

02402 Statistik (Polyteknisk grundlag)

Uge 2:

Python til statistik,
Stokastiske variable (fortsat),
Diskrete sandsynlighedsfordelinger

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

1 Python til statistik

2 Konkrete diskrete fordelinger

- Binomialfordelingen
- Den Hypergeometriske fordeling
- Poissonfordelingen
- Diskrete fordelinger i Python

3 Kontinuære stokastiske variable

02402 Statistik (Polyteknisk grundlag)

Python til statistik

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Python

I 02402 bruger vi VS Code og jupyter notebooks.

få hjælp: <https://pythonsupport.dtu.dk/> !!

I denne uge:

- 1 Deskriptiv statistik + data visualisering
- 2 Simulering af *tilfældige tal* (stokastiske variable)
- 3 Kendte statistiske fordelinger

Python til deskriptiv statistik

- Indlæse data fra en stikprøve
 - Beregne diverse statistiske nøgletal
 - Data visualisering (plots)
-
- Gå til Python notebook "week2.ipynb" i VS Code
Afsnit "Descriptive statistics in Python"



Visual Studio Code

Python forkortelser i 02402/02323

Vi bruger i dette kursus en række Python "libraries" og anvender følgende forkortelser:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats.power as smp
```

Når I bruger Python til øvelserne, anbefaler vi at I kopierer disse import statements ind i toppen af jeres notebook.

Simulering af stokastiske variable

Python til simulering

Vi kan bruge Python til at *simulere* observationer af en stokastisk variabel. Kaldes også at "*trække tilfældige tal*" (fra en population).

For at simulere observationer af en stokastisk variabel skal vi kende udfaldsrummet og $f(x)$ (eller $F(x)$).

For eksempel kan vi *simulere* kast med en terning (i stedet for at udføre eksperimentet i virkeligheden).

Simulering er et stærkt værktøj til at udføre "tanke eksperimenter" og undersøge hvordan en stokastisk variabel opfører sig *under en antagelse om* dens fordeling.

Python til simulering

Numpy's `random.choice` function kan bruges til at simulere en **diskret** stokastisk variabel, hvor vi selv definerer udfaldsrummet og tæthedsfunktionen.

```
np.random.choice(a, size=None, replace=True, p=None)
```

`a`: en vektor med udfaldsrummet, eller *populationen*.

`size`: antal observationer vi ønsker at simulere/"trække" (evt med struktur).

`replace=True`: angiver om observationer trækkes med eller uden tilbagelægning.

`p`: en vektor med sandsynlighed for hvert udfald.

- Find også dokumentation online for `numpy.random.choice` (det prøver vi lige).
- Til simuleringer der trækker tilfældige tal, kan man sætte et "seed", hvis man ønsker præcis samme resultater hver gang:

```
np.random.seed(42)
```

Python til simulering

Lad os simulere den stokastiske variabel X med tæthedsfunktion, $f(x)$:

x	0	1	2	3
$f(x)$	0.1	0.3	0.4	0.2

I Python bruger vi Numpy's `random.choice` function:

```
np.random.choice(a, size=None, replace=True, p=None)
```

- Gå til Python notebook "week2.ipynb" i VS Code
Afsnit "First simulation in Python"



Visual Studio Code

Eksempler med simulering

Prøv at lege med Example 2.15 og Example 2.19 i bogen

02402 Statistik (Polyteknisk grundlag)

Konkrete diskrete fordelinger

- Binomialfordelingen
- Den Hypergeometriske fordeling
- Poissonfordelingen
- Diskrete fordelinger i Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Diskrete stokastiske variable

(Sandsynligheds-)fordelingen for en diskret stokastisk variabel X kan beskrives med:

- Tæthedsfunktionen $f(x) = P(X = x)$ (pdf)
- Fordelingsfunktionen $F(x) = P(X \leq x)$ (cdf)

I dag gennemgås tre konkrete diskrete fordelinger:

- Binomialfordelingen
- Den hypergeometriske fordeling
- Poissonfordelingen

02402 Statistik (Polyteknisk grundlag)

- Binomialfordelingen
- Den Hypergeometriske fordeling
- Poissonfordelingen
- Diskrete fordelinger i Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Binomialfordelingen, Setup

- Vi betragter et eksperiment med to udfald: "succes" og "fiasko"
- p angiver sandsynligheden for at få en "succes".
- $(1 - p)$ angiver sandsynligheden for at få en "fiasko".
- Eksperimentet gentages n gange (uafhængige gentagelser).
- Lad X være antallet af succes'er.

Binomialfordelingen, Setup

- Vi betragter et eksperiment med to udfald: "succes" og "fiasko"
- p angiver sandsynligheden for at få en "succes".
- $(1 - p)$ angiver sandsynligheden for at få en "fiasko".
- Eksperimentet gentages n gange (uafhængige gentagelser).
- Lad X være antallet af succes'er.

Den stokastiske variabel X følger da en binomialfordeling:

$$X \sim B(n, p)$$

n : antal gentagelser

p : sandsynligheden for succes i hver gentagelse

Binomialfordelingens tæthedsfunktion

||| Definition 2.20 Binomial distribution

Let the random variable X be binomial distributed

$$X \sim B(n, p), \quad (2-19)$$

where n is number of independent draws and p is the probability of a success in each draw.

The binomial *pdf* describes probability of obtaining x successes

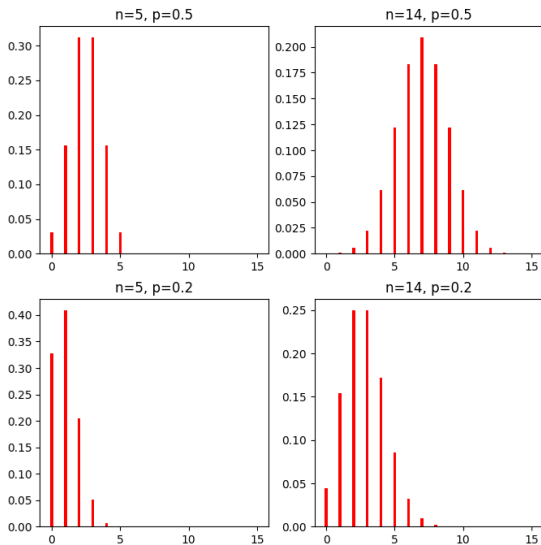
$$f(x; n, p) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad (2-20)$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \quad (2-21)$$

is the number of distinct sets of x elements which can be chosen from a set of n elements. Remember that $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$.

Eksempler på binomialfordelinger



Binomialfordelingens middelværdi og varians

||| Theorem 2.21 Mean and variance

The mean of a binomial distributed random variable is

$$\mu = np, \quad (2-22)$$

and the variance is

$$\sigma^2 = np(1 - p). \quad (2-23)$$

Eksempel med binomialfordelingen

I et firma prøver man at forbedre kundetilfredsheden. Det er især vigtigt at indrapporterede fejl bliver udbedret inden der er gået 24 timer.

Sandsynligheden for, at fejl udbedres indenfor 24 timer, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres indenfor 24 timer?

Hvad skal repræsenteres af den stokastiske variabel X ?

Hvad er fordelingen af X (og hvad er parametrene)?

Hvilken sandsynlighed skal udregnes?

02402 Statistik (Polyteknisk grundlag)

- Binomialfordelingen
- Den Hypergeometriske fordeling
- Poissonfordelingen
- Diskrete fordelinger i Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Den Hypergeometriske fordeling

- Vi betragter et eksperiment med to udfald (igen "succes" og "fiasko"), men denne gang fra en lukket population af i alt N mulige udfald.
- Det typiske billede er *trækning uden tilbagelægning*:
Fra en population med i alt N enheder, findes a "succes'er" og $N - a$ "fiasko'er".
- Man trækker n enheder - *uden tilbagelægning*.
Dvs. *sandsynligheden for at få en succes ændrer sig efter hver trækning*.
- Lad X være antallet af succes'er ud af de n trækninger.

Den Hypergeometriske fordeling

- Vi betragter et eksperiment med to udfald (igen "succes" og "fiasko"), men denne gang fra en lukket population af i alt N mulige udfald.
- Det typiske billede er *trækning uden tilbagelægning*:
Fra en population med i alt N enheder, findes a "succes'er" og $N - a$ "fiasko'er".
- Man trækker n enheder - *uden tilbagelægning*.
Dvs. *sandsynligheden for at få en succes ændrer sig efter hver trækning*.
- Lad X være antallet af succes'er ud af de n trækninger.

Den stokastiske variabel X følger da en Hypergeometrisk fordeling:

$$X \sim H(n, a, N)$$

n er antallet af trækninger (gentagelser)

a er antallet af succes'er i (hele) populationen

N er antallet af enheder i (hele) populationen

Den Hypergeometriske fordelings tæthedsfunktion

||| Definition 2.24 Hypergeometric distribution

Let the random variable X be the number of successes in n draws without replacement. Then X follows the hypergeometric distribution

$$X \sim H(n, a, N), \quad (2-24)$$

where a is the number of successes in the N elements large population. The probability of obtaining x successes is described by the hypergeometric *pdf*

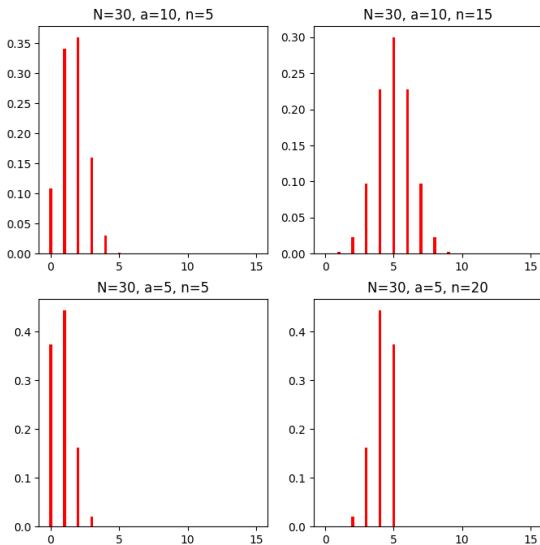
$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}}. \quad (2-25)$$

The notation

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \quad (2-26)$$

represents the number of distinct sets of b elements which can be chosen from a set of a elements.

Eksempler på hypergeometriske fordelinger



Den Hypergeometriske fordelings middelværdi og varians

||| Theorem 2.25 Mean and variance

The mean of a hypergeometric distributed random variable is

$$\mu = n \frac{a}{N}, \quad (2-27)$$

and the variance is

$$\sigma^2 = n \frac{a(N-a)}{N^2} \frac{N-n}{N-1}. \quad (2-28)$$

Eksempel med den Hypergeometriske fordeling

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer, men dette ved modtageren ikke.

Modtageren udtager 3 tilfældige harddiske til test.

Hvad er sandsynligheden for at mindst en af dem har skrammer?

Hvad skal repræsenteres af den stokastiske variabel X ?

Hvad er fordelingen af X (og hvad er parametrene)?

Hvilken sandsynlighed skal udregnes?

02402 Statistik (Polyteknisk grundlag)

- Binomialfordelingen
- Den Hypergeometriske fordeling
- **Poissonfordelingen**
- Diskrete fordelinger i Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Poisson fordelings tæthedsfunktion

- Vi betragter et eksperiment hvor vi observerer *tilfældige events* (hændelser)
- Events forekommer på tilfældige tidspunkter, men med en *gennemsnitlig rate* λ .
- λ måles typisk i *events per tid*, men generelt som *antal per interval*.
- Lad X være antallet af events der observeres i løbet af et bestemt (tids-)interval.
- Der er ingen naturlig øvre grænse på antallet X .

Poisson fordelings tæthedsfunktion

- Vi betragter et eksperiment hvor vi observerer *tilfældige events* (hændelser)
- Events forekommer på tilfældige tidspunkter, men med en *gennemsnitlig rate* λ .
- λ måles typisk i *events per tid*, men generelt som *antal per interval*.
- Lad X være antallet af events der observeres i løbet af et bestemt (tids-)interval.
- Der er ingen naturlig øvre grænse på antallet X .

Den stokastiske variabel X følger da en Poisson fordeling:

$$X \sim Po(\lambda)$$

λ er det gennemsnitlige antal hændelser per interval

Poisson fordelings tæthedsfunktion

||| Definition 2.27 Poisson distribution

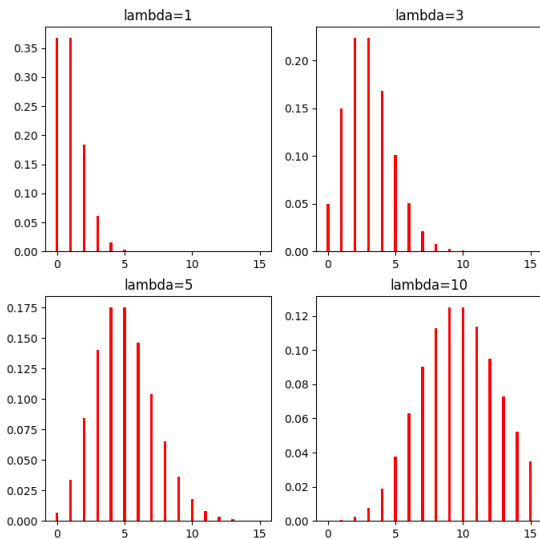
Let the random variable X be Poisson distributed

$$X \sim Po(\lambda), \quad (2-29)$$

where λ is the rate (or intensity): the average number of events per interval. The Poisson *pdf* describes the probability of x events in an interval

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}. \quad (2-30)$$

Eksempler på Poisson fordelinger



Poisson fordelings middelværdi og varians

||| Theorem 2.28 Mean and variance

A Poisson distributed random variable X has exactly the rate λ as the mean

$$\mu = \lambda, \quad (2-31)$$

and variance

$$\sigma^2 = \lambda. \quad (2-32)$$

Eksempel med Poisson fordelingen

Det antages, at der på en bestemt hospitalsafdeling i gennemsnit bliver indlagt 0.3 nye patienter pr. dag.

Hvad er sandsynligheden for at der på en vilkårlig dag bliver indlagt højst 2 nye patienter på afdelingen?

Hvad skal repræsenteres af den stokastiske variabel X ?

Hvad er fordelingen af X (og hvad er parametrene)?

Hvilken sandsynlighed skal udregnes?

At identificere den rigtige fordeling

Det er vigtigt at kunne identificere hvilken fordeling der er korrekt at bruge til en bestemt situation. Dette kan være ret svært og kræver øvelse.

- ① Antal udbedrede fejl. Er der et maksimalt antal? Med eller uden "tilbagelægning" (ændrer sandsynligheden sig efter hver udtrækning)?
- ② Antal harddiske med skrammer? Er der et maksimalt antal? Med eller uden "tilbagelægning" (ændrer sandsynligheden sig efter hver udtrækning)?
- ③ Antal indlagte patienter. Er der et maksimalt antal? Er der tale om en *rate*?

Hvad skal man være opmærksom på, når man skal identificere den rigtige fordeling?

Kahoot!

(x2)

02402 Statistik (Polyteknisk grundlag)

- Binomialfordelingen
- Den Hypergeometriske fordeling
- Poissonfordelingen
- Diskrete fordelinger i Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konkrete fordelinger i Python: `Scipy.stats`

Vi bruger `Scipy.stats` til de fleste kendte fordelinger i dette kursus

For eksempel kan vi bruge `scipy.stats.binom` til diverse beregninger i binomialfordelingen.

Dokumentation for `Scipy.stats` kan findes på internettet (det prøver vi lige).

Eksempel med binomialfordelingen i Python

Lad os tage eksemplet fra tidligere: Den stokastiske variabel X repræsenterer antallet af udbedrede fejl ud af de i alt 6 indrapporterede fejl. X følger en binomialfordeling med parametre: $n = 6$ og $p = 0.70$.

Vi ønsker at beregne $P(X = 6) = f(6) = \binom{n}{6}p^6(1-p)^{n-6}$

Denne beregning kan udføres på lommeregner, men vi kan også bruge Python's funktion:

```
stats.binom.pmf(k=6, n=6, p=0.70)
```

I Python bruges `pmf` i stedet for *pdf*, når der er tale om diskrete fordelinger.

- Gå til Python notebook "week2.ipynb" i VS Code
Afsnit "Probability distributions and scipy.stats"



Visual Studio Code

Fordelinger i Scipy.stats

`Scipy.stats.binom`

`Scipy.stats.hypergeom`

`Scipy.stats.poisson`

Generelle 'methods' for diskrete fordelinger:

<code>.rvs</code>	'random variates' (simulér tilfældige tal)
<code>.pmf</code>	'probability mass function' (pmf/pdf/tæthedsfunktion)
<code>.cdf</code>	'cumulative distribution function' (fordelingsfunktion)
<code>.ppf</code>	'percent point function' (invers cdf / fraktilfunktion)
<code>.mean / .var / .std</code>	'mean'/'variance'/'standard deviation'

Se også i formelsamling!

(Notation i bog og Python kan være forskellig)

Kahoot!
(x2)

Kontinuære stokastiske variable

Sandsynligheder - kontinuerte tilfælde

Hvad er sandsynligheden for et bestemt udfald af en kontinuer stokastisk variabel?

Overvej eksemplet med Benzinforbrug af tilfældigt udvalgt biltype.

Hvad er $P(X = 20\text{km/L})$?

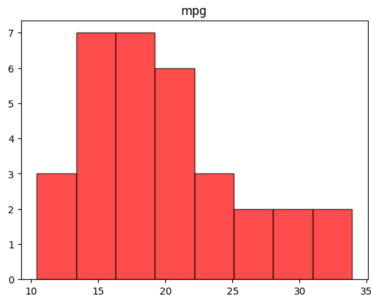
Hvad er $P(X \leq 20\text{km/L})$?

(Skriv dine overvejelser her)

Sandsynligheder - kontinuerte tilfælde

Vi har brug for mere information om den stokastiske variabel for at kunne svare på spørgsmålene.

Her er et histogram af "miles per gallon" (mpg) for en stikprøve med data for 32 forskellige biler:



Hvordan kan vi beskrive den **teoretiske** stokastiske variabel for "benzinforbrug af bil"?

Tæthedsfunktionen (kontinuære tilfælde)

(eng: *Probability density function*, **pdf**)

||| Definition 2.32 Density and probabilities

The *pdf* of a continuous random variable X is a non-negative function for all possible outcomes

$$f(x) \geq 0 \text{ for all } x, \quad (2-36)$$

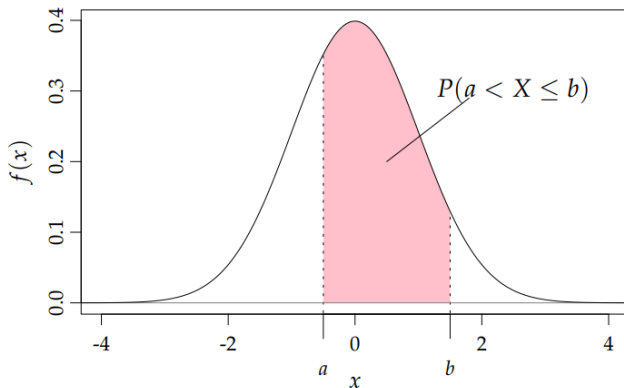
and has an area below the function of one

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (2-37)$$

It defines the probability of observing an outcome in the range from a to b by

$$P(a < X \leq b) = \int_a^b f(x) dx. \quad (2-38)$$

Eksempel på tæthedsfunktioner



Arealet under (hele) tæthedsfunktionen er 1.

Det skraverede areal illustrerer sandsynligheden for at den stokastiske variabel X vil antage en værdi der ligger mellem værdierne a og b .

Mere om sandsynligheder for kontinuære stokastiske variable

Sandsynligheden for at den stokastiske variabel X vil antage en værdi der ligger mellem værdierne a og b er givet ved integralet:

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

Vi kan derfor også se at sandsynligheden for at den stokastiske variabel antager én præcis værdi (med uendelig præcision) er lig nul.

OBS: I praksis vil rigtige data altid afspejle afrundede værdier - dvs. værdier der ligger indenfor et interval svarende til afrundingens præcision.

Fordelingsfunktionen (kontinuære tilfælde) (eng: *Cumulative distribution function*, **cdf**)

||| Definition 2.33 Distribution

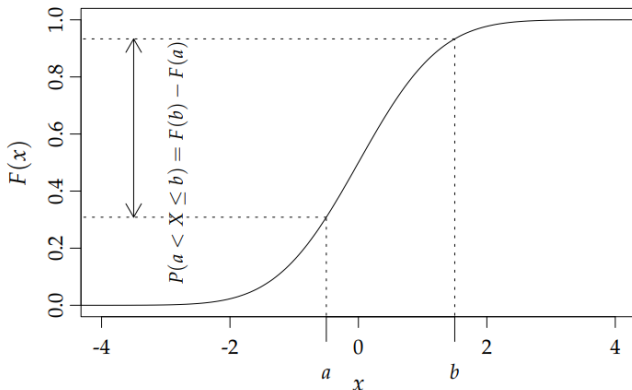
The *cdf* of a continuous variable is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du, \quad (2-40)$$

and has the properties (in both the discrete and continuous case): the *cdf* is non-decreasing and

$$\lim_{x \rightarrow -\infty} F(x) = 0 \text{ and } \lim_{x \rightarrow \infty} F(x) = 1. \quad (2-41)$$

Eksempel på fordelingsfunktioner



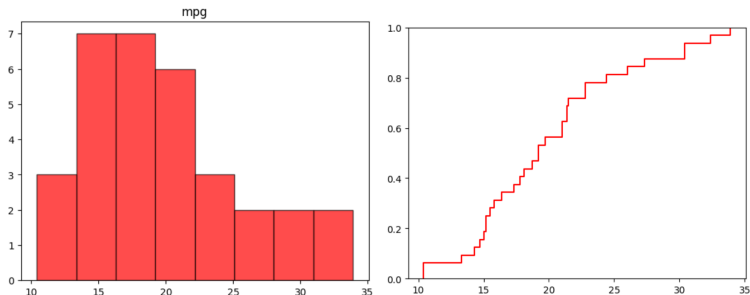
Fordelingsfunktionen er en alternativ måde at visualisere sandsynlighedsfordelingen af udfaldene på.

Fordelingsfunktionen vokser altid fra 0 til 1 (på y -aksen).

Histogram og ECDF

Har man et rigtigt datasæt vil histogrammet give en ide om hvordan en underliggende (teoretisk) tæthedsfunktion kunne se ud.

På tilsvarende vis kan et "empirisk cdf" (ecdf) give en ide om hvordan en underliggende (teoretisk) fordelingsfunktion kunne se ud:



Forventningsværdi, $E[X]$, og varians, $V[X]$ (kontinuære tilfælde)

||| Definition 2.34 Mean and variance

For a continuous random variable the mean or expected value is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx, \quad (2-44)$$

hence similar as for the discrete case the outcome is weighted with the *pdf*.
The variance is

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx, \quad (2-45)$$

For kontinuære stokastiske variable bruger vi integraler i stedet for summer.

Eksempel

En ingeniør skal bruge bærende stålprofiler til et byggeprojekt. Konstruktionskravet er, at stålprofilerne skal kunne bære mindst **350 kN**.

Ingeniøren skal vælge mellem to leverandører på baggrund af oplysninger om bæreevne (den belastning (i kN), som en stålprofil kan modstå, før brud opstår).

Fra de to leverandører (A og B) oplyses middelværdi samt variationskoefficient for bæreevnen af deres stålprofiler:

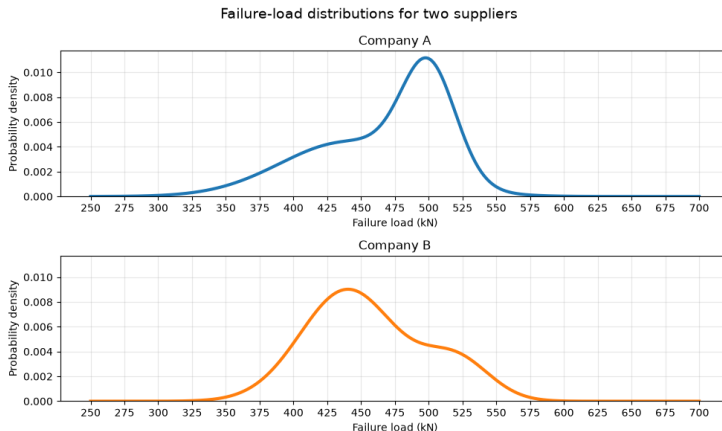
	leverandør A	leverandør B
Middelværdi	467.0 kN	456.8 kN
Variationskoefficient (CV)	10.6%	10.2%

Hvilken leverandør producerer de stærkeste stålprofiler?

Hvilken leverandør er "bedst"?

Eksempel

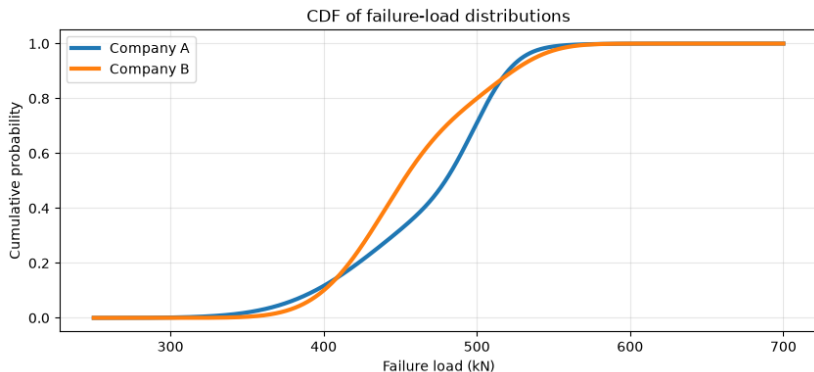
Ingeniøren beder om mere information og modtager fra leverandørerne et plot af bæreevne-fordelingen for deres stålprofiler:



Hvad viser plottet?

Eksempel

Her er samme data visualiseret med **kumuleret** fordeling:



Hvilken leverandør stiller bedst sikkerhed for at stålprofilerne har en bæreevne på mindst **350 kN**?

Eksempel

Ingeniøren tilføjer følgende oplysning til tabellen:

	leverandør A	leverandør B
Middelværdi	467.0 kN	456.8 kN
Variationskoefficient (CV)	10.6%	10.2%
$P(X < 350 \text{ kN})$	2.0%	0.4%

Her er X en **stokastisk variabel** der repræsenterer bæreevnen af stålprofilerne.

Oplysningen om $P(X < 350 \text{ kN})$ føres til protokol som argument for at vælge leverandør B til byggeprojektet.

Eksempel med kontinuær stokastisk variabel

En stokastisk variabel X følger
en tæthedsfordling $f(x)$:

$$f(x) = \begin{cases} x & 0 \leq x < 1 \\ 2 - x & 1 \leq x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Tegn $f(x)$ (pdf):

Kahoot!
(x5)

Hvad er sandsynligheden for at X er større end 1, $P(X \geq 1)$?

Hvad er sandsynligheden for at X er 0.5, $P(X = 0.5)$?

Hvad er $\mathbf{E}[X]$ (og kan du opskrive et udtryk for $\mathbf{V}[X]$)?

Tjekliste

Efter i dag skal du kunne:

- DATA m. PYTHON: Indtaste data fra en stikprøve i Python og beregne diverse nøgletal med Python.
- DATA m. PYTHON: Producere diverse plots til data visualisering med Python.
- TEORI: Kende og beskrive kendte diskrete sandsynlighedsfordelinger.
- TEORI: Identificere den relevante sandsynlighedfordeling til en given kontekst/anvendelse.
- TEORI m. PYTHON: Benytte Python til at finde diverse størrelser relateret til kendte konkrete sandsynlighedsfordelinger (fx sandsynligheder, fraktiler og lign.)
- TEORI m. PYTHON: Simulere tilfældige tal fra en given sandsynlighedfordeling.

Eksamensopgave III.1 fra Juni 2021

Exercise III

The card game Casino is played with an ordinary deck of cards, i.e. 52 cards, 13 of each colour (clubs, hearts, diamonds and spade), and 4 of each value (e.g. 4 aces). The game starts by placing 4 cards face up on the table.

Question III.1 (7)

What is the probability that at least 3 of the 4 opening cards are aces?

- 1 ☐ 0.00057
- 2 ☐ $2.7 \cdot 10^{-6}$
- 3 ☐ 0.025
- 4 ☐ 0.00071
- 5 ☐ $3.7 \cdot 10^{-6}$

Eksamensopgave III.1 fra Juni 2021

```
print(1 - stats.hypergeom.cdf(2, M=56, n=4, N=4))  
0.00057
```

```
print(1 - stats.hypergeom.cdf(3, M=56, n=4, N=4))  
2.7e-06
```

```
print(stats.hypergeom.pmf(2, M=52, n=4, N=4))  
0.025
```

```
print(1 - stats.hypergeom.cdf(2, M=52, n=4, N=4))  
0.0007
```

```
print(1 - stats.hypergeom.cdf(3, M=52, n=4, N=4))  
3.7e-06
```

Tirsdage 10-12, Bygning 116 (hjælpelærere ankommer 10.15)

- "Klasseundervisning":
 - **010**: Elektroteknologi, Softwareteknologi, Cyberteknologi
 - **011**: Bæredygtigt energidesign, Miljøteknologi, Design og innovation
 - **012**: Kemi og teknologi, Medicin og teknologi, Teknisk biomedicin, Bioteknologi, Life science
 - **013**: Byggeteknologi, Bygningsdesign, Produktion og konstruktion, General Engineering
 - **015**: Fysik og ingeniørvidenskab, Geofysik og rumteknologi, Anvendt matematik
 - **016**: Kunstig intelligens og data
- Arbejd selv og få hjælp efter behov:
 - **Foyer areas** (with circulating TA's): Blandet