STK-IN1050 - Statistikk for informatikere

6. Stokastiske variabler

Matematisk institutt

2026-02-04

Stokastiske variabler for å beskrive stokastiske prosesser

- Hvordan kan stokastiske variabler defineres som et teoretisk konstrukt?
- Hvordan karakteriseres fordelingen til diskrete og kontinuerlige stokastiske variabler?
- Hva er noen eksempler på diskrete og kontinuerlige fordelinger?
- Hvordan overføres begrepet uavhengighet til stokastiske variabler?

Matematiske variabler vs. stokastiske variabler

I sannsynlighetsteori utgjør stokastiske variabler teoretiske modeller som beskriver stokastiske prosesser, hvor en **stokastisk prosess** er en prosess som styres etter sannsynlighetslover.

En stokastisk variabel er, i likhet med en variabel innen matematikk, noe som kan variere: variabelen kan altså ha ulike verdier.

I motsetning til matematiske variabler vil en stokastisk variabel i tillegg ha en sannsynlighetsfordeling som forteller hvor sannsynlige de ulike verdiene er.

Karakteristikker vs. stokastiske variabler

Deskriptiv statistikk	Sannsynlighetsteori
karakteristikker	stokastiske variabler
verdier	verdier (fra en verdimengde)
absolutte og relative frekvenser	sannsynligheter for (mengder av) muliger verdier

- Merk: mulige verdier av en stokastisk variabel må være tall.
- Skille: diskrete vs. kontinuerlige stokastiske variabler.

Definisjon: stokastisk variabel

Stokastisk variabel

En **stokastisk variabel** (eller **tilfeldig variabel**) X er en funksjon som tilordner hvert mulig utfall $\omega \in \Omega$ en reell verdi fra en **verdimengde** $\mathcal{X} \subseteq \mathbb{R}$.

Dvs. en stokastisk variabel oversetter utfallet i et stokastisk eksperiment til et tall.

Eksempel: To terningkast

I to terningkast er utfallsrommet $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$.

Vi kan definiere forskjellige stokastiske variabler for å beskrive de to treningskastene. For eksempel,

- X = summen av de to kastene
  - verdimengde $\mathcal{X} = \{2, 3, 4, \ldots, 12\}$
- Y = den absolute differensen av de to kastene
  - verdimengde $\mathcal{Y} = \{0, 1, 2, \ldots, 5\}$
- $Z = 1$ hvis summen er et partall og $Z = 0$ hvis summen er et oddetall
  - verdimengde $\mathcal{Z} = \{0, 1\}$

Diskrete stokastiske variabler

Diskret stokastisk variabel

Den stokastiske variablen X er **diskret** hvis den antar et endelig eller tellbart sett av verdier $x_1, x_2, \ldots$, dvs. hvis $\mathcal{X} = \{x_1, x_2, \ldots\}$.

Sannsynlighetsfordelingen (eller **fordelingen**) til X er beskrevet ved **sannsynlighetsfunksjonen**

$$P(X = x_i) = p_i, \text{ for } i = 1, 2, \ldots.$$

Diskrete stokastiske variabler

Det gjelder at

- alle p_i er sannsynligheter: $0 \leq p_i \leq 1$;
- summen av sannsynlighetene er 1: $\sum_i p_i = 1$;
- vi får sannsynligheten for en begivenhet ved å summere:

$$P(X \in A) = \sum_{i:x_i \in A} p_i \quad \text{for } A \subseteq \mathcal{X};$$

og spesielt gjelder at

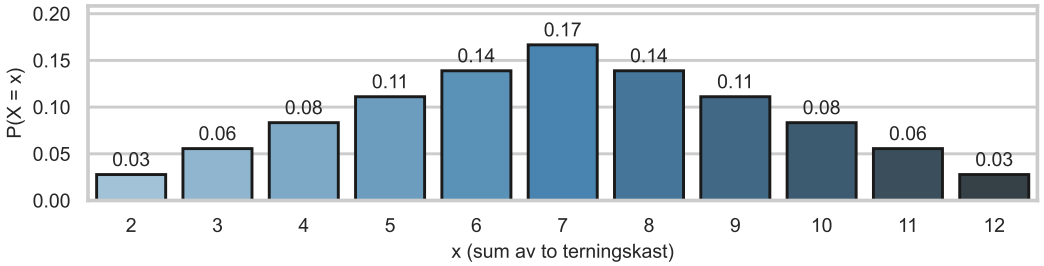
$$P(a \leq X \leq b) = \sum_{i:a \leq x_i \leq b} p_i.$$

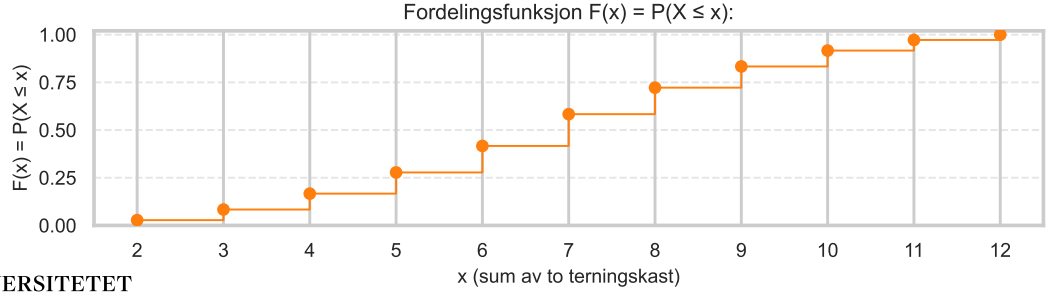
Eksempel: Sum av to terningskast

La X være summen av to terningskast. Da er $\mathcal{X} = \{2, 3, \ldots, 12\}$ og de mulige verdiene har sannsynligheter

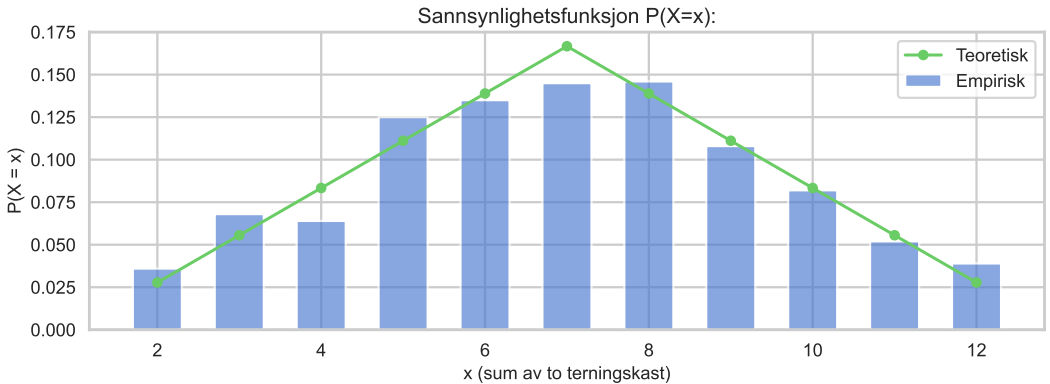
- $P(X = 2) = 1/36$
- $P(X = 3) = 2/36$
- $P(X = 4) = 3/36$
- $P(X = 5) = 4/36$
- $P(X = 6) = 5/36$
- $P(X = 7) = 6/36$
- $P(X = 8) = 5/36$
- $P(X = 9) = 4/36$
- $P(X = 10) = 3/36$
- $P(X = 11) = 2/36$
- $P(X = 12) = 1/36$

hvor summen av sannsynlighetene er lik 1.

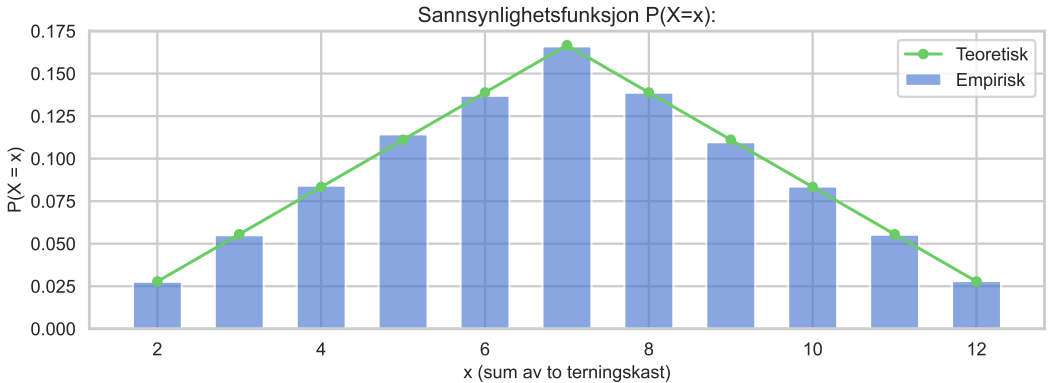

Sannsynlighetsfunksjon P(X=x):
P(X = x)
0.20
0.15
0.10
0.05
0.00
0.03
0.06
0.08
0.11
0.14
0.17
0.14
0.11
0.08
0.06
0.03
2
3
4
5
6
7
8
9
10
11
12
x (sum av to terningskast)


Fordelingsfunksjon F(x) = P(X ≤ x):
F(x) = P(X ≤ x)
1.00
0.75
0.50
0.25
0.00
2
3
4
5
6
7
8
9
10
11
12
x (sum av to terningskast)

Vi kan også sammenligne den teoretiske sannsynlighetsfunksjonen mot empiriske sannsynligheter (relative frekvenser) basert på 1000 simuleringer:



Hvis vi i stedet kjører 100 000 simuleringer, så nærmer de empiriske sannsynlighetene seg de tilsvarende teoretiske verdiene:



Kontinuerlige stokastiske variabler

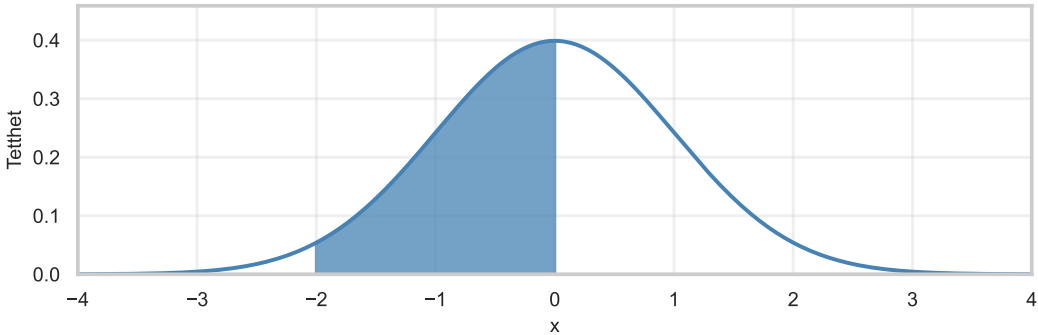
- Kontinuerlige tilfeldige variabler kan anta alle reelle verdier i et intervall.
- Dette gjør det umulig å angi sannsynligheter for individuelle verdier på en meningsfull måte (for mange mulige verdier).
- I stedet beskrives fordelingen av en tetthetsfunksjon $f(x) \geq 0$.

Kontinuerlig stokastisk variabel

Den stokastiske variablen X er **kontinuerlig** hvis det eksisterer en **tetthetsfunksjon** $f(x) \geq 0$ slik at for ethvert intervall $[a, b]$ gjelder at

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

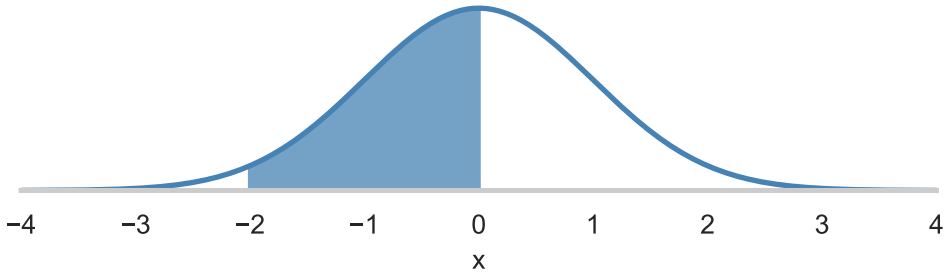
Illustrasjon av tetthet



Eksempel på en tetthetsfunksjon f hvor $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ og

$$P(-2 \leq X \leq 0) = \int_{-2}^{0} f(x)\,dx \approx 0.48.$$

Egenskaper

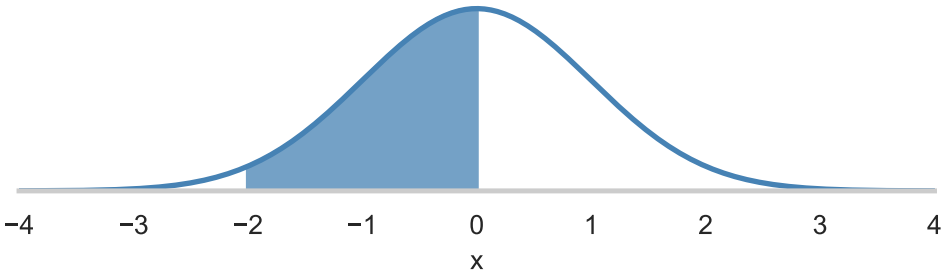


For kontinuerlige stokastiske variabler gjelder at

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

og $P(X = x) = 0$ for alle $x \in \mathbb{R}$.

Egenskaper



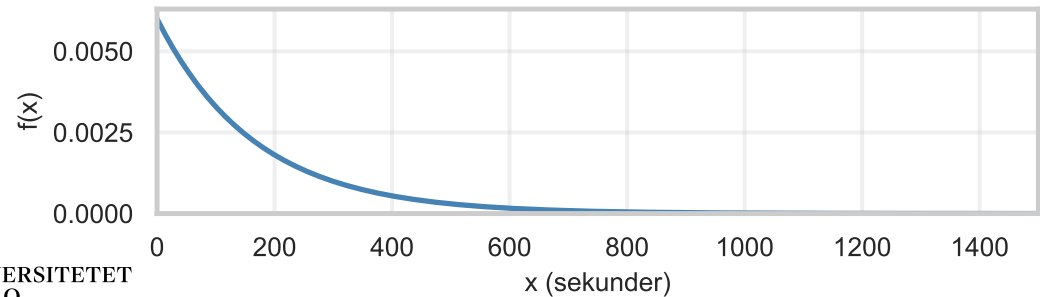
- Normaliseringsegenskapen $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ samsvarer $\sum_i p_i = 1$ for diskrete stokastiske variabler.
- Det er mulig å ha $f(x) > 1$.
- Verdimengden til X er gitt ved $\mathcal{X} = \{x : f(x) > 0\}$.

Eksempel: Callsenter

Anta at X er varigheten på en samtale med et callsenter (i sekunder).

Vi antar at vi kan beskrive sannsynlighetsfordelingen til X ved følgende tetthet:

$$f(x) = \begin{cases} 0.006\, e^{-0.006x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$



Fordelingsfunksjon

Fordelingsfunksjonen til en stokastisk variabel X er gitt ved

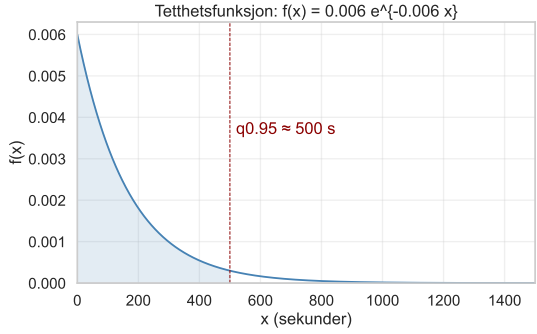
$$F(x) = P(X \leq x).$$

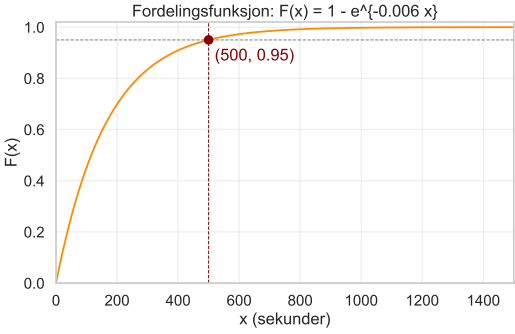
For et intervall $(a, b]$ gjelder at

$$P(a < X \leq b) = F(b) - F(a).$$

Når F er kjent, trenger vi altså ofte ikke å beregne integraler direkte.

Eksempel: Callsenter





Vi har at $P(X \leq 500) = \int_0^{500} f(x)dx = F(500) \approx 0.95$.

Kvantiler

Anta at X er en stokastisk variabel. For $0 \leq \alpha \leq 1$ er q_α et α-kvantil hvis:

$$P(X \leq q_\alpha) \geq \alpha$$

og

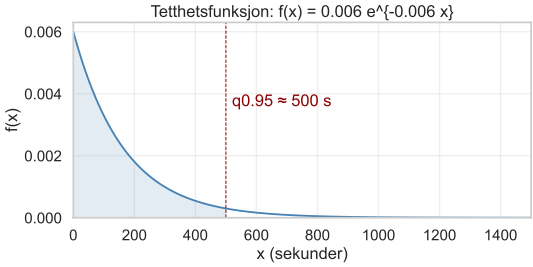
$$P(X \geq q_\alpha) \geq 1 - \alpha.$$

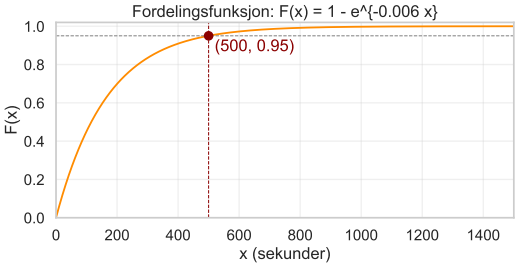
Kvantilen er entydig bestemt for kontinuerlige stokastiske variabler og gitt ved løsningen av ligningen

$$F(q_\alpha) = \alpha.$$

Generelt: kvantilen er entydig bestemt hvis F er en strengt voksende funksjon.

Eksempel: Callsenter





Vi finner $q_{0.95}$ ved å løse $F(q_{0.95}) = 1 - e^{-0.006 q_{0.95}} = 0.95$. Det gir:

$$q_{0.95} = -\frac{1}{0.006} \ln(0.05) \approx 500.$$

Tolkning: ca. 95% av samtalene varer mindre enn 500 s.

Uavhengighet av stokastiske variabler

To stokastiske variabler X og Y er **stokastisk uavhengige** hvis

$$P(X \leq x,\ Y \leq y) = P(X \leq x)\, P(Y \leq y) \quad \text{for alle } x, y \in \mathbb{R}.$$

Mer generelt sier vi at n stokastiske variabler $X_1, \ldots, X_n$ er stokastisk uavhengige hvis

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n) \quad \text{for vilkårlige } x_1, \ldots, x_n \in \mathbb{R}.$$

Uavhengighet av stokastiske variabler

Hvis de stokastiske variablene er diskrete, da kan vi også kreve at

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

eller

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

for vilkårlige verdier fra den tilsvarenden verdimengden.

For kontinuerlige stokastiske variabler har vi lignende kriterier basert på tetthetsfunksjonene:

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

eller

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

Uavhengige og identisk fordelte (iid) stokastiske variabler

Stokastiske variabler $X_1, \ldots, X_n$ er **uavhengige og identisk fordelt** (iid) hvis:

- de er stokastisk uavhengige, og
- hver X_i har samme fordeling.

Denne situasjonen oppstår, for eksempel, ved

- n uavhengige gjentakelser av et tilfeldig eksperiment
- trekning med tilbakelegging n ganger.

I statistikk brukes uavhengige og identisk fordelte stokastiske variabler ofte i statistisk modellering.

Oppsummering: Stokastiske variabler

- Stokastiske variabler beskriver resultatet av et stokastisk eksperiment gjennom tall.
- Diskrete og kontinuerlige fordelinger karakteriseres forskjellig (sannsynlighetsfunksjon vs. tetthetsfunksjon).
- Fordelingsfunksjonen F er et enhetlig begrep for begge typer.
- Uavhengige og identisk fordelte (iid) variabler er et grunnleggende modellgrep i statistikk.