02402 Statistik (Polyteknisk grundlag)

Uge 2:
Stokastiske variable,
Python til statistik,
Konkrete diskrete fordelinger

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Agenda

1. Stokastiske variable (fortsat)
2. Python til statistik
   - Deskriptiv statistik og data visualisering med Python
   - Simulering af stokastiske variable
3. Konkrete diskrete fordelinger
   - Binomialfordelingen
   - Den Hypergeometriske fordeling
   - Poissonfordelingen
   - Diskrete fordelinger i Python

02402 Statistik (Polyteknisk grundlag)

Stokastiske variable (fortsat)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Tæthedsfunktionen, **pdf** (diskrete tilfælde)

Definition 2.6 The *pdf* of a discrete random variable

For a discrete random variable X the *probability density function* (*pdf*) is

$$f(x) = P(X = x). \tag{2-9}$$

It assigns a probability to every possible outcome value x.
A discrete *pdf* fulfils two properties: there are no negative probabilities for any outcome value

$$f(x) \geq 0 \text{ for all } x, \tag{2-10}$$

and the probabilities for all outcome values sum to one

$$\sum_{\text{all } x} f(x) = 1. \tag{2-11}$$

Fordelingsfunktionen, **cdf** (diskrete tilfælde)

Definition 2.9 The *cdf* of a discrete random variable

The *cumulated distribution function* (*cdf*) for the discrete case is the probability of realizing an outcome below or equal to the value x

$$F(x) = P(X \leq x) = \sum_{j \text{ where } x_j \leq x} f(x_j) = \sum_{j \text{ where } x_j \leq x} P(X = x_j). \qquad (2\text{-}12)$$

Forventningsværdi, $E[X]$ (diskrete tilfælde)

Definition 2.13 Mean value

The mean of a discrete random variable X is

$$\mu = \mathrm{E}(X) = \sum_{j=1}^{\infty} x_j f(x_j), \tag{2-15}$$

where x_j is the value and $f(x_j)$ is the probability that X takes the outcome value x_j.

En "vægtet" sum, hvor værdierne (x_j) vægtes med deres sandsynlighed $(f(x_j))$.

Varians, $V[X]$ (diskrete tilfælde)

Definition 2.16 Variance

The variance of a discrete random variable X is

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}[(X-\mu)^2] = \sum_{i=1}^{\infty}(x_i-\mu)^2 f(x_i), \qquad (2\text{-}17)$$

where x_i is the outcome value and $f(x_i)$ is the *pdf* of the ith outcome value. The *standard deviation* σ is the square root of the variance.

Tæthedsfunktionen, **pdf** (kontinuærte tilfælde)

Definition 2.32 Density and probabilities

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

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

and has an area below the function of one

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

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

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx. \tag{2-38}$$

Fordelingsfunktionen, **cdf** (kontinuærte tilfælde)

Definition 2.33 Distribution

The *cdf* of a continuous variable is defined by

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

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

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

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

Definition 2.34 Mean and variance

For a continuous random variable the mean or expected value is

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx, \tag{2-44}$$

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

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

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

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 pdf(x):

Hvad er sandsynligheden for at X er større end 1?; $P(X > 1)$?

Kahoot!
(x1)

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

02402 Statistik (Polyteknisk grundlag)

Python til statistik

- Deskriptiv statistik og data visualisering med Python
- Simulering af stokastiske variable

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Få hjælp hos Python Supporten!

https://pythonsupport.dtu.dk/

Python

I 02402 bruger vi VS Code og jupyter notebooks.

Visual Studio Code

I denne uge:

- Python til deskriptiv statistik (beregne nøgletal, visualisering)
- Simulering af *tilfældige tal* (stokastiske variable)
- Simulering fra konkrete fordelinger

DataCamp

Som forsøg vil I få adgang til **www.DataCamp.com**,
hvor I kan opfriske eller udvide jeres Python kompetencer.

Brugen af DataCamp er frivillig.

DataCamp er en online platform med en masse små kurser indenfor kodning og Data Science. Prøv fx kurserne:

- "**Introduction to Python**": Python basics, data typer og intro til Numpy arrays
- "**Intermediate Python**" (Chapter 1): Data visualisering med Matplotlib
- "**Python for R Users**": Hvis man har været vandt til at bruge R

Python til eksamen

Til eksamen er det ikke tilladt at medbringe computer!

Der vil være opgaver hvor man skal kunne
læse og forstå Python kode og output.

Vi laver nogle eksempler i løbet af kurset.

Til eksamen må man medbringe en simpel lommeregner

Vi anbefaler at man anskaffer en TI30xs - der kommer afgørelse om lommeregner snarest muligt.

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.

02402 Statistik (Polyteknisk grundlag)

- Deskriptiv statistik og data visualisering med Python
- Simulering af stokastiske variable

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Python til deskriptiv statistik

- Indlæse data fra en stikprøve
- Beregne diverse statistiske nøgletal
- Data visualisering (plots)

- Gå til Python notebook "descriptive_statistics.ipynb" i VS Code (+ Kahoots)

02402 Statistik (Polyteknisk grundlag)

- Deskriptiv statistik og data visualisering med Python
- Simulering af stokastiske variable

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

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 eksempel kan vi *simulere* 10 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.

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 "first_simulation_in_python.ipynb" i VS Code (+ Kahoot)

Eksempler med simulering

Prøv at lege med Example 2.15 og Example 2.19 i bogen

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

$$X \sim B(n, p), \tag{2-19}$$

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

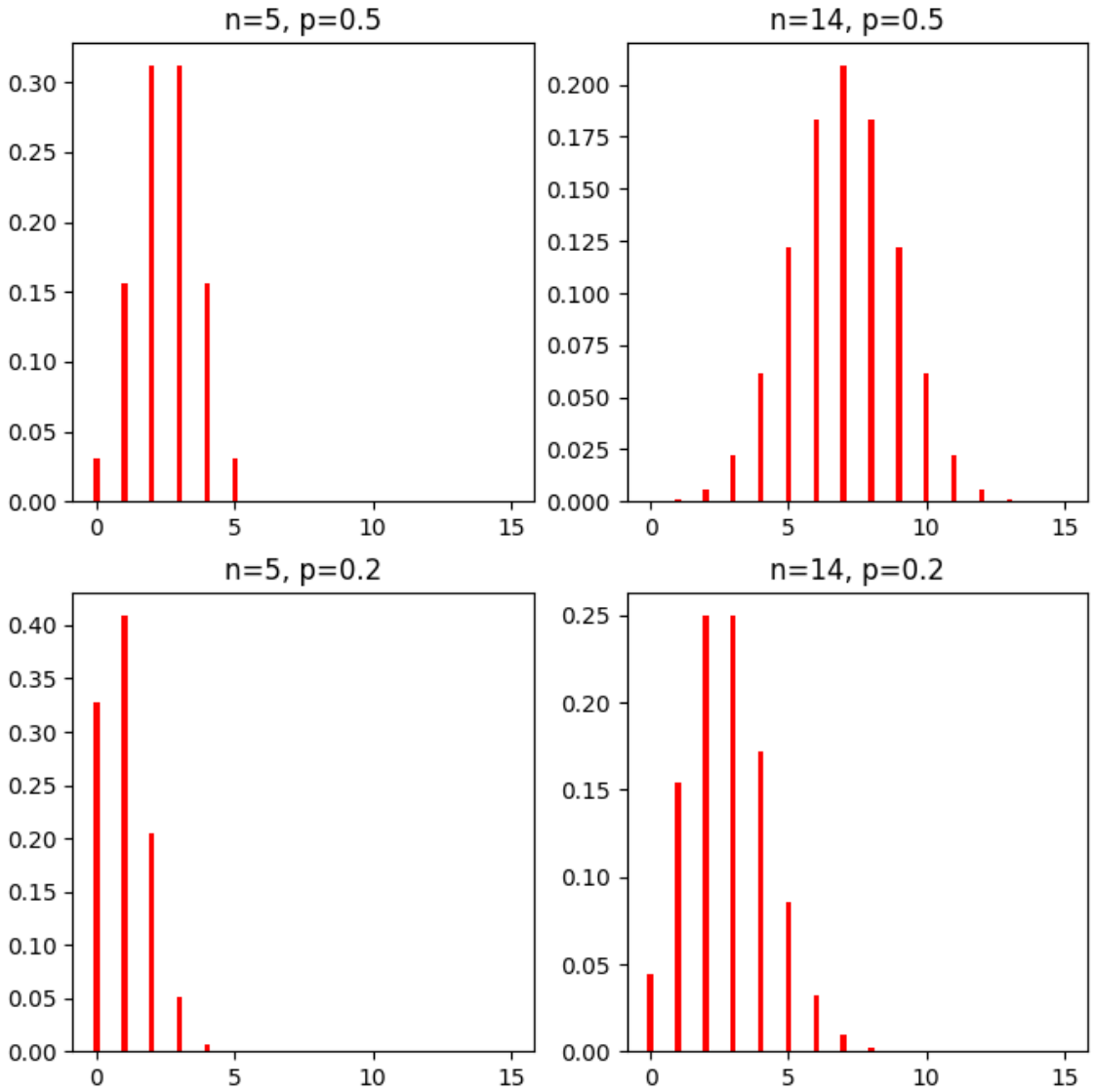
$$f(x; n, p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2-20}$$

where

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

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

Eksempler på binomialfordelinger



Binomialfordelingens middelværdi og varians

Theorem 2.21 Mean and variance

The mean of a binomial distributed random variable is

$$\mu = np, \tag{2-22}$$

and the variance is

$$\sigma^2 = np(1-p). \tag{2-23}$$

Eksempel med binomialfordelingen

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt at indrapporterede fejl bliver udbedret samme dag.
Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

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

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

$$X \sim H(n, a, N), \tag{2-24}$$

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*

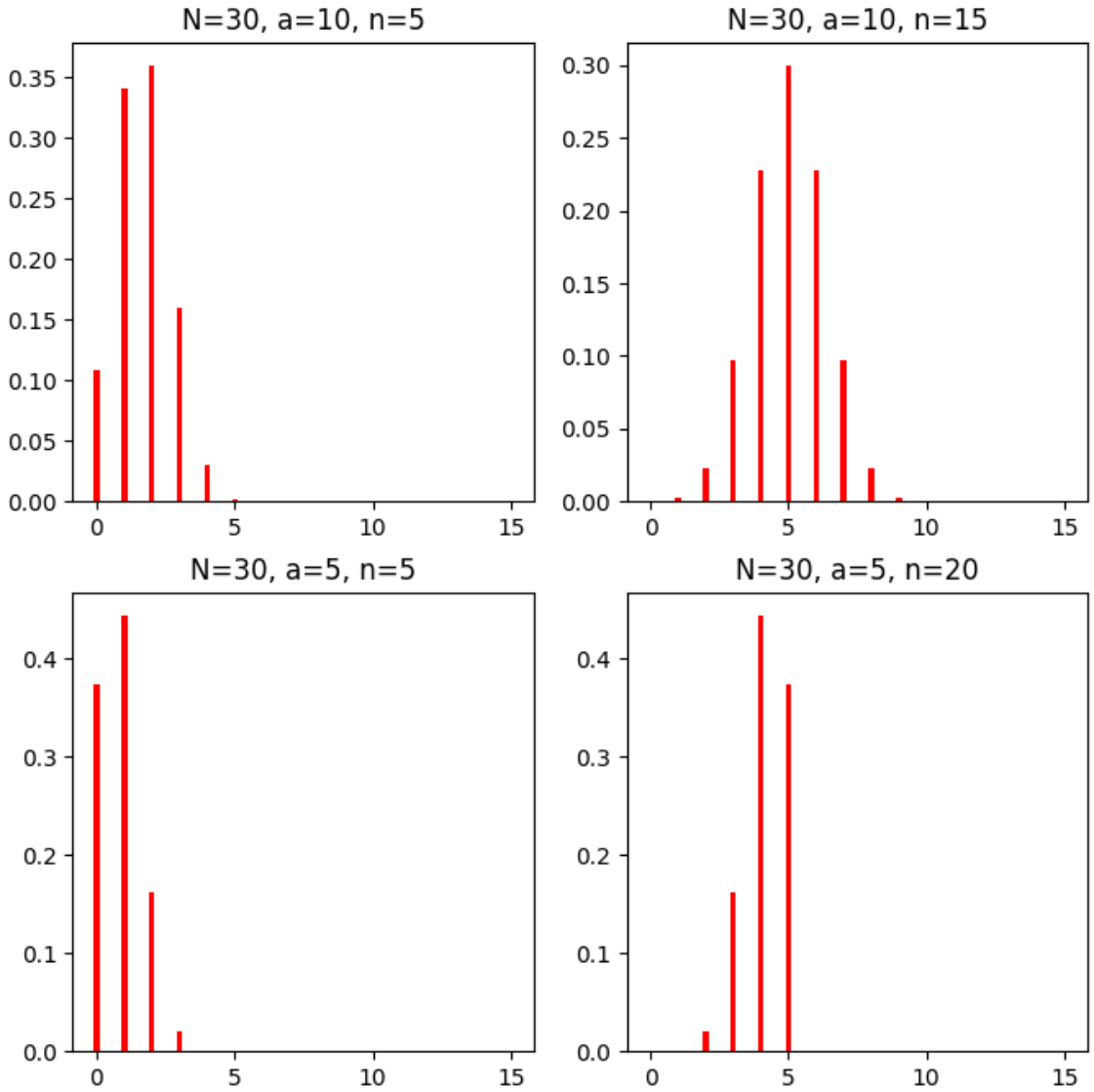
$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}. \tag{2-25}$$

The notation

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

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

$$\mu = n\frac{a}{N}, \tag{2-27}$$

and the variance is

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

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 fordelingens 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 fordelingens 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 fordelingens tæthedsfunktion

Definition 2.27 Poisson distribution

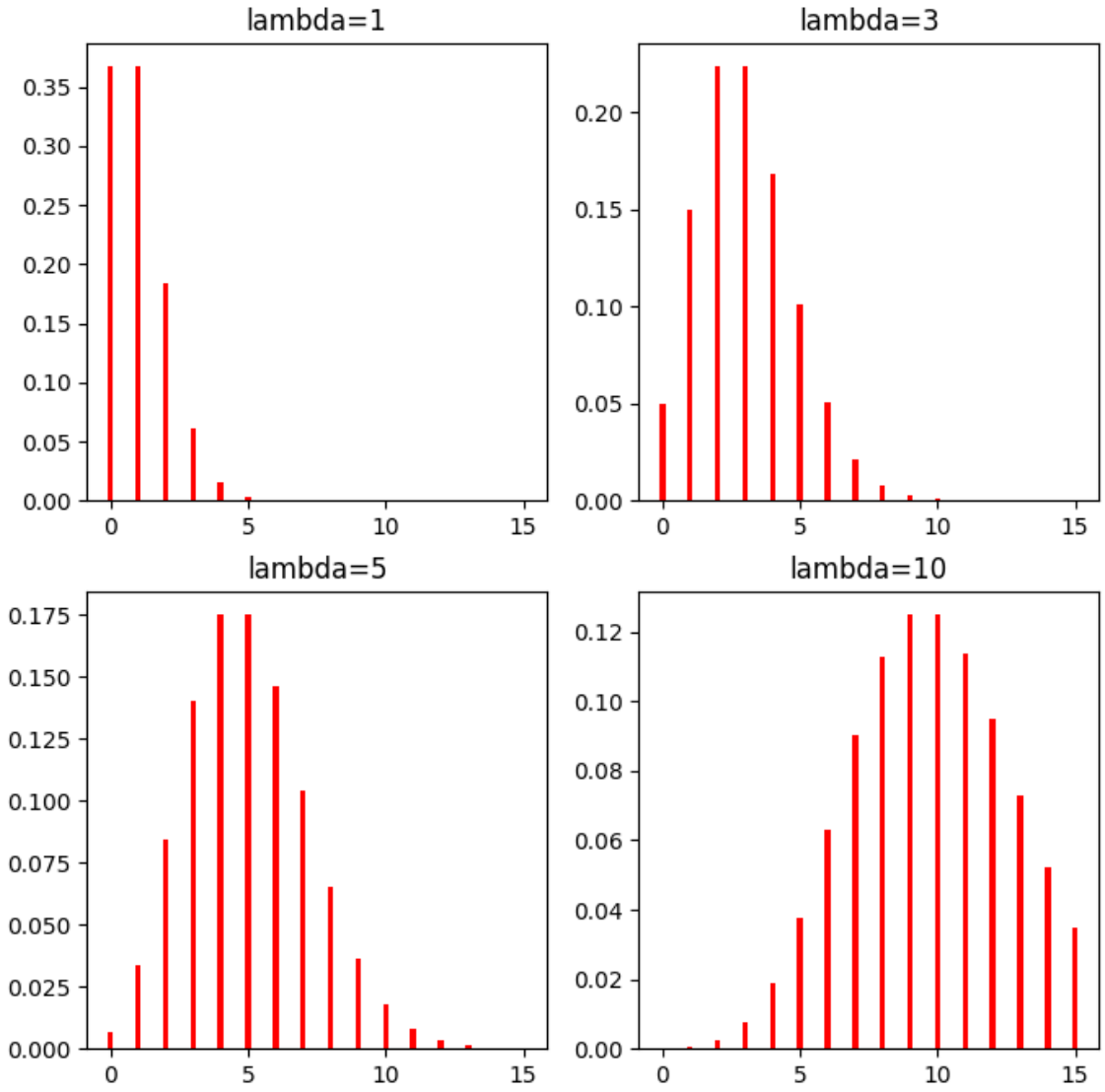
Let the random variable X be Poisson distributed

$$X \sim Po(\lambda), \tag{2-29}$$

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

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

Eksempler på Poisson fordelinger



Poisson fordelingens middelværdi og varians

Theorem 2.28 Mean and variance

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

$$\mu = \lambda, \tag{2-31}$$

and variance

$$\sigma^2 = \lambda. \tag{2-32}$$

Eksempel med Poisson fordelingen

Det antages, at der i gennemsnit bliver indlagt 0.3 patienter pr. dag på københavnske hospitaler som følge af luftforurening.

Hvad er sandsynligheden for at der på en vilkårlig dag bliver indlagt højst 2 patienter som følge af luftforurening?

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

Til eksamen kan der være opgaver hvor man selv skal identificere hvilken fordeling der er korrekt at bruge ud fra en kontekst. 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?

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 konkrete fordelinger i dette kursus (men også `Numpy.random.choice`).

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 a 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 "binomial_distribution.ipynb" i VS Code

Fordelinger i Scipy.stats

`Scipy.stats.binom`
`Scipy.stats.hypergeom`
`Scipy.stats.poisson`

Generelle 'methods' for diskrete fordelinger:

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

Se også Appendix A.2.1 i bogen.

Kahoot!
(x1)

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.

Øvelser

- Bygning 306 1. sal.
  - 105 (øvelseslokale 96). Hjælpelærer: Nuria
  - 122 (øvelseslokale 98). Hjælpelærer: Ali (KID students)
  - 119 (øvelseslokale 99). Hjælpelærer: Alfred (KID students, overflow)
  - 108A. Hjælpelærer: Afonso
  - 108B. Hjælpelærer: Uffe
  - Ved pladsmangel kan man også sidde i foyerområdet ved trappen.
- Bygning 324, stueetagen.
  - 060. Hjælpelærer: Phd-student Kyril
  - 040. Hjælpelærer: Phd-student Thea
  - 050. Hjælpelærer: Jakob
  - 020. Hjælpelærer: Drin
  - 030. Hjælpelærer: Maliha
  - Man kan også sidde i foryerområde 003, 004, 005 og 008.