STK-IN1050 - Statistikk for informatikere

8. Diskrete fordelinger

Matematisk institutt

2026-02-16

Introduksjon

Vi skal se på tre grunnleggende diskrete sannsynlighetsfordelinger til å beskrive fordelingen til diskrete stokastiske variabler:

- Bernoulli
- binomisk
- Poisson

Sentrale begreper for å beskrive sannsynlighetsfordelinger:

- sannsynlighetsfunksjon: $P(X = x)$
- fordelingsfunksjon: $P(X \leq x) = F(x)$
- forventning: $E(X)$
- varians: $\text{Var}(X)$

Bernoulli-fordelingen

Bernoulli-fordelingen brukes til å beskrive **binære** data (1/0, Ja/Nei, Suksess/Feil).

Da stokastiske variabler tar tall-verdier, så koder vi utfallet som 1/0, hvor 1 vanligvis betyr ja/suksess. Dvs. X har verdimengde $\mathcal{X} = \{0, 1\}$.

Vi sier at X er **Bernoulli fordelt** og skriver $X \sim \text{Bernoulli}(p)$ som betyr at

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

Det gir

$$\begin{aligned} \mathbb{E}(X) &= 1 \cdot p + 0 \cdot (1 - p) = p, \\ \text{Var}(X) &= (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p). \end{aligned}$$

Bernoulli-fordelingen i Python

```
from scipy.stats import bernoulli

p = 0.3
rv = bernoulli(p)
type(rv)
```

```
scipy.stats._distn_infrastructure.rv_discrete_frozen
```

```python
import numpy as np

x = np.array([0, 1])
pmf = rv.pmf(x)
cdf = rv.cdf(x)
mean = rv.mean()
var = rv.var()
sample = rv.rvs(size=1000,
                random_state=42)
```

```
x: [0 1]
pmf: [0.7 0.3]
cdf: [0.7 1. ]
mean: 0.3
var: 0.21
sample mean: 0.288
sample var: 0.205
```

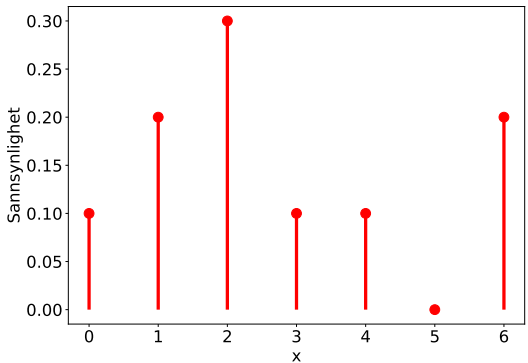
Her er

- pmf = probability mass function = sannsynlighetsfunksjonen
- cdf = cumulative distribution function = fordelingsfunksjonen
- rvs = random value sample

Eksempel: Skreddersydd diskret fordeling med 7 kategorier

```
import numpy as np
from scipy import stats
xk = np.arange(7)
pk = (0.1, 0.2, 0.3, 0.1, 0.1, 0.0, 0.2)
custm = stats.rv_discrete(name='custm',
          values=(xk, pk))
```

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
ax.plot(xk, custm.pmf(xk), 'ro',
        ms=12, mec='r')
ax.vlines(xk, 0, custm.pmf(xk),
          colors='r', lw=4)
ax.set_xlabel('x', fontsize=20)
ax.set_ylabel('Sannsynlighet',
              fontsize=20)
ax.tick_params(axis='both',
               labelsize=20)
plt.show()
```



Skreddersydd diskret fordeling: Forventning og varians

```
import numpy as np

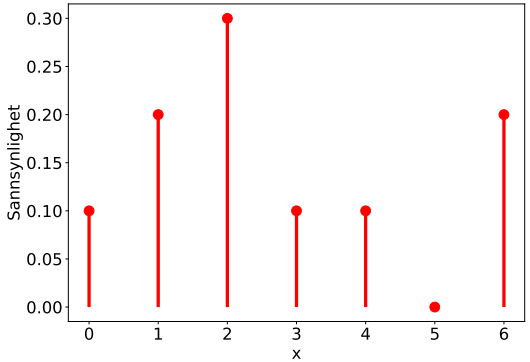
xk = np.arange(7)
pk = (0.1, 0.2, 0.3, 0.1, 0.1, 0.0, 0.2)
# Expected value
EX = np.sum(xk * pk)
# Second moment
EX2 = np.sum(xk**2 * pk)
# Variance
VarX = EX2 - EX**2

print("E[X] =", EX)
print("E[X^2] =", EX2)
print(f"Var(X) = {VarX:.3g}")
```

```
E[X] = 2.7
E[X^2] = 11.1
Var(X) = 3.81
```

```
custm = stats.rv_discrete(name='custm',
          values=(xk, pk))
print("E[X] =", custm.mean())
print(f"Var(X) = {custm.var():.3g}")
```

```
E[X] = 2.7
Var(X) = 3.81
```



Skreddersydd diskret fordeling: Simulering

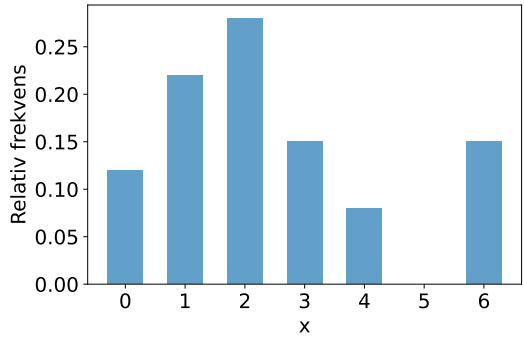
```
import numpy as np
from scipy import stats

xk = np.arange(7)
pk = (0.1, 0.2, 0.3, 0.1, 0.1, 0.0, 0.2)
custm = stats.rv_discrete(name='custm',
          values=(xk, pk))

np.random.seed(0)
R = custm.rvs(size=100)
values, counts = np.unique(R, return_counts=True)
rel_freq = counts / len(R)

print("Verdier =", values)
print("Relativ frekvens =", rel_freq)
```

```
Verdier = [0 1 2 3 4 6]
Relativ frekvens = [0.12 0.22 0.28 0.15 0.08 0.15]
```



Binomisk fordeling: et eksempel

I en søskenflokk er det fire barn som ikke er tvillinger, trillinger eller firlinger

Hva er sannsynligheten for at det er to gutter og to jenter i søskenflokken?

La X = «antall gutter i søskenflokken».

Vi vil finne $P(X = 2)$.

To eldste gutter, to yngste jenter: *GGJJ*

Eldste og yngste gutt, to midterste jenter: *GJJG*

Andre rekkefølger som gir to gutter og to jenter: *GJGJ*, *JGGJ*, *JGJG* og *JJGG*

Binomisk fordeling: et eksempel

Vi antar at barnas kjønn er uavhengig av hverandre

$$P(GGJJ) = 0.514 \cdot 0.514 \cdot 0.486 \cdot 0.486 = 0.514^2 \cdot 0.486^2$$
$$P(GJJG) = 0.514 \cdot 0.486 \cdot 0.486 \cdot 0.514 = 0.514^2 \cdot 0.486^2$$

Hver av de fire andre rekkefølgene som gir to gutter og to jenter har også sannsynligheten $0.514^2 \cdot 0.486^2$.

Vi finner $P(X = 2)$ ved å legge sammen sannsynlig-hetene for de seks rekkefølgene som gir to gutter og to jenter:

$$P(X = 2) = 6 \cdot 0.514^2 \cdot 0.486^2 = 0.374.$$

Binomisk fordeling: et eksempel

Vi så at det var seks rekkefølger som gir to gutter og to jenter.

Det kunne vi funnet ut uten å skrive opp alle rekkefølgene.

For å velge en bestemt rekkefølge er det samme som å velge 2 plasser blant 4 der det skal stå *G*.

Det kan vi gjøre på

$$\binom{4}{2} = 6$$

måter.

Binomisk forsøk

- Vi gjør n forsøk

  (I eksemplet er hvert barn et «forsøk»)

- I hvert forsøk er det to muligheter: Enten inntreffer S ellers så inntreffer F

  (I eksemplet er hvert barn enten en gutt eller en jente)

- Forsøkene er uavhengige

  (I eksemplet har vi antatt uavhengighet siden vi ser bort fra tvillinger, osv.)

- I hvert forsøk er sannsynligheten lik p for at S skal inntreffe og $1 - p$ for at F skal inntreffe

  (I eksemplet er sannsynligheten for gutt 51.4% og sannsynligheten for jente 48.6%)

Binomisk fordeling

La X være antall ganger S inntreffer i et binomisk forsøk.

Ved å resonnere som i eksemplet finner vi at X har sannsynlighetsfunksjon (eller **punktsannsynlighet**)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

for $k = 0, 1, \ldots, n$.

Vi sier at X er **binomisk fordelt**.

Vi skriver $X \sim Bin(n, p)$.

Binomisk fordeling med Python

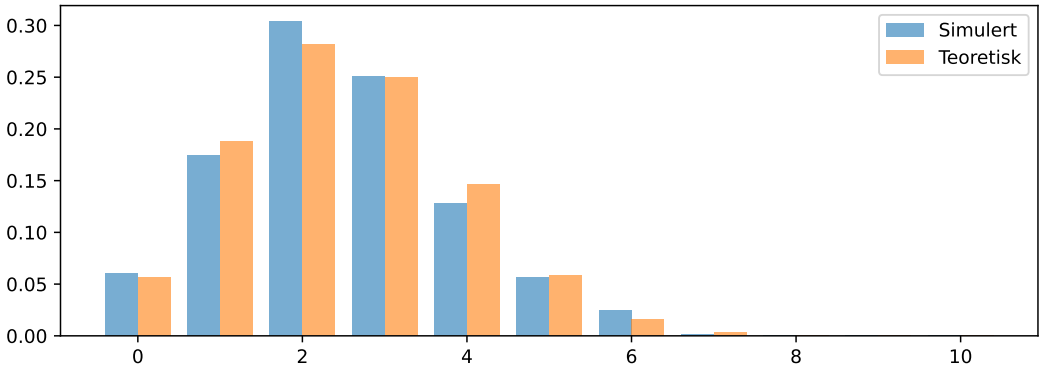
```
import numpy as np
from scipy.stats import binom

n = 10
p = 0.25
rv = binom(n, p)
k = np.arange(0, n+1)
pmf = rv.pmf(k)
cdf = rv.cdf(k)
mean = rv.mean()
var = rv.var()
```

```
k: [ 0  1  2  3  4  5  6  7  8  9 10]
pmf: [0.056 0.188 0.282 0.250 0.146 0.058 0.016 0.003 0.000 0.000]
cdf: [0.056 0.244 0.526 0.776 0.922 0.980 0.996 1.000 1.000 1.000]
mean: 2.5 var: 1.875
```

Binomisk fordeling med Python

```
samples = rv.rvs(size=1000, random_state=123)
counts = np.bincount(samples, minlength=n+1) / samples.size
```



Eksempel: Frø

En type frø spirer med 70% sannsynlighet.

Vi sår 20 frø.

- Hva er sannsynligheten for at 15 frø vil spire?
- Hva er sannsynligheten for at minst 15 frø vil spire?

La X være antall frø som spirer.

Hvis frøene spirer uavhengig av hverandre, da er $X \sim Bin(n = 20, p = 0.7)$:

$$P(X = 15) = \binom{20}{15} 0.70^{15} 0.30^{5} = 0.179$$

$$P(X \geq 15) = P(X > 14) = 1 - P(X \leq 14) = 0.416$$

Eksempel: Frø

```
import numpy as np
from scipy.stats import binom

n = 20
p = 0.7
rv = binom(n, p)

P_15 = rv.pmf(15)
P1 = 1 - rv.cdf(14)
P2 = rv.sf(14)
```

```
P(X = 15): 0.179
1 - P(X <= 14):  0.416
P(X > 14):  0.416
```

Eksempel: Meningsmåling

Anta at et politisk parti har oppslutning fra 19% av velgerne.

Et meningsmålingsinstitutt spør et tilfeldig utvalg på 1000 personer hva de ville ha stemt hvis det var valg.

Hva er sannsynligheten for at partiets oppslutning på meningsmålingen blir mellom 17 og 21 prosent?

- Hva er den stokastiske variablen X?
- Hvilken fordeling har X?
- Hvilken sannsynlighet skal vi finne?

Eksempel: Meningsmåling

Her er X = «antall som ville ha stemt på partiet» med $X \sim Bin(1000, 0.19)$.

Partiets oppslutning på meningsmålingen blir mellom 17% og 21% hvis $170 \leq X \leq 210$. Dvs. vi skal finne

$$P(170 \leq X \leq 210) = P(X \leq 210) - P(X \leq 169)$$

```
from scipy.stats import binom

print(f"{stats.binom.cdf(210, 1000, 0.19)
          - stats.binom.cdf(169, 1000, 0.19): .3f}")
```

```
 0.902
```

Eksempel: Firebarnsmødre født 1935-1991 (data fra SSB)

Passer binomisk fordeling med $n = 4$ og $p = 0.514$?

Mødre som har født minst 4 barn	Antall kvinner Absolutt	Prosent
I alt	95 993	100.0
4 jenter	6 048	6.3
3 jenter og 1 gutt	22 484	23.4
2 jenter og 2 gutter	34 095	35.5
1 jente og 3 gutter	25 446	26.5
4 gutter	7 920	8.3

```
import numpy as np
import scipy.stats as stats

x = np.arange(0,5)
rv = binom(4, 0.514)
p = rv.pmf(x)
print(p * 100)
```

```
[5.6 23.6 37.4 26.4 7.0]
```

Binomisk fordeling gir litt for få familier der barna har samme kjønn.

Forventning og varians i binomisk fordeling

Anta $X \sim Bin(n, p)$.

La S_i være en binær tilfeldig variabel som indikerer om observasjon i er en suksess ($S_i = 1$) eller ikke ($S_i = 0$).

Da er $S_1, \ldots, S_n$ uavhengig og identisk fordelt med $S_i \sim Bernoulli(p)$, $\mathbb{E}S_i = p$ og $\text{Var}(S_i) = p(1-p)$.

Da er også $X = S_1 + S_2 + \cdots + S_n$ (antall suksesser).

Forventningen til X er

$$\mathbb{E}X = \mathbb{E}(S_1 + \cdots + S_n) = \mathbb{E}S_1 + \cdots + \mathbb{E}S_n = np$$

og variansen blir

$$\text{Var}(X) = \text{Var}(S_1 + \cdots + S_n) = \text{Var}(S_1) + \cdots + \text{Var}(S_n) = np(1-p).$$

Når utfallene ikke lenger er uavhengig

Anta at vi har en populasjon av størrelse N med andel suksess p.

Vi trekker et utvalg av størrelse n, hvor X er antall suksesser i utvalget.

I noen tilfeller kan utfallet av 2. observasjon avhenge av utfallet av 1. observasjon.

Eksempel: $N = 52$ kort, trekker $n = 2$ kort. Da er

$$P(\text{1. Rødt}) = 26/52 = 0.5, \quad P(\text{2. Rødt} \mid \text{1. Rødt}) = 25/51 < 0.5$$

Vi har avhengighet mellom observasjonene.

MEN: Hvis N er mye større enn n ($N > 20n$), kan man neglisjere slik avhengighet, og X er tilnærma $Bin(n, p)$-fordelt.

Presisjonen til denne tilnærminga er bedre jo større forholdet N/n er.

Poisson-fordelingen

Poisson viste at hvis $n \to \infty$ og $p \to 0$ på en slik måte at $np \to \lambda > 0$ så vil

$$\binom{n}{k} p^k (1-p)^{n-k} \to \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson brukte dette til å finne en tilnærming til den binomiske fordelingen når n er stor og p er liten.



Poisson-fordelingen er oppkalt etter den franske fysikeren og matematikeren Siméon Denis Poisson (1781-1840)

Poisson-fordeling

Seinere fant en ut at en kunne bruke den grensen Poisson fant som sannsynlighetsfunksjon til en stokastisk variabel.

Vi sier at en stokastisk variabel X er **Poisson-fordelt** med parameter λ, eller $X \sim Po(\lambda)$, hvis den har sannsynlighetsfunksjon

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \ldots$$

Merk at $\sum_{k=0}^{\infty} P(X = k) = 1$ slik det skal være for en sannsynlighetsfunksjon.

Videre gjelder at

$$\mathbb{E}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

Betingelser for Poisson-fordeling

Poisson-fordeling brukes ofte når vi har tellevariable som kan gå i det uendelige, i hvert fall i teoretisk forstand.

Den krever noen betingeleser:

- Antallet suksesser som skjer i to ikke-overlappende måleenheter (f.eks. to tidsperioder eller geografisk områder) er uavhengige.
- Sannsynligheten for at en suksess inntreffer inne i en måleenhet er den samme for alle enheter av samme størrelse. Sannsynligheten er da også proporsjonal med størrelsen til enheten.
- Sannsynligheten for at mer enn én hendelse skjer innenfor en enhet er neglisjerbar for veldig små enheter. Med andre ord, hendelsene skjer én av gangen.

Poisson-fordeling med Python

```
import numpy as np
from scipy.stats import poisson

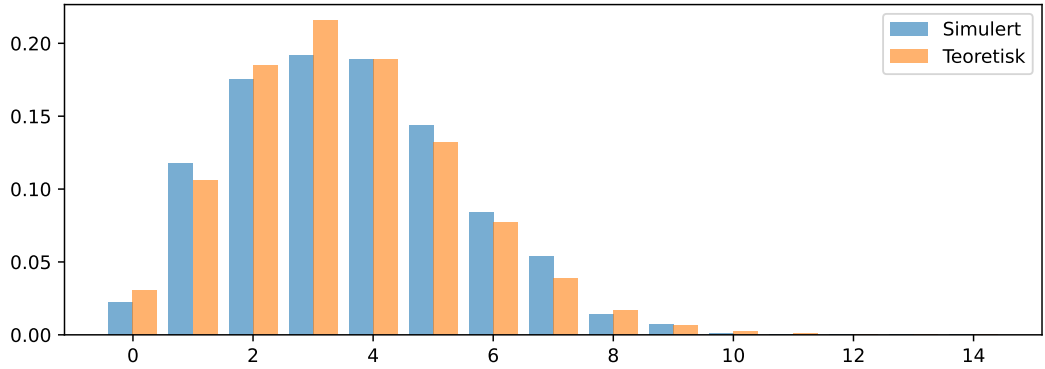
lam = 3.5
rv = poisson(lam)

k = np.arange(0, 15)
pmf = rv.pmf(k)
cdf = rv.cdf(k)
mean = rv.mean()
var = rv.var()
```

```
pmf (k=0..10): [0.03 0.11 0.18 0.22 0.19 0.13 0.08 0.04 0.02 0.01 0.00]
cdf (k=0..10): [0.03 0.14 0.32 0.54 0.73 0.86 0.93 0.97 0.99 1.00 1.00]
mean: 3.5 var: 3.5
```

Poisson-fordeling med Python

```
samples = rv.rvs(size=1000, random_state=2026)
counts = np.bincount(samples, minlength=k.max()+1)[:len(k)] / samples.size
```



Sammenligning binomisk og Poisson

Vi sammenligner

- $Bin(n = 20, p = 0.1)$
- $Bin(n = 200, p = 0.01)$
- $Po(\lambda = 2)$

```
x=np.arange(0,10)
b20= stats.binom.pmf(x,20,0.1)
b200=stats.binom.pmf(x,200,0.01)
pois= stats.poisson.pmf(x,2)
```

```
 b20: [0.122 0.270 0.285 0.190 0.090 0.032 0.009 0.002 0.000 0.000]
b200: [0.134 0.271 0.272 0.181 0.090 0.036 0.012 0.003 0.001 0.000]
pois: [0.135 0.271 0.271 0.180 0.090 0.036 0.012 0.003 0.001 0.000]
```

Eksempel: Antall drepte av hestespark i det preusiske kavaleriet (1875-1894)

Eksempelet ble publisert av Ladislaus von Bortkiewicz i 1898 og var et av de første eksemplene på bruk av Poisson-fordelingen til å beskrive et datamateriale

Det er registrert antall soldater som ble drept av hestespark i 10 kavalerikorps i en periode på 20 år (totalt 200 «korps-år»)

Antall drepte	0	1	2	3	4	5+
Antall korps-år	109	65	22	3	1	0

Totalt antall drepte

$$0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot = 122$$

Kan dataene beskrives med Poisson-fordelingen?

Eksempel: Antall drepte av hestespark i det preusiske kavaleriet (1875-1894)

La X være antall drepte i et korps-år. Hvis X er Poisson-fordelt har vi at

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

For λ bruker vi gjennomsnittlig antall drepte per korps-år som er $\frac{122}{200} = 0.61$.

Forventet antall korps-år med x drepte er $n \cdot P(X = x)$, hvor $n = 200$.

Antall drepte	0	1	2	3	4	5+
Antall korps-år	109	65	22	3	1	0
$P(X = x)$	0.54	0.33	0.10	0.02	0.003	0.007
Forventet antall	109	66	20	4	1	0

Poisson-fordelingen passer bra.

Eksempel: Fødsler ved Ahus 1999-2014

I en artikkel i Legeforeningens tidsskrift studeres fordelingen av antall ikke-planlagte fødsler per dag ved Ahus i perioden 1999-2014 (totalt vel 50 000 fødsler)[1].

De fant at antall fødsler per dag er Poisson-fordelt, der forventningen λ øker over perioden 1999-2015 (siden antall fødsler ved Ahus har økt i perioden) og avhenger i noen grad av tid på året (siden det blir født flere barn om våren og sommeren enn om høsten og vinteren).

For en «gjennomsnittsmåned» i 2014 fant de $\lambda = 10$.

[1] https://tidsskriftet.no/2015/12/originalartikkel/sesongjusterte-fodselsfrekvenser-er-poisson-fordelte

Eksempel: Fødsler ved Ahus 1999-2014

Hvis antall fødsler per dag er Poisson-fordelt med $\lambda = 10$, hvor mange fødeplasser må sykehuset ha?

- Hvis sykehuset har 10 fødeplasser, er sannsynligheten 58.3% for at alle som føder en dag får plass
- Hvis sykehuset har 15 fødeplasser, er sannsynligheten 95.1% for at alle som føder en dag får plass
- Hvis sykehuset har 18 fødeplasser, er sannsynligheten 99.3% for at alle som føder en dag får plass

x	$P(X \leq x)$
8	0.333
9	0.458
10	**0.583**
11	0.697
12	0.793
13	0.865
14	0.917
15	**0.951**
16	0.973
17	0.986
18	**0.993**

Oppsummering

Når velge hvilken fordeling?

- Bernoulli: enkelt binært utfall (1/0).
- binomisk: sum av uavhengige Bernoulli-forsøk med lik p (f.eks. antall suksesser i n forsøk).
- Poisson: ofte modell for antall hendelser i en tids-/områdeenhet når hendelser er sjeldne og uavhengige; kan også være grense for binomisk fordeling når $n \to \infty$, $p \to 0$ men $np \to \lambda$.

Python tips

Numerisk bruk i SciPy:

- `.pmf(k)` for sannsynlighetsfunksjonen i enkelte punkter
- `.cdf(k)` for fordelingsfunksjonen (kumulativ fordelingsfunksjon)
- `.sf(k)` som numerisk mer stabil alternativ til $1 - \mathrm{cdf}$ for store k
- `.rvs(size=...)` for simuleringer
- `.mean()` og `.var()` for forventning og varians gitt parameterverdier

Videre lesning: SciPy-dokumentasjon (scipy.stats)