STK-IN1050 - Statistikk for informatikere

7. Egenskaper til stokastiske variabler

Matematisk institutt

2026-02-09

Egenskaper til stokastiske variabler

- Hvordan beskrive lokasjon, variasjon og andre egenskaper av en stokastisk variabels fordeling?
- Hvilke regneregler og verktøy er nyttige når vi jobber med fordelinger til stokastiske variabler?
- Hvordan definere simultanfordeling og mål for sammenheng mellom to (eller flere) stokastiske variabler?

Forventningsverdi

En diskret stokastisk variabel X har **forventningsverdi** (eller **forventning**):

$$\mathbb{E}(X) = \sum_x x\, P(X = x)$$

En kontinuerlig stokastisk variabel X har forventning:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

Forventningsverdi

I det diskrete tilfelle tilsvarer forventningsverdien det aritmetiske gjennomsnittet

$$\bar{x} = \sum_j a_j f_j,$$

hvor $a_1, a_2, \ldots$ er de mulige verdiene og f_j er den relative frekvensen av a_j.

Men: forventningsverdien beskriver en egenskap til den teoretiske sannsynlighetsfordelingen til en tilfeldig variabel (før vi har observert data).

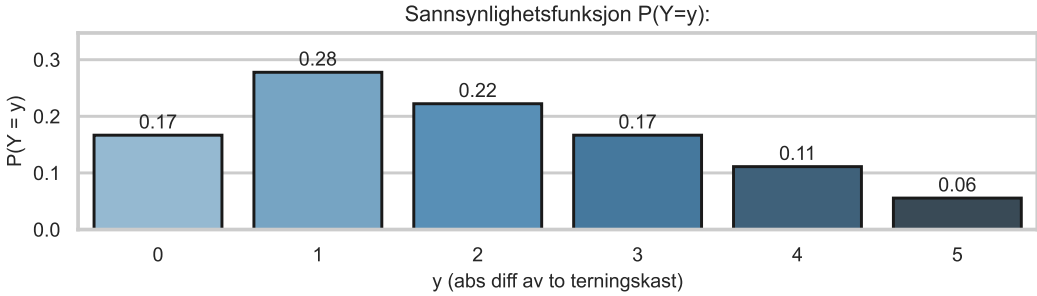
Forventningsverdien beskriver spesifikt sannsynlighetsfordelingens lokasjon (tyngdepunkt).

Eksempel: Differens av to terningskast

La Y være den absolute differensen av to terningskast. Da er $\mathcal{Y} = \{0, 1, \ldots, 5\}$ og de mulige verdiene har sannsynligheter

- $P(Y = 0) = 6/36$
- $P(Y = 1) = 10/36$
- $P(Y = 2) = 8/36$
- $P(Y = 3) = 6/36$
- $P(Y = 4) = 4/36$
- $P(Y = 5) = 2/36$

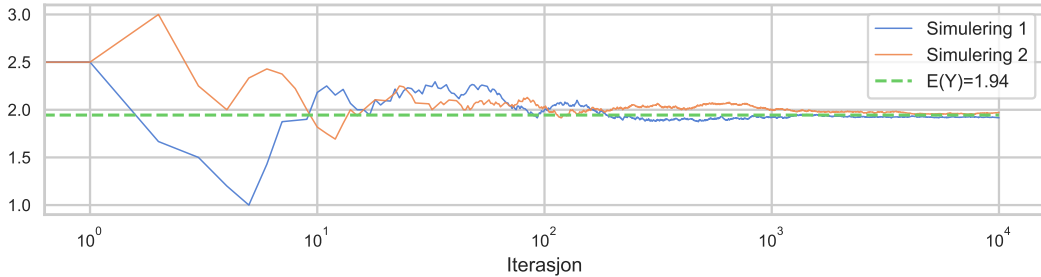
hvor summen av sannsynlighetene er lik 1.



Y har forventning:

$$\mathbb{E}(Y) = \sum_{i=1}^{6} y_i \, P(Y = y_i) = 0{\cdot}0.17+1{\cdot}0.28+2{\cdot}0.22+3{\cdot}0.17+4{\cdot}0.11+5{\cdot}0.06 \approx 1.94.$$

Hvis vi gjentar eksperimentet flere ganger, så vil det aritmetiske gjennomsnittet nærme seg forventningsverdien:



Vi kan altså tolke forventningen som gjennomsnittlig eller typisk realisering av den stokastiske variabelen med hyppig repetisjon.

Store talls lov

Store talls lov

Anta at $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variabler med $\mathbb{E}(X_i) = \mu$ for alle $i = 1, \ldots, n$. Da gjelder at

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \quad \text{når } n \to \infty.$$

Dvs. det aritmetiske gjennomsnittet konvergerer mot forventningsverdien når utvalgsstørrelsen øker.

Forventning: Egenskaper

Obs: Noen fordelinger mangler forventningsverdi!

Spesifikt, så eksisterer forventningsverdien hvis

$$\sum_i |x_i| P(X = x_i) < +\infty$$

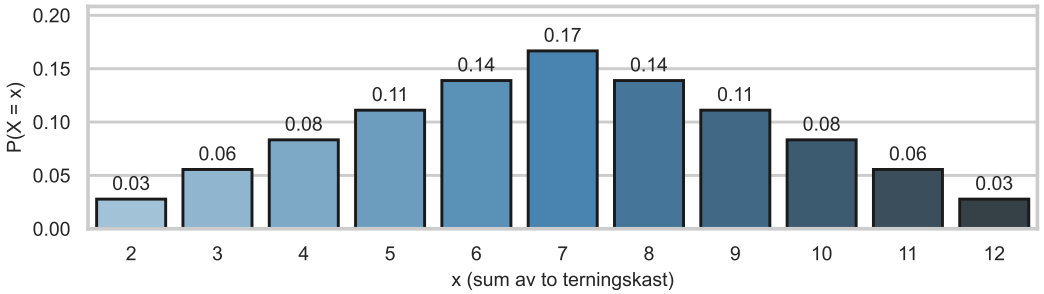
når X er diskret og hvis

$$\int_{-\infty}^{+\infty} |x| f(x) dx < +\infty$$

når X er kontinuerlig.

Hvis fordelingen er symmetrisk omkring a (og $E(X)$ eksisterer), så er $E(X) = a$.

Eksempel: La X være summen av to terningskast. Da har X fordeling gitt ved:



Den er symmetrisk om 7 og X har forventningsverdi

$$\mathbb{E}(X) = \sum_{i=1}^{12} x_i p_i = 2{\cdot}0.03{+}3{\cdot}0.06{+}4{\cdot}0.08{+}5{\cdot}0.11{+}6{\cdot}0.14{+}7{\cdot}0.17{+}8{\cdot}0.14{+}9{\cdot}0.11{+}10{\cdot}0.08{+}11{\cdot}0.06{+}12{\cdot}0.03 = 7.$$

Forventning: Transformasjoner

Lineær transformasjon:

For $Y = aX + b$ (med X stokastisk variabel og a, b reelle tall) gjelder at

$$\mathbb{E}(Y) = a\mathbb{E}(X) + b.$$

Ikke-lineære transformasjoner:

For $Y = g(X)$ (hvor g er en funksjon og X en stokastisk variabel) gjelder at

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \begin{cases} \sum g(x)P(X = x), & \text{hvis X er diskret,} \\ \int g(x)f(x)\,dx, & \text{hvis X er kontinuerlig.} \end{cases}$$

Generelt: $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

Eksempel: Transformasjoner

Anta at X er en stokastisk variabel med verdimengde $\mathcal{X} = \{0, 1\}$ hvor

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p,$$

for $0 \leq p \leq 1$. Anta videre at $g(x) = x^2$. Da gjelder

$$\mathbb{E}(X) = 0 \cdot (1 - p) + 1 \cdot p = p,$$
$$\mathbb{E}(g(X)) = \mathbb{E}(X^2) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p,$$
$$g(\mathbb{E}(X)) = g(p) = p^2.$$

Forventningsverdi: Linearitet og produkter

Forventningverdien til summen av stokastiske variabler:

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n)$

Produktregel for uavhengige stokastiske variabler:

- Hvis X og Y er uavhengige, så gjelder at

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

- Obs: Uavhengighet er avgjørende her!

Varians og standardavvik

En stokastisk variabel X har **varians** gitt ved

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \\ &= \begin{cases} \sum (x - \mathbb{E}(X))^2 P(X = x), & \text{hvis X er diskret,} \\ \int (x - \mathbb{E}(X))^2 f(x)\, dx, & \text{hvis X er kontinuerlig.} \end{cases}\end{aligned}$$

Standardavviket til X er gitt ved $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$.

Variansen (og standardavviket) er summariske mål som sier noe om «spredningen», dvs. hvor mye verdien til en stokastisk variabel vil variere fra forsøk til forsøk.

Varians: Egenskaper

Obs: Noen fordelinger mangler varians!

Spesifikt, så eksisterer variansen hvis

$$\sum_i |x_i|^2 P(X = x_i) < +\infty$$

når X er diskret og hvis

$$\int_{-\infty}^{+\infty} |x|^2 f(x) dx < +\infty$$

når X er kontinuerlig.

Regneregler for varians

En alternativ formel for variansen er

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$$

Lineær transformasjon: Hvis $Y = aX + b$ (med X stokastisk variabel og a, b reelle tall), da blir

$$\mathrm{Var}(Y) = a^2 \mathrm{Var}(X)$$

og

$$\mathrm{sd}(Y) = |a| \, \mathrm{sd}(X).$$

Regneregler for varians

Anta at X og Y er uavhengige (a_1, a_2 reelle tall). Da gjelder:

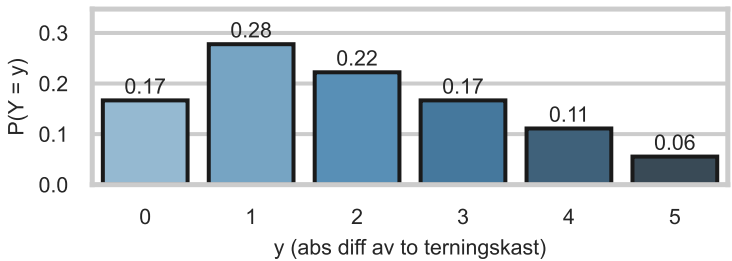
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$
$$\mathrm{Var}(a_1 X + a_2 Y) = a_1^2 \mathrm{Var}(X) + a_2^2 \mathrm{Var}(Y)$$

Generelt for uavhengige $X_1, \ldots, X_n$ ($a_1, \ldots, a_n$ reelle tall):

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i \Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$
$$\mathrm{Var}\Big(\sum_{i=1}^{n} a_i X_i \Big) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i)$$

Eksempel: Differens av to terningskast

La Y være den absolute differensen av to terningskast. Da har Y fordeling



Vi har at $\mathbb{E}(Y) \approx 1.94$. Dermed blir $\text{Var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 \approx 5.95 - 1.94^2 \approx 2.19$:

$$\mathbb{E}(Y^2) = \sum_{i=1}^{6} y_i^2 \, P(Y = y_i) = 0^2 \cdot 0.17 + 1^2 \cdot 0.28 + 2^2 \cdot 0.22 + 3^3 \cdot 0.17 + 4^2 \cdot 0.11 + 5^2 \cdot 0.06 \approx 5.95.$$

Eksempel: Prøv selv

Anta at X og Y er uavhengige stokastiske variabler hvor

- X har forventning μ_X og Y har forventning μ_Y
- X har varians σ_X^2 og Y har varians σ_Y^2

1. Hva er forventningen og variansen til $X - Y$?
2. Hva er forventningen og variansen til $X + Y$?
3. Hva er forventningen og variansen til $2X$?

Eksempel: Chuck-a-Luck (terningspill)

- Tre terninger kastes.
- Mål: Å gjette hvilket tall (fra 1 til 6) som vil vises på de tre terningene.
- Innsats: Spillere plasserer innsatser på ett eller flere tall.

Utbetaling:

- En terning viser tallet ditt: Du vinner 1:1 (får innsatsen tilbake + like mye i gevinst).
- To terninger viser tallet ditt: Du vinner 2:1.
- Tre terninger viser tallet ditt: Du vinner 3:1.
- Hvis tallet ditt ikke dukker opp, taper du innsatsen.

Chuck-a-Luck: Sannsynlighetsfordeling (én innsats på 6)

Gevinst X_6	Kombinasjoner	Antall	Sannsynlighet
3	666	1	$1/216$
2	$66a, 6a6, a66$ ($a \in \{1,\ldots,5\}$)	15	$15/216$
1	$6ab, a6b, ab6$ ($a, b \in \{1,\ldots,5\}$)	75	$75/216$
-1	abc (ingen 6, $a, b, c \in \{1,\ldots,5\}$)	125	$125/216$
Sum		216	1

Forventningsverdien er gitt ved

$$E(X_6) = 3 \cdot \frac{1}{216} + 2 \cdot \frac{15}{216} + 1 \cdot \frac{75}{216} - 1 \cdot \frac{125}{216} = -0.078,$$

dvs. vi forventer å tape 7.8% av det vi satser (i det lange løpet).

Chuck-a-Luck: Spillstrategier for to kast

For å vurdere forskjellige spillstrategier for to kast (gjentakelser av spillet), definerer vi noen stokastiske variabler:

- $X_1, \ldots, X_6$: gevinst hvis én innsats på henholdsvis $1, 2, \ldots, 6$ i første kast.
- $Y_1, \ldots, Y_6$: gevinst hvis én innsats på henholdsvis $1, 2, \ldots, 6$ i andre kast.

Vi skal se nærmere på følgende strategier:

- $2X_6$ (dobbel innsats på 6 i første kast)
- $X_1 + X_6$ (én innsats på 1 og én innsats på 6 i første kast)
- $X_6 + Y_6$ (én innsats på 6 i begge kast)

For å sammenligne disse strategiene, må vi se på både forventninger og varianser (risiko).

Chuck-a-Luck: Forventet gevinst

Vi har $\mathbb{E}(X_i) = \mathbb{E}(Y_i) = -17/216$.

Fra regnereglene for forventning følger dermed at

$$\mathbb{E}(2\,X_6) = 2\,\mathbb{E}(X_6) = -\frac{34}{216}$$

$$\mathbb{E}(X_1 + X_6) = \mathbb{E}(X_1) + \mathbb{E}(X_6) = -\frac{34}{216}$$

$$\mathbb{E}(X_6 + Y_6) = \mathbb{E}(X_6) + \mathbb{E}(Y_6) = -\frac{34}{216}$$

Forventningsverdien er den samme for alle tre strategier!

Chuck-a-Luck: Sannsynlighetsfordelinger

Selv om forventningsverdien er den samme for alle tre strategier, så er det større forskjeller mellom disse strategiene.

De tre strategiene har forskjellige verdimengder med forskjellige sannsynligheter for enkelte verdier:

Strategi	Verdimengde	$P(\{-2\})$
$2X_6$	{ -2, 2, 4, 6}	0.579
$X_1 + X_6$	{ -2, 0, 1, 2, 3}	0.296
$X_6 + Y_6$	{ -2, 0, 1, 2, 3, 4, 5, 6}	0.335

Chuck-a-Luck: Varians

Variansen er også ganske forskjellig for de tre strategiene:

$$E(X_6^2) = 3^2 \cdot \frac{1}{216} + 2^2 \cdot \frac{15}{216} + 1^2 \cdot \frac{75}{216} + (-1)^2 \cdot \frac{125}{216} = 1.245,$$
$$\text{Var}(X_6) = \mathbb{E}(X_6^2) - (\mathbb{E}(X_6))^2 = 1.245 - (-0.078)^2 = 1.239$$

$$\text{Var}(2 \cdot X_6) = 2^2 \cdot \text{Var}(X_6) = 4 \cdot 1.239 = 4.956$$
$$\text{Var}(X_6 + Y_6) = \text{Var}(X_6) + \text{Var}(Y_6) = 1.239 + 1.239 = 2.478$$

X_6 og Y_6 er uavhengig da de henviser til to uavhengige kast.

Chuck-a-Luck: Varians

X_1 og X_6 henviser til samme kast og er avhengige. Vi kan derfor ikke bruke regnereglene og må regne ut variansen til $X_1 + X_6$ direkte:

s	-2	0	1	2	3
$P(X_1 + X_6 = s)$	0.296	0.444	0.111	0.120	0.028

$$\begin{aligned}\mathbb{E}((X_1 + X_6)^2) &= (-2)^2 \cdot 0.296 + 0^2 \cdot 0.444 + 1^2 \cdot 0.111 + 2^2 \cdot 0.120 + 3^2 \cdot 0.028 \\ &= 2.027\end{aligned}$$

$$\text{Var}(X_1 + X_6) = \mathbb{E}((X_1 + X_6)^2) - (\mathbb{E}(X_1 + X_6))^2 = 2.027 - (-0.157)^2 = 2.002$$

Chuck-a-Luck: Konklusjon

Strategi	Verdimengde	$\mathbb{E}$	Var
$2X_6$	{ -2, 2, 4, 6}	-0.157	4.956
$X_1 + X_6$	{ -2, 0, 1, 2, 3}	-0.157	2.002
$X_6 + Y_6$	{ -2, 0, 1, 2, 3, 4, 5, 6}	-0.157	2.478

- Strategi 1 har størst risiko.
- Strategi 2 har minst risiko, men den maksimale gevinsten er også lavere.

Mye av dataanalyse og modellering går ut på å undersøke og modellere sammenhenger mellom par av variabler

Det er en **sammenheng** mellom to variabler målt på de samme individene (de er **assosiert**) hvis det å kjenne verdien av en av variablene sier deg noe du ellers ikke ville visst om verdien til den andre variabelen.

En **responsvariabel** måler utfallet av en studie. En **forklaringsvariabel** forklarer eller forårsaker forandringer i responsvariabelen.

Assosiasjon mellom to variabler

Viktige aspekter ved et datasett når man skal undersøke assosiasjonen mellom to variabler:

- **Individer**: Identifiser individene og hvor mange de er i datasettet.
- For å undersøke assosiasjonen mellom to variable må man
  - registrere begge variablene for hvert individ
  - vite hvilke målinger av de to variablene som hører til hvert individ
- Klassifiser hver variabel som kategorisk eller kvantitativ.
- Når det er relevant, klassifiser hver variabel som forklarings- eller responsvariabel.

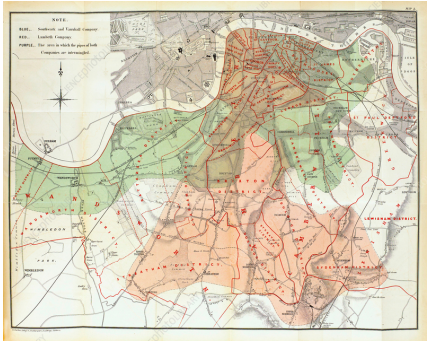
Eksempel: Spredning av kolera

- Midten av 1800-tallet: Koleraepidemi i London
- Ledende teori på den tiden: Spredning via miasma, dvs. ved «vond dunst, forurensning, besudling, smitte» overføres en «sykdomsoverførende materia som oppstår ved forråtnelsesprosesser i luft og vann».
- I virkeligheten skjedde spredningen via bakterielt forurenset vann fra den lokale vannforsyningen.
- Undersøkelsene av epidemien og dens forbindelse med den lokale vannforsyningen regnes som epidemiologiens fødsel[1].

[1] Rothman, Kenneth J., Krista F. Huybrechts, and Eleanor J. Murray, Pioneers in Epidemiology and Public Health, Epidemiology: An Introduction, 3rd edn.

Vannleverandør	Husholdninger	Koleradødsfall	$\frac{\text{Koleradødsfall}}{\text{1000 hushold}}$
Southwark/Vauxhall	40 046	1 263	32
Lambeth	26 107	98	4

Sammenhengen mellom vannleverandører og koleradødsfall gir at vi har en statistisk korrelasjon.



Southwark/Vauxhall: grønn
Lambeth: rød

Årsakesammenheng

Når vi undersøker (om det er) en sammenheng mellom to variable, dukker det opp et viktig spørsmål:

Er formålet bare å undersøke hvordan sammenhengen er, eller ønsker vi å vise at en av variablene kan forklare variasjonen i den andre?

Kausalitet er årsakesammenheng: Sammenheng mellom to variable, hvor den ene (responsen) er en konsekvens av den andre (forklaringsvariabelen).

Kausalitet

Kausale sammenhenger er ofte det egentlige målet med studien, men det er vanskelig å bevise:

- Er sammenhengen forårsaket av andre variabler (**confoundere**, eller sammenblandende variabler, eller «lurkende» variabler)?
- Hvilken vei går egentlig årsakesammenhengen?

Selv om vi finner en assosiasjon eller sammenheng så betyr ikke det at vi har påvist årsak.

I dette kurset fokuserer vi på assosiasjon og ikke kausalitet.

Eksempel: assosiasjon ≠ årsak

Det er en sterk sammenheng mellom barns skostørrelse og deres leseferdigheter

- Er det en assosiasjon mellom skostørrelse og leseferdigheter?
- Er det en kausal sammenheng?
- Er det en confounder?

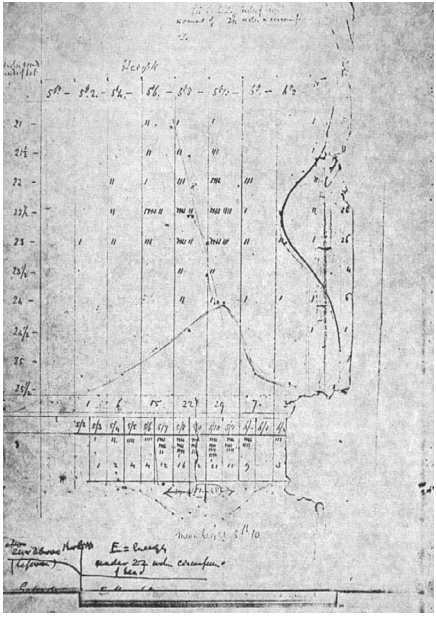
Eksplorativ dataanalyse

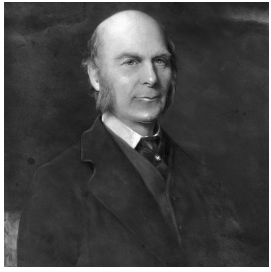
Et **spredningsplott** er det nyttigste grafiske verktøyet for å vise sammenhengen mellom to kvantitative variabler

Et spredningsplott viser sammenhengen mellom to kvantitative variable målt på de samme individene.

Hvert individ tilsvarer ett punkt i grafen.

Verdiene til en variabel vises langs den horisontale aksen, og verdiene til den andre vises langs den vertikale aksen.





Sir Francis Galton (1822-1911) sitt første spredningsplott viser sammenhengen mellom høyde og hodeomkrets. Fra boken «Special Peculiarities».

Bilde fra Hulton Archive/Getty images.

Å lage et spredningsplott

1. Bestem hvilken variabel som skal være langs hvilken akse.

   Hvis det er en naturlig oppdeling i forklarings- og responsvariable, velg forklaringsvariabelen langs x-aksen og responsvariabelen langs y-aksen.
2. Gi navn og skala til aksene.
3. Plott de individuelle data-verdiene.

Eksempel: Leseferdigheter

```
import pandas as pd

df = pd.read_csv("data/readingSkills.txt", sep=r"\s+",
                 quotechar='"', engine="python", header=0)
print(df.head())
type(df)
```

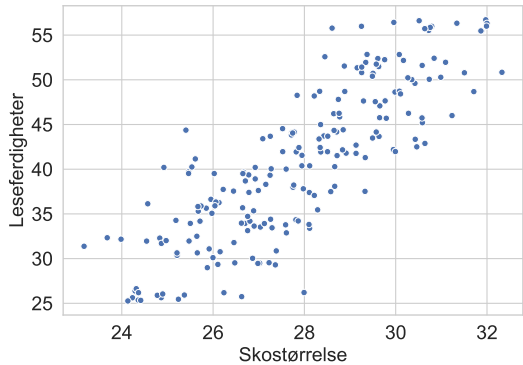
```
    nativeSpeaker  age   shoeSize      score
"1"         "yes"    5  24.831889  32.293850
"2"         "yes"    6  25.952378  36.631049
"3"          "no"   11  30.421700  49.605927
"4"         "yes"    7  28.664501  40.284556
"5"         "yes"   11  31.882070  55.460851

pandas.core.frame.DataFrame
```

Leseferdigheter og skostørrelse

```
import matplotlib.pyplot as plt
import seaborn as sns

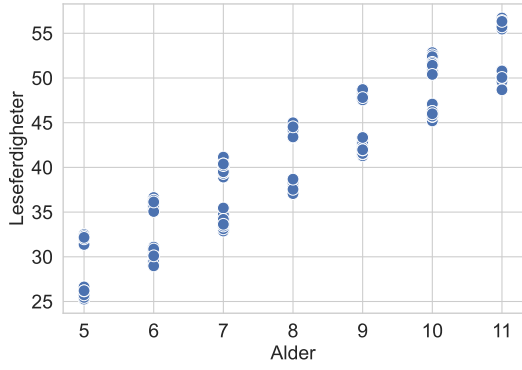
sns.set(style="whitegrid")
plt.figure(figsize=(7,5))
sns.scatterplot(data=df,
                x="shoeSize",
                y="score")
plt.xlabel("Skostørrelse",
           fontsize=18)
plt.ylabel("Leseferdigheter",
           fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.tight_layout()
plt.show()
```



Leseferdigheter og alder

```
import matplotlib.pyplot as plt
import seaborn as sns

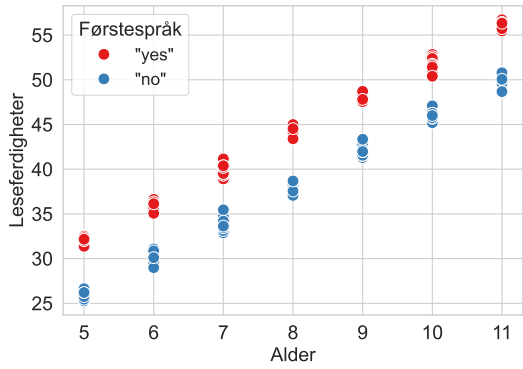
sns.set(style="whitegrid")
plt.figure(figsize=(7,5))
sns.scatterplot(data=df, x="age",
                y="score", s=120)
plt.xlabel("Alder",
           fontsize=18)
plt.ylabel("Leseferdigheter",
           fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.tight_layout()
plt.show()
```



Leseferdigheter og alder: sammenhenger innenfor kategorier (om opplæringsspråk er førstespråk)

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")
plt.figure(figsize=(7,5))
sns.scatterplot(data=df, x="age", y="score",
                s=120, hue="nativeSpeaker",
                palette="Set1")
plt.xlabel("Alder", fontsize=18)
plt.ylabel("Leseferdigheter", fontsize=18)
plt.legend(title="Førstespråk", fontsize = 16,
           title_fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.tight_layout()
plt.show()
```



Finpussing av visualiseringen: katagorinavn

Kategorinavnene skal gjerne være på norsk også:

```
df['native_label'] = df['nativeSpeaker'].map({'"yes"': 'ja', '"no"': 'nei'})
df['native_label'] = pd.Categorical(df['native_label'],
                                    categories=['ja', 'nei'])

print(df.head())
```

```
    nativeSpeaker  age   shoeSize      score native_label
"1"         "yes"    5  24.831889  32.293850           ja
"2"         "yes"    6  25.952378  36.631049           ja
"3"          "no"   11  30.421700  49.605927          nei
"4"         "yes"    7  28.664501  40.284556           ja
"5"         "yes"   11  31.882070  55.460851           ja
```

Finpussing av visualiseringen: alder

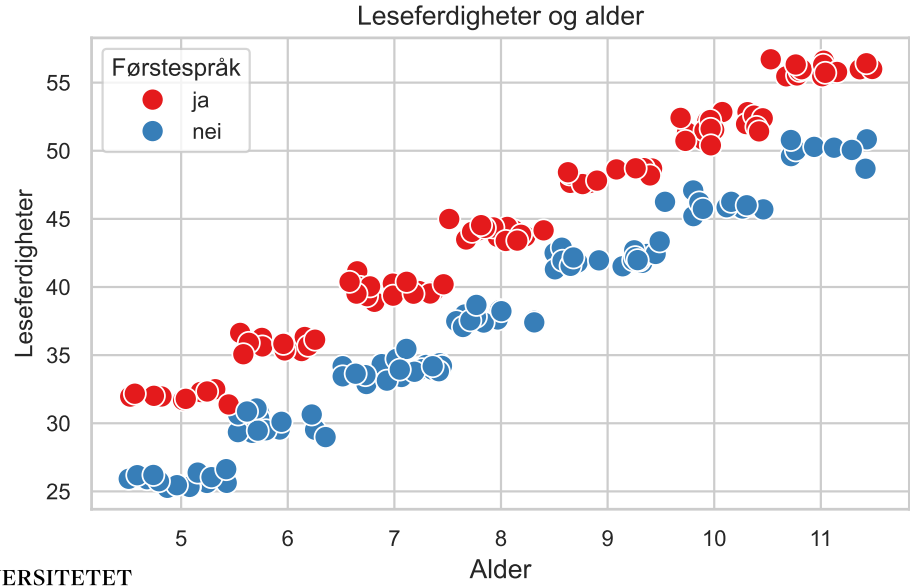
Alder er egentlig en kontinuerlig og ikke kategorisk variabel.

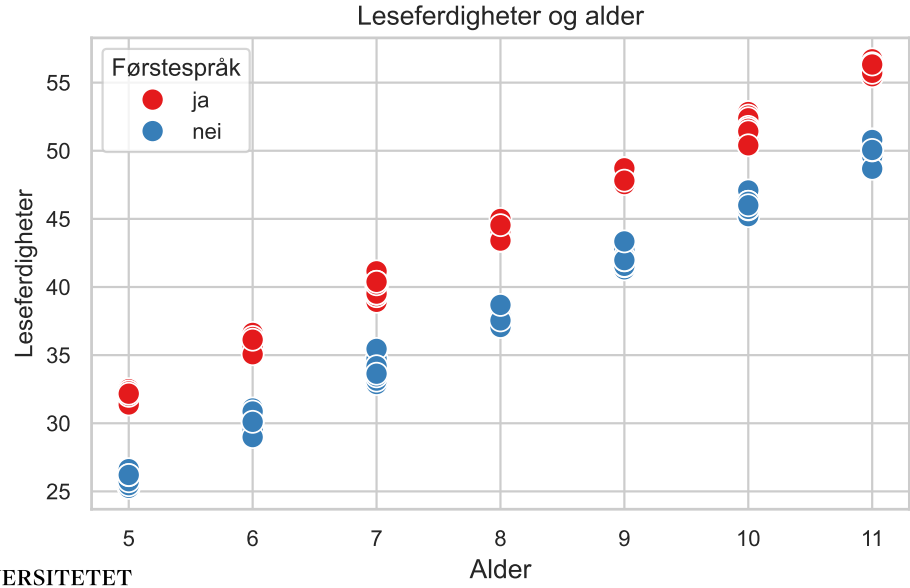
```
import numpy as np

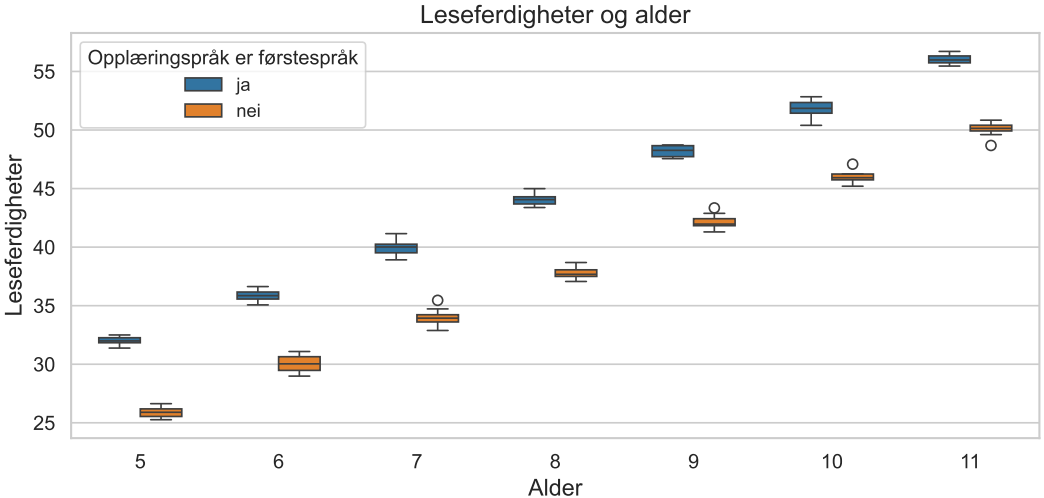
rng = np.random.default_rng(123)
df['age_jitter'] = df['age'] + rng.uniform(low=-0.5, high=0.5, size=len(df))

print(df.head())
```

```
    nativeSpeaker  age   shoeSize      score native_label  age_jitter
"1"          "yes"    5  24.831889  32.293850           ja    5.182352
"2"          "yes"    6  25.952378  36.631049           ja    5.553821
"3"           "no"   11  30.421700  49.605927          nei   10.720360
"4"          "yes"    7  28.664501  40.284556           ja    6.684372
"5"          "yes"   11  31.882070  55.460851           ja   10.675906
```


Leseferdigheter og alder
Førstespråk
ja
nei
Leseferdigheter
55
50
45
40
35
30
25
5
6
7
8
9
10
11
Alder

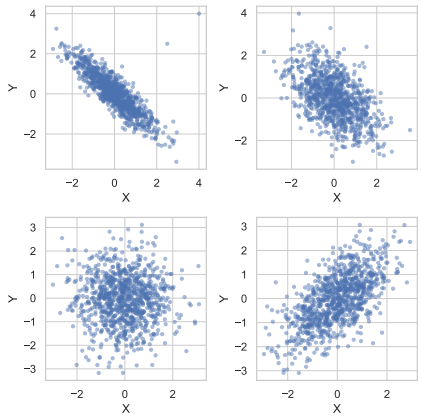

Leseferdigheter og alder
Førstespråk
ja
nei
Leseferdigheter
25
30
35
40
45
50
55
5
6
7
8
9
10
11
Alder


Leseferdigheter og alder
Opplæringspråk er førstespråk
ja
nei
Leseferdigheter
55
50
45
40
35
30
25
5
6
7
8
9
10
11
Alder

Tolkning av spredningsplott

For å tolke et spredningsplott, se etter mønstre og viktige avvik fra disse mønstrene:

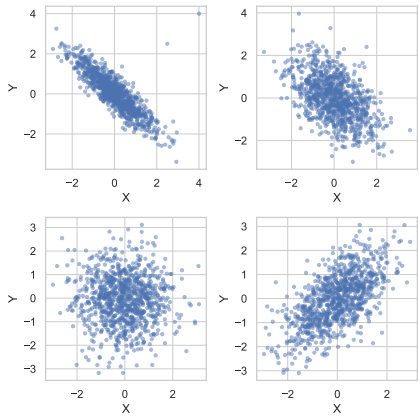
- Vi kan beskrive det overordnede mønsteret til et spredningsplott ved **formen**, **retningen**, og **styrken** til sammenhengen.
- En viktig type avvik er en **outlier**: en individuell verdi som faller utenfor det overordnede mønsteret til sammenhengen.



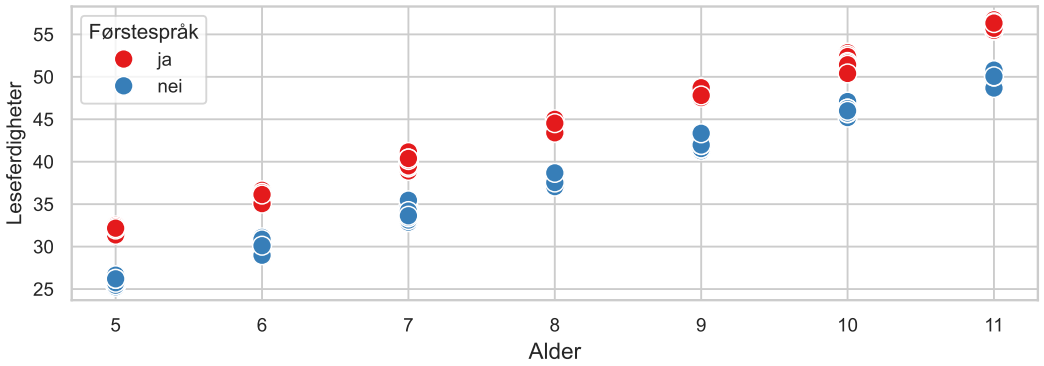
Tolkning av spredningsplott

Positiv og negativ lineær sammenheng:

- To variabler har en **positiv sammenheng** når de følger hverandre i samme retning (begge øker eller minker).
- To variabler har en **negativ sammenheng** når de beveger seg i motsatt retning (én øker mens den andre synker).

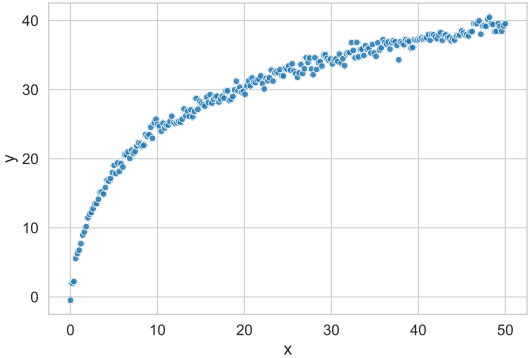


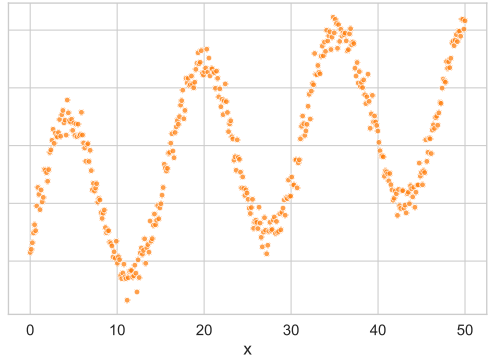
Eksempel: Leseferdigheter



Det er en **sterk**, **positiv**, **lineær** sammenheng mellom barns alder og leseferdigheter.

Det finnes andre former for sammenhenger enn lineære





Korrelasjonskoeffisient

Korrelasjonskoeffisienten r måler styrken til den lineære sammenhengen mellom to kvantitative variabler.

Anta vi har data $x_1, \ldots, x_n$ og $y_1, \ldots y_n$ hvor (x_i, y_i) er et par, dvs. forskjellig informasjon om den samme individen i. Vi kan oppsummere x-ene og y-ene, henholdsvis, ved gjennomsnitt og standardavvik:

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

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

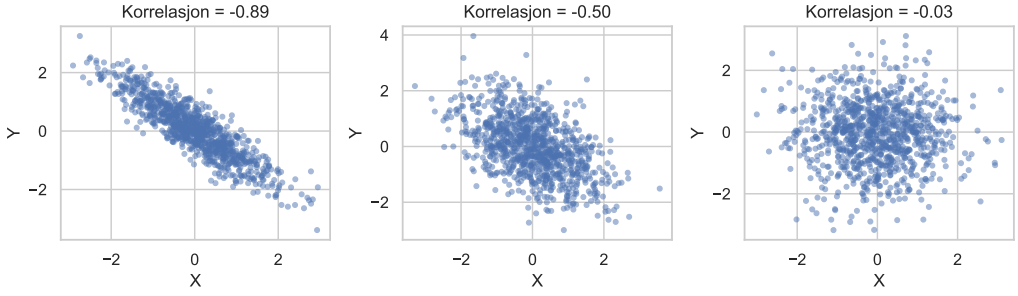
Korrelasjonskoeffisient

Da er **korrelasjonskoeffisienten** (eller **korrelasjonen**) r (eller r_{xy}) gitt ved

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}
\end{aligned}
$$

Egenskaper til korrelasjonen

- r er alltid et tall mellom –1 og 1. $r > 0$ indikerer en positiv sammenheng, $r < 0$ indikerer en negativ sammenheng.
- Verdier av r nær 0 indikerer en veldig svak lineær sammenheng. Styrken til den lineære sammenhengen vokser når r beveger seg bort fra 0 mot –1 eller 1.
- De ekstreme verdiene $r = -1$ og $r = 1$ inntreffer bare dersom det er en perfekt lineær sammenheng.
- Korrelasjonen r er symmetrisk, dvs skiller ikke på forklarings- og responsvariable.
- r har ingen enhet og forandrer seg ikke når vi endrer måleenhet for x eller y eller begge.



UNIVERSITETET
I OSLO

Korrelasjon er ikke en komplett oppsummering av to-variabel-data

- Korrelasjon krever at begge variablene er kvantitative.
- Korrelasjon beskriver ikke ikke-lineære sammenhenger mellom variable, uavhengig av hvor sterk sammenhengene er.
- Korrelasjonskoeffisienten *r* er ikke robust mot outliers, og kan bli sterkt påvirket av noen få outliers i datagrunnlaget.

Simultanfordeling

Anta at X og Y er stokastiske variabler. **Simultanfordelingen** for X og Y beskriver samvariasjonen til X og Y.

Hvis X og Y er diskrete bruker vi den simultane sannsynlighetsfunksjonen

$$P(X = x, Y = y),$$

hvor

- $P(X = x, Y = y) \geq 0$,
- $\sum_x \sum_y P(X = x, Y = y) = 1$.

Merk at $P(X = x, Y = y) \neq P(X = x)P(Y = y)$ (medmindre X og Y er uavhengig).

Simultanfordeling

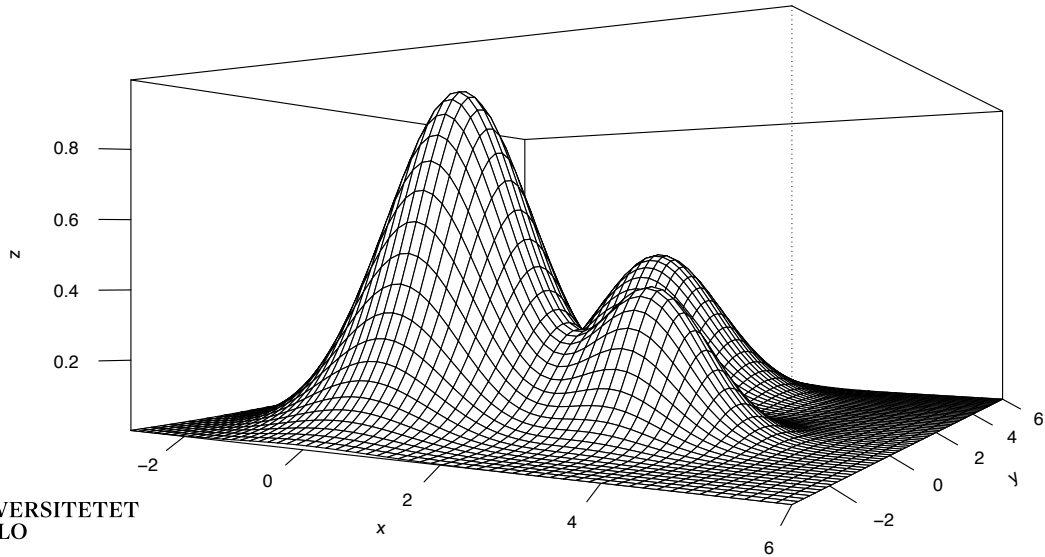
Hvis X og Y er kontinuerlige beskrives simultanfordelingen ved den simultane tetthetsfunksjonen $f(x,y)$ med

$$P(a \leq X \leq b,\ c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\, dy\, dx,$$

hvor

- $f(x,y) \geq 0$,
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)\, dy\, dx = 1$.

Eksempel på en simultan tetthetsfunksjon



Marginalfordeling

Fra simultanfordelingen for X og Y kan vi finne **marginalfordelingen** for X og Y, dvs. fordelingene for X og Y hver for seg.

Hvis X og Y er diskrete, da har vi at

$$P(X = x) = \sum_y P(X = x, Y = y),$$

og hvis X og Y er kontinuerlige, da gjelder at

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

(og tilsvarende for Y).

Betinget fordeling

Anta at X og Y er diskrete stokastiske variabler. Den **betingede fordelingen** til X gitt $Y = y$ er da

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Hvis X og Y er kontinuerlige stokastiske variabler, da er den betingede fordelingen til X gitt $Y = y$ gitt ved følgende tetthetsfunksjon:

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}.$$

Eksempel: marginal- og betinget fordeling

Anta at X og Y er diskrete stokastiske variabler med verdimengder $\mathcal{X} = \{0, 1, 2, 3, 4\}$ og $\mathcal{Y} = \{0, 1, 2, 3\}$, og simultanfordeling gitt ved følgende tabell:

$x \backslash y$	0	1	2	3
0	0.05	0.10	0.01	0.03
1	0.00	0.05	0.08	0.20
2	0.11	0.04	0.09	0.04
3	0.01	0.01	0.03	0.05
4	0.00	0.05	0.05	0.00

- Finn marginalfordelingen $P(X = x)$
- Finn den betingede fordelingen $P(X = x \mid Y = 3)$

Eksempel: marginal- og betinget fordeling

Vi finner marginalfordelingene ved å summere over rekker og kolonner:

$x \backslash y$	0	1	2	3	$P(X = x)$
0	0.05	0.10	0.01	0.03	0.19
1	0.00	0.05	0.08	0.20	0.32
2	0.11	0.04	0.09	0.04	0.28
3	0.01	0.01	0.03	0.05	0.10
4	0.00	0.05	0.05	0.00	0.10
$P(Y = y)$	0.17	0.25	0.26	0.32	1.00

Dermed får vi også

x	0	1	2	3	4
$P(X = x \mid Y = 3)$	$\frac{0.03}{0.32} = 0.09$	$\frac{0.20}{0.32} = 0.63$	$\frac{0.04}{0.32} = 0.12$	$\frac{0.05}{0.32} = 0.16$	$\frac{0.00}{0.32} = 0.00$

Uavhengige stokastiske variabler

To diskrete stokastiske variabler X og Y er **uavhengig** hvis og bare hvis

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

To kontinuerlige stokastiske variabler X og Y er **uavhengig** hvis og bare hvis

$$f(x, y) = f_X(x)f_Y(y).$$

Hvis X og Y er uavhengig, da gjelder at

- $P(X = x \mid Y = y) = P(X = x)$ hvis de er diskrete
- $f_{X|Y}(x \mid y) = f_X(x)$ hvis de er kontinuerlige

(og tilsvarende for Y).

Kovarians for stokastiske variabler

Anta at X og Y er stokastiske variabler. Da er **kovariansen** til X og Y gitt ved

$$\text{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big].$$

Kovarians er et mål på hvordan to stokastiske variabler samvarierer når man gjentar det underliggende stokastiske forsøket.

Det gjelder at

$$\text{Cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y)$$

hvor

$$\mathbb{E}(X \cdot Y) = \begin{cases} \sum_x \sum_y x\, y\, P(X = x, Y = x), & \text{diskret} \\ \int \int x\, y\, f(x, y) dx dy, & \text{kontinuerlig} \end{cases}$$

Korrelasjon for stokastiske variabler

Korrelasjonskoeffisienten til X og Y er tilsvarende gitt ved

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{sd}(X)\,\mathrm{sd}(Y)}.$$

Den har følgende egenskaper:

- $-1 \leq \rho \leq 1$
- $|\rho| = 1$ hvis og bare hvis Y er lineær funksjon av X

Hvis $\mathrm{Cov}(X, Y) = 0$ (og dermed $\rho(X, Y) = 0$) da er X og Y **ukorrelert**.

Hvis X og Y er uavhengig, da er de også ukorrelert, men ikke nødvendigvis omvendt.

Oppsummering

- Fordelinger for stokastiske variabler beskrives av lokasjons- og spredningsmål som i deskriptiv statistikk.
- Felles oppførsel av to variabler beskrives med simultan-, marginal- og betingede fordelinger.
- Kovarians og korrelasjon måler styrken til det lineære forholdet mellom to stokastiske variabler.