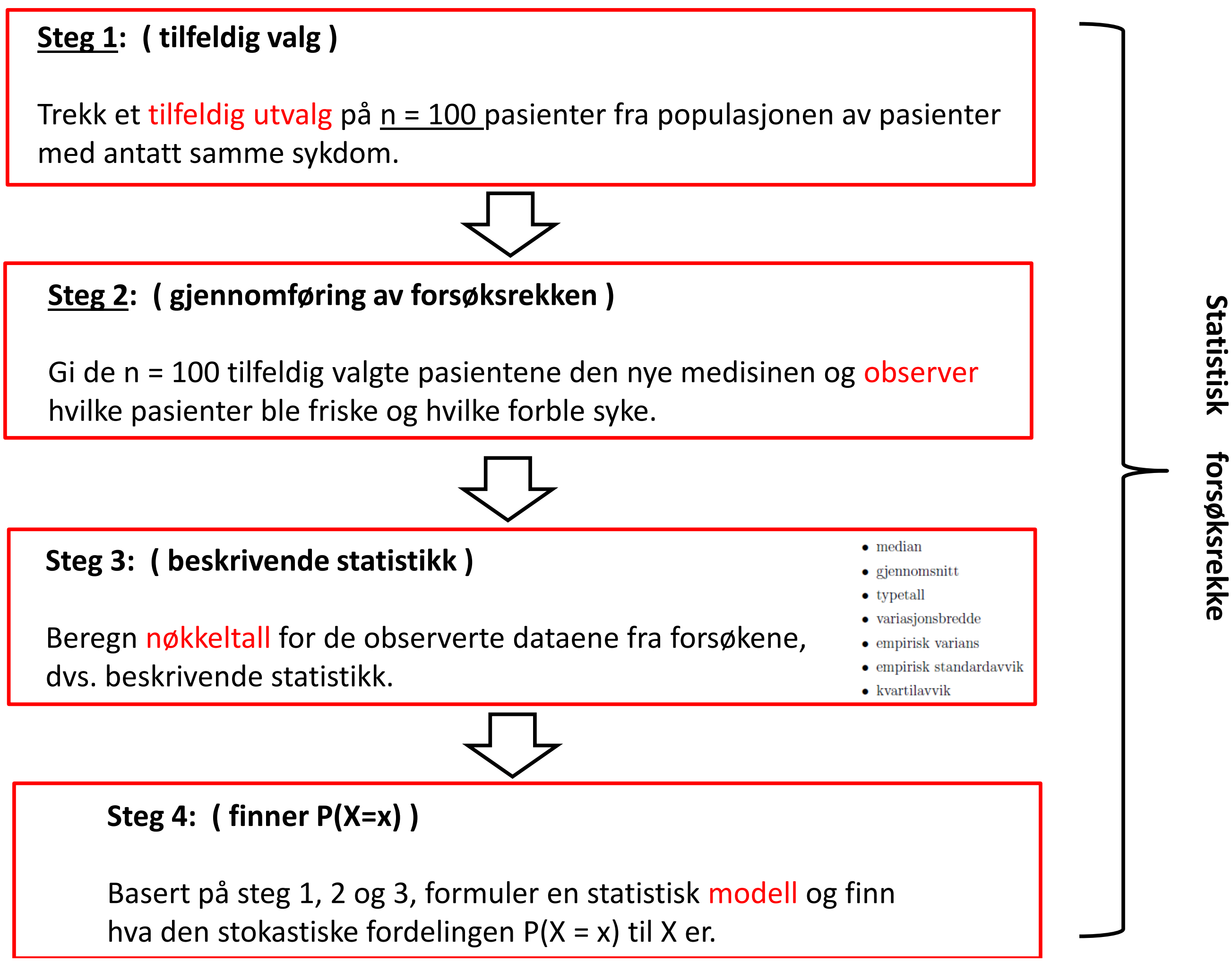
Statistisk inferens



Figur 5.1: Statistisk inferens.

5.1 Fra sannsynlighetsteori til statistisk inferens

Statistisk forsøksrekke (4 steg)



Figur 5.2: 4 steg.

Definisjon: (populasjon)

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

■

Definisjon: (utvalg)

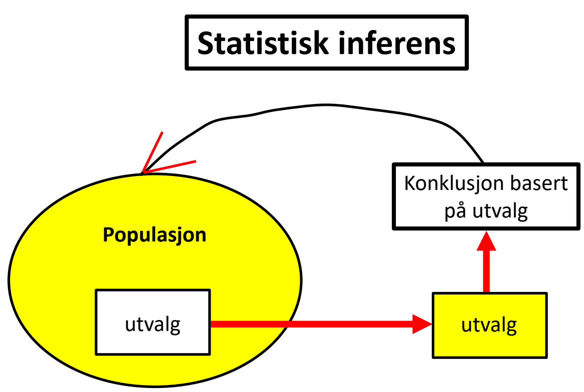
Utvalg = en delmengde av populasjonen,
dvs. en samling av data som er hentet fra en populasjon

■

Definisjon: (statistisk inferens)

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

■



Figur 5.3: Statistisk inferens.

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

■

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 [1]
2. X_i-ene er gjensidig uavhengige

■

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 $\underbrace{x_1, x_2, \ldots, x_n}_{\text{tall}}$.

■

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

5.1.1 Lokaliseringsmål

Definisjon: (median)

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

$$\boxed{\text{median} = \begin{cases} \underline{\text{midtre}} \text{ observasjonen} & , \quad n = \text{odde} \\ \\ \text{gjennomsnitt av to } \underline{\text{midterste}} \text{ observasjonene} & , \quad n = \text{like} \end{cases} \qquad (5.1)}$$

■

Definisjon: (gjennomsnitt)

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

$$\boxed{\overline{x} \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (5.2)}$$

■

Definisjon: (typetall) [2]

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

$$\boxed{\text{typetall} \;=\; \text{den verdien som forekommer } \underline{\text{hyppigst}} \qquad (5.3)}$$

■

[2] Kalles også modus eller modalverdi.

5.1.2 Spredningsmål

Definisjon: (empirisk varians)[3]

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:

$$\boxed{s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (5.4)}$$

■

Definisjon: (empirisk standardavvik)[4]

Det empiriske standardavviket er:

$$\boxed{s_x = \sqrt{s_x^2} \qquad (5.5)}$$

■

[3] Kalles også **utvalgsvariansen**.
[4] Kalles også **utvalgstandardavviket**.

Definisjon: (variasjonsbredde)

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

$$\boxed{\text{variasjonsbredde} \;=\; \text{differansen mellom } \underline{\text{største}} \text{ og } \underline{\text{minste}} \text{ verdi} \qquad (5.6)}$$

■

Definisjon: (kvartilavvik)

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

$$k_1 \;=\; \text{nedre kvartil , dvs.} \quad 25\% \text{ av observasjonene har verdi } \leq \; k_1 \tag{5.7}$$

$$k_2 \;=\; \text{medianen , dvs.} \begin{cases} 50\% \text{ av observasjonene har verdi } \leq \; k_2 \\ \\ 50\% \text{ av observasjonene har verdi } \geq \; k_2 \end{cases} \tag{5.8}$$

$$k_3 \;=\; \text{øvre kvartil , dvs.} \quad 75\% \text{ av observasjonene har verdi } \leq \; k_3 \tag{5.9}$$

Da er

$$\boxed{\text{kvartilavvik} \;=\; k_3 \;-\; k_1 \qquad (5.10)}$$

■

5.2 Statistisk modell

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. [5] stokastiske forsøksvariabler for utfallet av forsøkene og X er tilhørende populasjonsvariabel.

En *statistisk modell* for forsøksrekken er da:

$$\boxed{X_1,\, X_2,\, X_3,\, \ldots\, ,X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \qquad (5.11)}$$

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

$$V = \text{verdimengden til de stokastiske forsøksvariablene} \qquad (5.12)$$

$$P_\theta(X = x) = \text{parametrisk familie av sannsynlighetsfordelinger for } V \qquad (5.13)$$

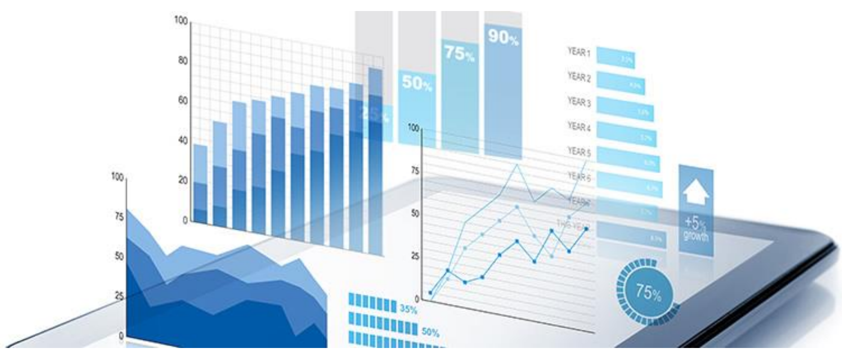
$$\Theta = \text{parametermengden, dvs. de mulige verdiene for parametrene} \qquad (5.14)$$

■

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

Kapittel 6

Estimering og konfidensintervaller



Figur 6.1: Estimering og konfidensintervaller.

6.1 Estimatorer

Felles antagelser: (for definisjoner og setninger i avsnitt 6.1)

Anta at:

$$X_1,\ X_2,\ X_3,\ \ldots\ , X_n \ \overset{\text{i.i.d.}}{\sim}\ X \ \sim\ P_\theta(X = x) \tag{6.1}$$

for $\theta \in \Theta$.

Definisjon: (estimator)

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

$$\boxed{\hat{\theta} \ = \ \hat{\theta}(X_1, X_2 \ldots X_n) \qquad (6.2)}$$

■

Definisjon: (forventningsrett estimator)

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

$$\boxed{E[\,\hat{\theta}\,] \ = \ \theta \qquad (6.3)}$$

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

I motsatt fall er den forventningskjev.

■

6.2 Konfidensintervaller

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

Anta at:

$$X_1,\, X_2,\, X_3,\, \ldots\,,X_n \overset{\text{i.i.d.}}{\sim} X \sim P_\theta(X = x) \tag{6.4}$$

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

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

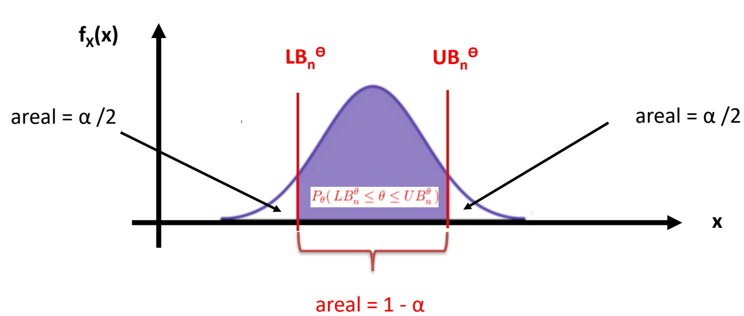
Et $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\, LB_n^\theta\,,\, UB_n^\theta\,] \qquad (6.5)}$$

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.:

$$\boxed{P_\theta(\, LB_n^\theta \leq \theta \leq UB_n^\theta\,) \;=\; 1-\alpha \qquad (6.6)}$$

■



Figur 6.2: Konfidensintervall.

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

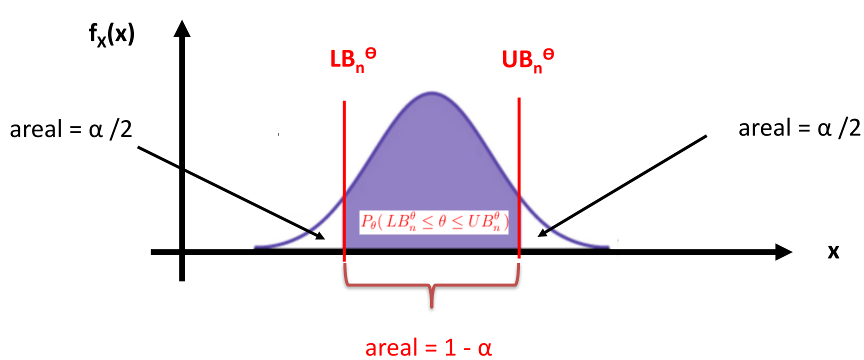
Et asymptotisk $(1-\alpha)\,100\,\%$-konfidensintervall for θ er et *stokastisk intervall*

$$\boxed{[\,LB_n^\theta\,,\,UB_n^\theta\,] \qquad (6.7)}$$

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.:

$$\boxed{\lim_{n\to\infty} P_\theta(\,LB_n^\theta \leq \theta \leq UB_n^\theta\,) \;=\; 1-\alpha \qquad (6.8)}$$

■



Figur 6.3: Konfidensintervall.

6.3 Student's t-fordeling og χ_k^2-fordeling

Definisjon: (χ_k^2-fordeling)

La $Z_1, Z_2, \ldots, Z_n$ være i.i.d. standard normalfordelte stokastiske variabler, dvs. $Z_i \sim N[0,1]$. Da er:

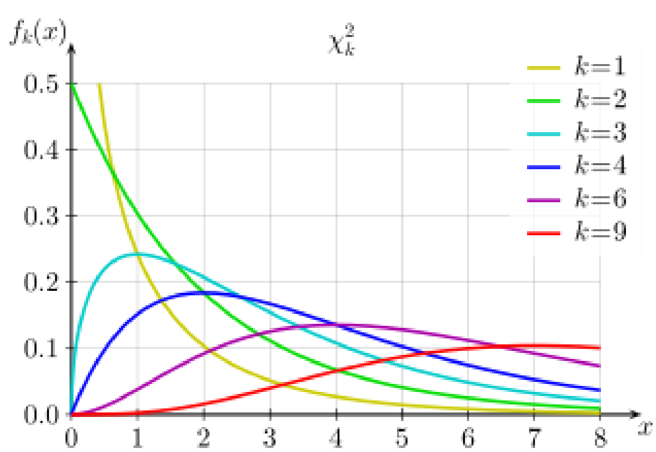
$$\boxed{Q \;=\; \sum_{i=1}^{n} Z_i^2 \;=\; Z_1^2 + \cdots + Z_n^2 \qquad (6.9)}$$

kji-fordelt med n frihetsgrader. Vi skriver:

$$\boxed{Q \;\sim\; \chi_n^2 \qquad (6.10)}$$

■

Figur 6.4 viser χ_k^2-fordelingen for ulike frihetsgrader k.



Figur 6.4: χ_k^2-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:

$$\boxed{Q \;=\; \frac{S_y^2}{\sigma^2/(n-1)} \;\sim\; \chi^2_{n-1} \qquad (6.11)}$$

χ^2_{n-1}-fordelt med $n-1$ frihetsgrader, hvor

$$S_y^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \tag{6.12}$$

■

Definisjon: (Student's t-fordeling)

La $Z \sim N[0,1]$ og $Q \sim \chi_n^2$ være stokastiske variabler.
Da er:

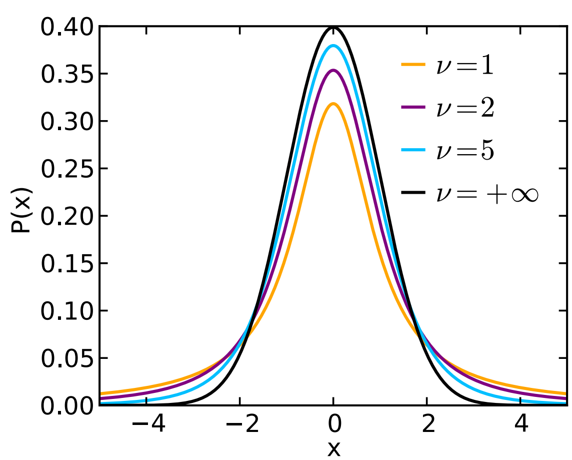
$$\boxed{T = \frac{\sqrt{n}Z}{\sqrt{Q}} \qquad (6.13)}$$

Student's t-fordelt med n frihetsgrader. Vi skriver

$$\boxed{T \sim t_n \qquad (6.14)}$$

■

Figur 6.5 viser Student's t-fordelingen for ulike ulike frihetsgrader.



Figur 6.5: 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

$$\boxed{T \;=\; \frac{\overline{Y} - \mu}{S_y/\sqrt{n}} \;\sim\; t_{n-1} \qquad (6.15)}$$

Student's t-fordelt med $n-1$ frihetsgrader.

■

Konfidensintervall for p

Asymptotisk: (CLT , normalfordeling)

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

= [0.8163 , 0.9437] (α = 0.05)

Konfidensintervall for μ

Asymptotisk: (CLT , normalfordeling)

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

= [0.8886 , 0.9048] (α = 0.05)

Eksakt: (Student's *t* fordeling)

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

= [0.8155 , 0.9444] (α = 0.05)

Konfidensintervall for σ

Asymptotisk: (CLT , normalfordeling)

$$[\, LB_n^\sigma \, , \, UB_{n'}^\sigma \,] = \left[\sqrt{\frac{n-1}{(n-1) + z_{1-\alpha/2}\sqrt{2(n-1)}}} \, S_y \, , \, \sqrt{\frac{n-1}{(n-1) + z_{\alpha/2}\sqrt{2(n-1)}}} \, S_y \right]$$

= [0.0365 , 0.0486] (α = 0.05)

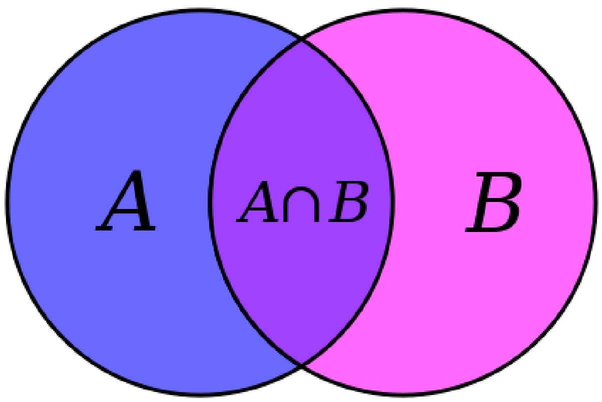
Eksakt: (χ^2 fordeling)

$$[\, LB_n^\sigma , UB_n^\sigma \,] = \left[\sqrt{\frac{n-1}{\chi_{1-\alpha/2}}} \, S_y \, , \, \sqrt{\frac{n-1}{\chi_{\alpha/2}}} \, S_y \right]$$

= [0.0362 , 0.0479] (α = 0.05)

Tillegg A

Mengdelære

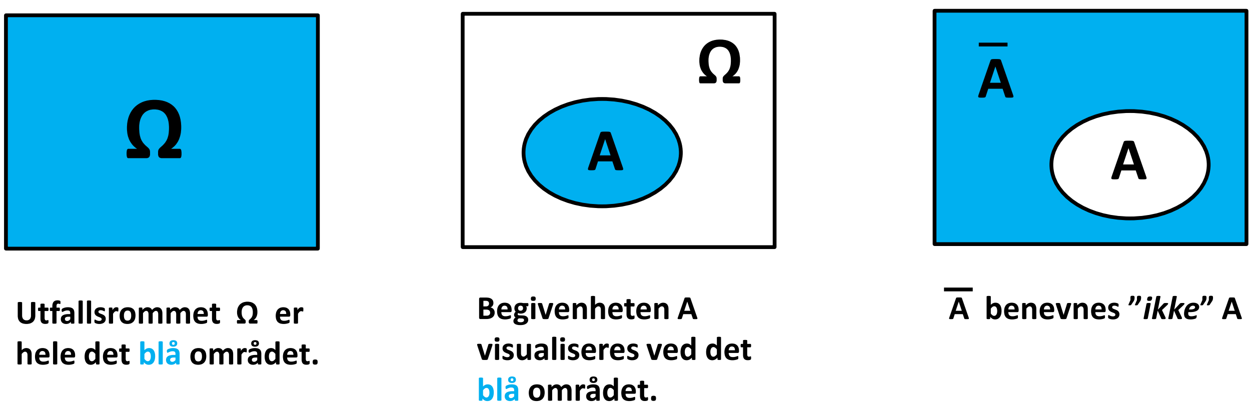


Figur A.1: Mengdelære.

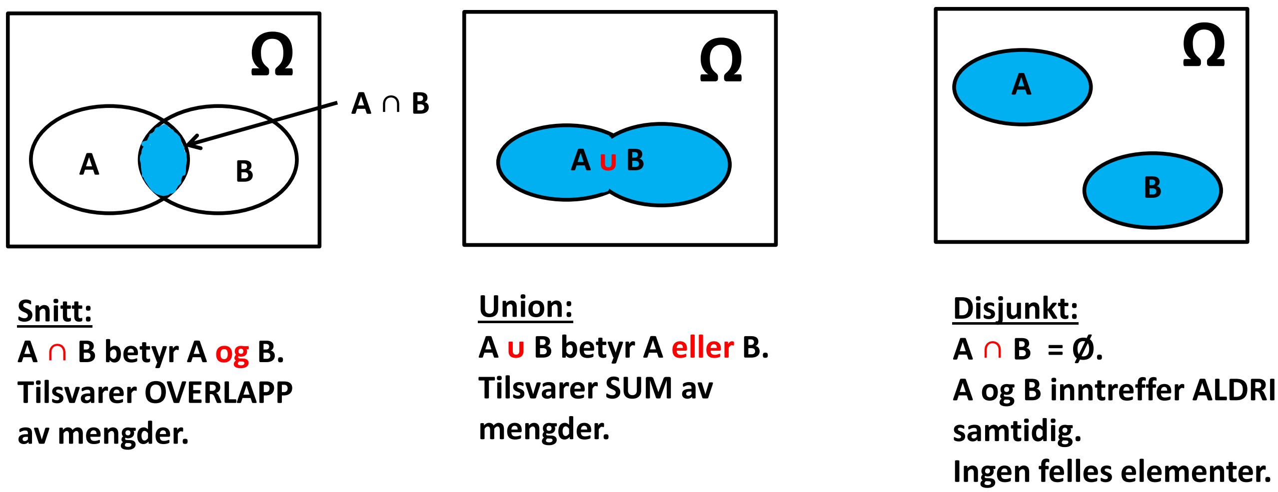
A.1 Venndiagrammer

Venn-diagrammer:

For et eksperiment, la A og B være to begivenheter i utfallsrommet Ω.



Figur A.2: Utfallsrommet Ω, begivenheten A og komplementet $\overline{A}$.



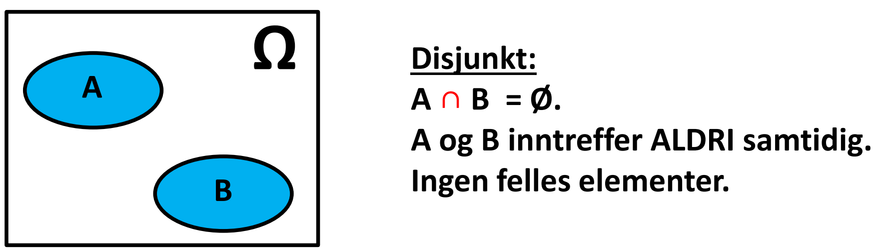
Figur A.3: Snitt, union og disjunkt.

Ut fra disse Venn-diagrammene kan vi merke oss følgende huskeregler:

$\cap$	$=$	og	,	tilsvarer OVERLAPP av mengder
$\cup$	$=$	eller	,	tilsvarer SUM av mengder

Definisjon: (disjunkte mengder)

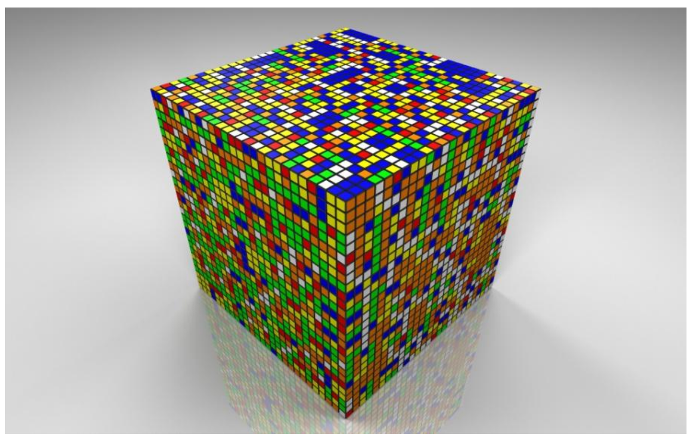
To mengder A og B er disjunkte dersom $A \overbrace{\cap}^{\text{og}} B = \varnothing$.



Figur A.4: A og B er **disjunkt**.

Tillegg B

Kombinatorikk



Figur B.1: Kombinatorikk.

B.1 Koblinger

Definisjon: (koblinger)

Koblinger = forhold som gjør at et bestemt valg kan påvirke utfallet av andre valg vi skal gjøre.

■

Grunnprinsipp i kombinatorikk: (antagelse)

Ingen kobling mellom mellom valgmulighetene. (Med koblinger blir det fort vanskelig).

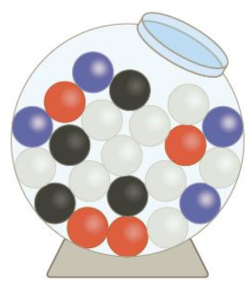
B.2 4 situasjoner (endelig populasjon)

Det kan være vanskelig å telle opp antall elementer i utfallsrommet.
Men dersom:

- uniformt utfallsrom
- utfallene ikke påvirker hverandre, dvs. de er uavhengige

så kan derfor ofte lønne seg å bruke

"**urnemodellen**"



	trekning	
	m/tilbakelegging	u/tilbakelegging
ordnet	situasjon 1	situasjon 2
ikke-ordnet	situasjon 4 (forekommer sjelden)	situasjon 3

Figur B.2: **4 situasjoner** for urnemodellen.

	trekning	
	m/tilbakelegging	**u/tilbakelegging**
ordnet	**Situasjon 1:** **# komb. =** N^s	**Situasjon 2:** **# komb. =** $\frac{N!}{(N-s)!}$
ikke-ordnet	**situasjon 4** (forekommer sjelden) (ikke pensum)	**Situasjon 3:** **# komb. =** $\binom{N}{s}$

Disse formlene gjelder for alle s: **s ≥ 0**

N = antall elementer totalt

s = antall <u>valgte</u> elementer

s=0:
Ingen valgte elementer.

s=1:
Situasjon 1 = situasjon 2 = situasjon 3 = situasjon 4
(alle formlene <u>reduserer seg til det samme</u>: # komb. = N)

s ≥ 2:
Formlene er forskjellige.

Figur B.3: Antall kombinasjoner.

B.3 Binomialkoeffisienten

Egenskaper til **binomialkoeffisienten** $\binom{N}{s}$: ("fra N trekk s")

$$\boxed{\underbrace{\binom{N}{s}}_{\text{binomialkoeff.}} = \frac{N!}{(N-s)!\,s!} \qquad \text{(B.1)}}$$

hvor f.eks.

$$s! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (s-2) \cdot (s-1) \cdot s \qquad \text{(B.2)}$$

$$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \qquad \text{(B.3)}$$

Legg merke til at

$$\binom{N}{0} = \frac{N!}{(N-0)!\,0!} = 1 \qquad \text{(B.4)}$$

$$\binom{N}{1} = \frac{N!}{(N-1)!\,1!} = N \qquad \text{(B.5)}$$

$$\binom{N}{N} = \frac{N!}{(N-N)!\,N!} = 1 \qquad \text{(B.6)}$$

siden $0! = 1$.

B.4 Kombinatoriske sannsynligheter

$$\boxed{P(A) \;=\; \frac{\text{antall } \textit{gunstige} \text{ kombinasjoner for } A}{\text{antall } \textit{mulige} \text{ kombinasjoner totalt}} \qquad \text{(B.7)}}$$

Tillegg C

Kvantiltabell for normalforldeling

Kvantiltabell for normalfordeling , $P(Z \leq z)$

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	–	–	–	–	–	–	–	–	–