

Forventning og varians.

Dekkes av kapittel 4 i læreboka.

Forventning (4.1)

Forventningsverdi = “gjennomsnitt i det lange løp”.

Definisjon: Forventningsverdien til en stokastisk variabel X er:

$$E(X) = \begin{cases} \sum_x x f(x) & ,X \text{ diskret} \\ \int_{-\infty}^{\infty} x f(x) dx & ,X \text{ kontinuert} \end{cases}$$

Bruker ofte μ eller μ_X som symbol for $E(X)$. $\square$

Eksempel: Myntkast, kaster en mynt 3 ganger.
 X = ”antall kron”. Vist tidligere at:

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

Gir at

$$E(X) = \sum_x x f(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \underline{\underline{\frac{3}{2}}} \quad \square$$

Eksempel: Y = ”funksjonstid til elektronisk komponent”. Y har sannsynlighetstetthet

$$f(y) = \begin{cases} \frac{1}{5}e^{-y/5} & , y \geq 0 \\ 0 & , \text{ellers} \end{cases}$$

som gir at

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f(y) dy = \int_0^{\infty} y \frac{1}{5} e^{-y/5} dy \\ &= [-ye^{-y/5}]_0^{\infty} - \int_0^{\infty} -e^{-y/5} dy \\ &= 0 - [5e^{-y/5}]_0^{\infty} = \underline{\underline{5}} \end{aligned}$$

Dvs tester vi mange komponenter av denne typen vil vi se at de i gjennomsnitt har en funksjonstid på rundt 5. □

Forventningsverdi til funksjoner av stokastiske variable finnes på tilsvarende måte:

Definisjon:

$$E(g(X)) = \begin{cases} \sum_x g(x) f(x) & , X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & , X \text{ kontinuert} \end{cases}$$

Eksempel: Se boka/øving.

Definisjon: Dersom X og Y er to stok. var. med simultanfordeling $f(x, y)$ er

$$E(g(X, Y)) = \begin{cases} \sum_x \sum_y g(x, y) f(x, y) & , \text{ disk.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy & , \text{ kont.} \end{cases}$$

Eksempel: To kontinuerlige stok. var. har simultanfordeling

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4} & , 0 < x < 2 , 0 < y < 1 \\ 0 & , \text{ellers} \end{cases}$$

Finn $E(Y/X)$:

$$E\left(\frac{Y}{X}\right) = \int_0^1 \int_0^2 \frac{y}{x} \frac{x(1+3y^2)}{4} dx dy = \underline{\underline{\frac{5}{8}}} \quad \square$$

Merk (viktig!): $E\left(\frac{Y}{X}\right) \neq \frac{E(Y)}{E(X)}$

Varians og kovarians (4.2)

Varians = mål på spredning omkring forventningsverdien.

Definisjon: Variansen til stok. var. X er:

$$\begin{aligned}\sigma^2 &= \sigma_X^2 = \text{Var}(X) = \text{E}[(X - \mu)^2] \\ &= \begin{cases} \sum_x (x - \mu)^2 f(x) & , X \text{ diskret} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & , X \text{ kontinuerlig} \end{cases}\end{aligned}$$

$\sigma = \text{SD}(X) = \sqrt{\text{Var}(X)}$ kalles standardavviket til X . $\square$

Teorem:

$$\text{Var}(X) = \text{E}(X^2) - \mu^2 = \text{E}(X^2) - \text{E}(X)^2 \quad \square$$

Bevis: Se boka.

Eksempel: Y = ”funksjonstid til elektronisk komponent”.

$$\begin{aligned}\text{E}(Y^2) &= \int_0^{\infty} y^2 \frac{1}{5} e^{-y/5} dy \\ &= \dots = 50\end{aligned}$$

$$\Rightarrow \text{Var}(Y) = \text{E}(Y^2) - \text{E}(Y)^2 = 50 - 5^2 = \underline{\underline{25}} \quad \square$$

Tommelfingerregel: Ofte (men ikke alltid...) vil ca 95% av observasjonene ligge i intervallet $[\mu - 2\sigma, \mu + 2\sigma]$.

Definisjon:

$$\text{Var}(g(X)) = E[(g(X) - \mu_{g(X)})^2] = E[g(X)^2] - \mu_{g(X)}^2 \quad \square$$

(Se eksempler i boka).

Definisjon: Kovariansen til X og Y er gitt ved:

$$\begin{aligned} \sigma_{XY} &= \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \\ &= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) & , \text{disk.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy & , \text{kont.} \end{cases} \end{aligned}$$

Kovarians = mål på samvariasjon.

$\text{Cov}(X, Y) > 0$, “positiv” sammenheng mellom X og Y (f.eks. høyde/skonummer),

$\text{Cov}(X, Y) < 0$, “negativ” sammenheng mellom X og Y (f.eks. alder bil/pris bil),

Merk: X og Y uavh. $\Rightarrow \text{Cov}(X, Y) = 0$.

NB! Motsatt gjelder ikke. To variable kan ha $\text{Cov}(X, Y) = 0$ og likevel være avhengige!

Teorem:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) \quad \square$$

Bevis: Se boka.

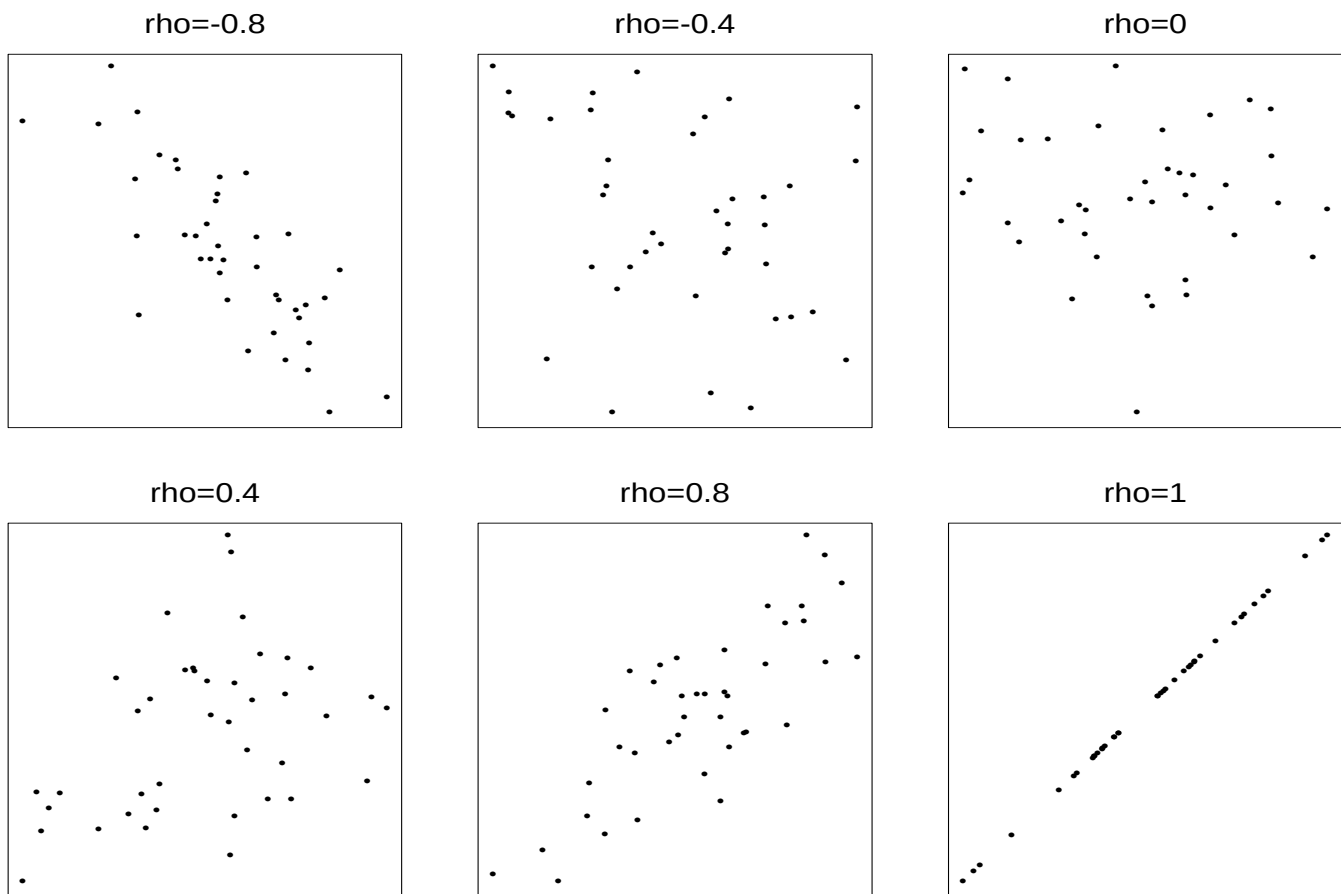
Problem: Kovariansen ofte vanskelig å tolke, avhenger av skala. Bruker derfor ofte korrelasjon som er en normalisering av kovariansen.

Definisjon: Korrelasjon

$$\rho(X, Y) = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$-1 \leq \rho(X, Y) \leq 1$$

Korrelasjon = grad av lineær sammenheng mellom X og Y :



Eksempel: Vann, X = “relativt vannforbruk”, Y = “relativt humusinnhold”.

$$f(x, y) = \begin{cases} x - \frac{1}{2}y + \frac{3}{4} & , 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & , \text{ellers} \end{cases}$$

Finn $\text{Corr}(X, Y)$:

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dy dx \\
&= \int_0^1 \int_0^1 x(x - \frac{1}{2}y + \frac{3}{4}) dy dx = \frac{7}{12} \\
E(Y) &= \int_0^1 \int_0^1 y(x - \frac{1}{2}y + \frac{3}{4}) dy dx = \frac{11}{24} \\
E(XY) &= \int_0^1 \int_0^1 xy(x - \frac{1}{2}y + \frac{3}{4}) dy dx = \frac{13}{48} \\
\text{Cov}(X, Y) &= \frac{13}{48} - \frac{7}{12} \frac{11}{24} = \frac{1}{288}
\end{aligned}$$

Tilsvarende utregninger gir at $E(X^2) = \frac{5}{12}$ og $E(Y^2) = \frac{7}{24}$, som videre gir at $\text{Var}(X) = \frac{5}{12} - (\frac{7}{12})^2 = \frac{11}{144}$ og $\text{Var}(Y) = \frac{7}{24} - (\frac{11}{24})^2 = \frac{47}{576}$ og dermed blir

$$\rho(X, Y) = \frac{1/288}{\sqrt{(11/144) \cdot (47/576)}} = \underline{\underline{0.04}}$$

dvs praktisk talt ingen korrelasjon mellom vannforbruk (X) og humusinnhold (Y). $\square$

Det foregående eksemplet viser at dersom man kjenner simultanfordelingen til to variable så kan man regne ut korrelasjonen mellom variablene.

I praksis er situasjonen ofte at man ikke kjenner simultanfordelingen, men man har noen data (målinger) over sammenhørende verdier. Da kan man fra dataene regne ut estimert, eller *empirisk* korrelasjon.

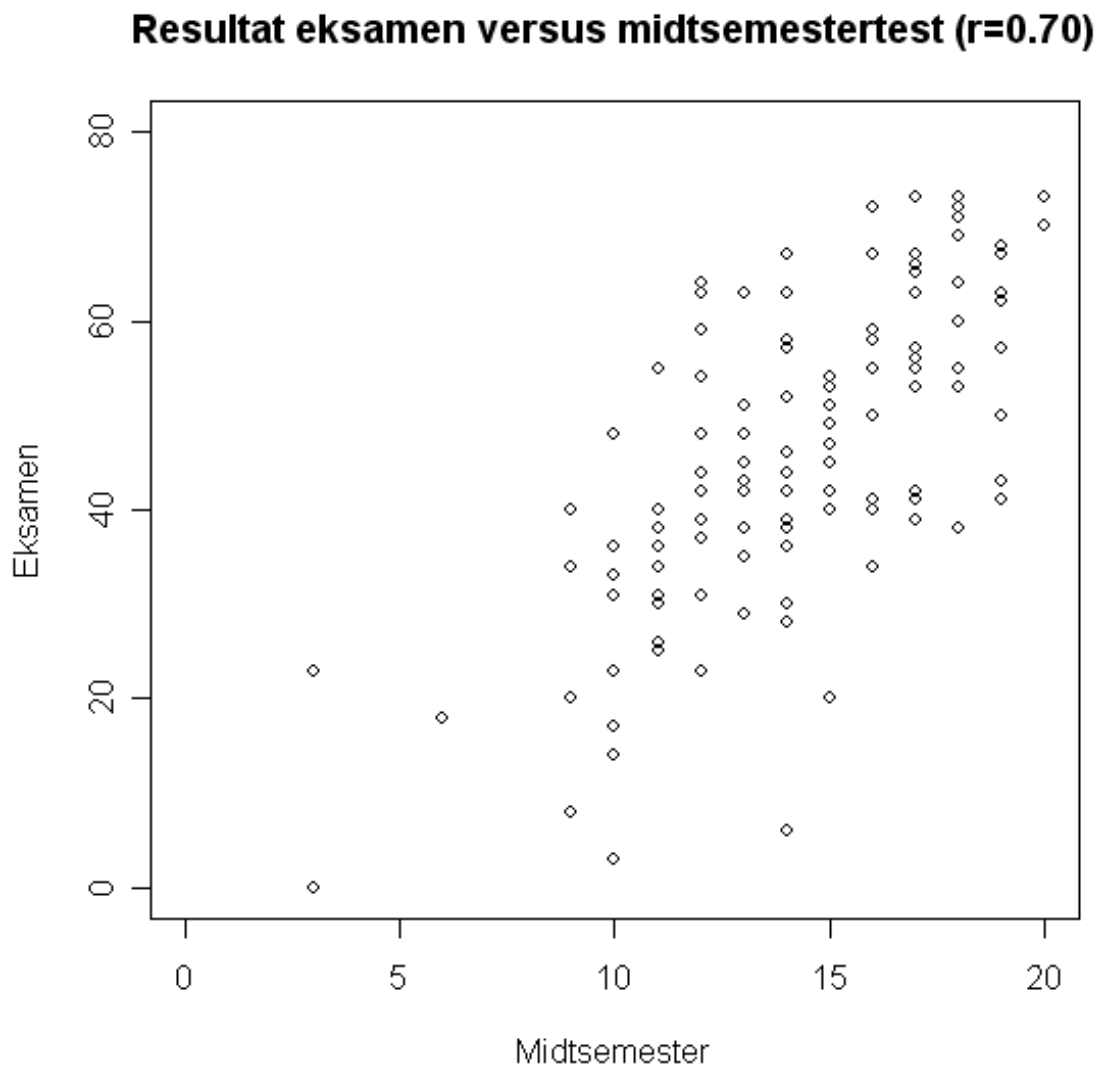
Definisjon: Empirisk korrelasjon (ofte bare kalt korrelasjon) for n målinger $(x_1, y_1), \dots, (x_n, y_n)$:

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$-1 \leq r \leq 1$$

Her er $\bar{x}$ og $\bar{y}$ gjennomsnittene av x - og y -målingene, s_{XY} empirisk kovarians og s_X^2 og s_Y^2 empiriske varianser.

Eksempel: Resultat på midtsemester og slutteksamen.



I del 2 av MOT110 kommer dere til å lære om autokovarians og autokorrelasjon, som er kovarians/korrelasjon mellom etterfølgende målinger i en tidsrekke.

Forventning og varians til lineærkombinasjoner (4.3)

Ved å bruke definisjonene av forventning og varians på lineærkombinasjoner av stokastiske variable får vi en rekke regneregler for lineærkombinasjoner (se boka for alle detaljer). Flesteparten av disse kan oppsummeres i følgende generelle regel (som står i formelsamlingen):

Teorem: La $Y = \sum_{i=1}^n a_i X_i + b$. Da er

$$\begin{aligned} E(Y) &= \sum_{i=1}^n a_i E(X_i) + b, \\ \text{Var}(Y) &= \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1}^{i-1} a_i a_j \text{Cov}(X_i, X_j). \quad \square \end{aligned}$$

Se bevis for en del av spesialtilfellene i boka.

Merk spesielt at regelen gir at $E(aX + b) = aE(X) + b$ og at $\text{Var}(aX + b) = a^2 \text{Var}(X)$ (f.eks. blir $\text{Var}(-X) = \text{Var}(X)$...). Husk også at ved uavhengighet mellom variable blir $\text{Cov} = 0$.

Eksempel: Vi skal bruke et måleinstrument til å måle pH-verdien til en løsning. Vi vet at målinger med dette instrumentet gir resultater som har en fordeling med forventningsverdi μ (sann pH-verdi) og varians $\sigma^2 = 0.01$. I stedet for å bare gjøre en måling bestemmer vi oss for å gjøre 10 målinger og bruke gjennomsnittet av disse 10 målingene som vårt resultat. Hva blir forventningsverdi og varians til dette gjennomsnittet? La $\bar{X} = \frac{1}{10}(X_1 + X_2 + \dots + X_{10}) = \frac{1}{10}X_1 + \dots + \frac{1}{10}X_{10}$ være gjennomsnittet av de 10 målingene. Vi antar at de 10 målingene er uavhengige.

$$\begin{aligned} E(\bar{X}) &= \frac{1}{10}E(X_1) + \dots + \frac{1}{10}E(X_{10}) \\ &= \frac{1}{10}\mu + \dots + \frac{1}{10}\mu = \mu \\ \text{Var}(\bar{X}) &= \left(\frac{1}{10}\right)^2 \text{Var}(X_1) + \dots + \left(\frac{1}{10}\right)^2 \text{Var}(X_{10}) \\ &= \left(\frac{1}{10}\right)^2 \sigma^2 + \dots + \left(\frac{1}{10}\right)^2 \sigma^2 \\ &= \frac{\sigma^2}{10} = \frac{0.01}{10} = 0.001 \end{aligned}$$

Dvs, gjennomsnittet har samme forventningsverdi som enkeltmålingene men lavere varians. Bruker vi tommelfingerregelen om at omtrent 95% av resultatene vil havne i intervallet $\mu \pm 2\sigma$ får vi at en enkeltmåling har ca 95% sannsynlighet for å havne innenfor $[\mu - 2 \cdot \sqrt{0.01}, \mu + 2 \cdot \sqrt{0.01}] = [\mu - 0.2, \mu + 0.2]$, mens gjennomsnittet av 10 målinger har ca 95% sannsynlighet for å havne innenfor $[\mu - 2 \cdot \sqrt{0.001}, \mu + 2 \cdot \sqrt{0.001}] = [\mu - 0.06, \mu + 0.06]$.