02402 Statistik (Polyteknisk grundlag)

Uge 1: Introduktion og Python

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

IntroStat team


Pernille Y. Nielsen


M S Khalid


Jan K. Møller


Nicolai S. Larsen


Peder Bacher

- 02402 Statistik (Polyteknisk Grundlag)
- 02323 Introduktion til statistik
- 02403 Introduktion til matematisk statistik

Special for Fall 2024

This semester 02323 will have lectures in English (given by M.S. Khalid).

- 02323 lectures in English: Friday 8-10.

Students following the course 02323, can come to lectures in 02402 if they prefer Danish (Tuesdays 8-10).

NOTE: There are small differences between the two courses (2-way ANOVA not covered in 02323).

Agenda

1. Praktiske informationer
2. Introduktion og motivation
3. Deskriptiv Statistik
   - Middelværdi og median (centralitetsmål)
   - Varians og standardafvigelse
   - Fraktiler
   - Kovarians og korrelation
4. Software: Python

Overview

Praktiske informationer

- Undervisning
  - Forelæsninger: Tirsdag 8-10
    - Bygning 306, Aud. 33, 34 (og 32).
  - Øvelser: Tirsdag 10-12
    - Bygning 303A, øvelsesområder HVEST/HOEST.
    - Bygning 324, stueetagen (Foyer øvelsesområder: 003, 004, 005 og 008. Lokaler: 020, 030, 040, 050, 060 og 070).
- Obligatoriske projekter
  - 2 projekter, som skal bestås for at kunne gå til eksamen.
  - Projekterne afleveres i kursus-uge 7 og 10 (tirsdag).
  - For hvert projekt vælges et af fire emner.
  - De som tidligere har bestået behøver ikke at lave projekterne igen.
- Eksamen
  - Søndag den 15. december 2024.
  - 4 timers multiple choice-prøve.

Praktiske informationer

- Generel ugeseddel
  - Før undervisningen: Læs de relevante afsnit bogen
  - Forelæsninger: Gennemgang af ugens pensum
  - Øvelser: Opgaveregning med brug af Python
  - Efter undervisningen: test-quizzer på hjemmesiden

Kahoot!

Praktiske informationer

- Kursushjemmeside: 02402.compute.dtu.dk
  - Undervisningsplan og pensum (agenda)
  - Bog
  - Øvelser og løsninger (engelsk)
  - Dias (dansk og engelsk)
  - Tidligere års forelæsninger (dansk og engelsk)
  - Quizzer
- DTU Learn
  - Announcements
  - Projekter - formulering og aflevering
- Ed Discussion
  - Spørgsmål og diskussioner

Python

Ud over papir og blyant skal vi bruge Python til at regne statistiske opgaver.

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

Brugen af DataCamp er frivillig.

Overview

Indledning

Statistik er grundlæggende en matematisk videnskab om indsamling, beskrivelse, analyse og fortolkning af data.
Man vil uddrage viden og lære fra observerede data.

Sandsynlighedsregning er en gren af matematik, der beskæftiger sig med beskrivelse og analyse af tilfældighed.
Man vil udlede viden og lære fra en teoretisk model.

Felterne er svære at adskille, og metoder fra begge felter bruges almindeligvis sammen i ingeniørarbejde.

Et fælles mål: Beskrive og forstå tilfældig variation og usikkerheder kvantitativt!

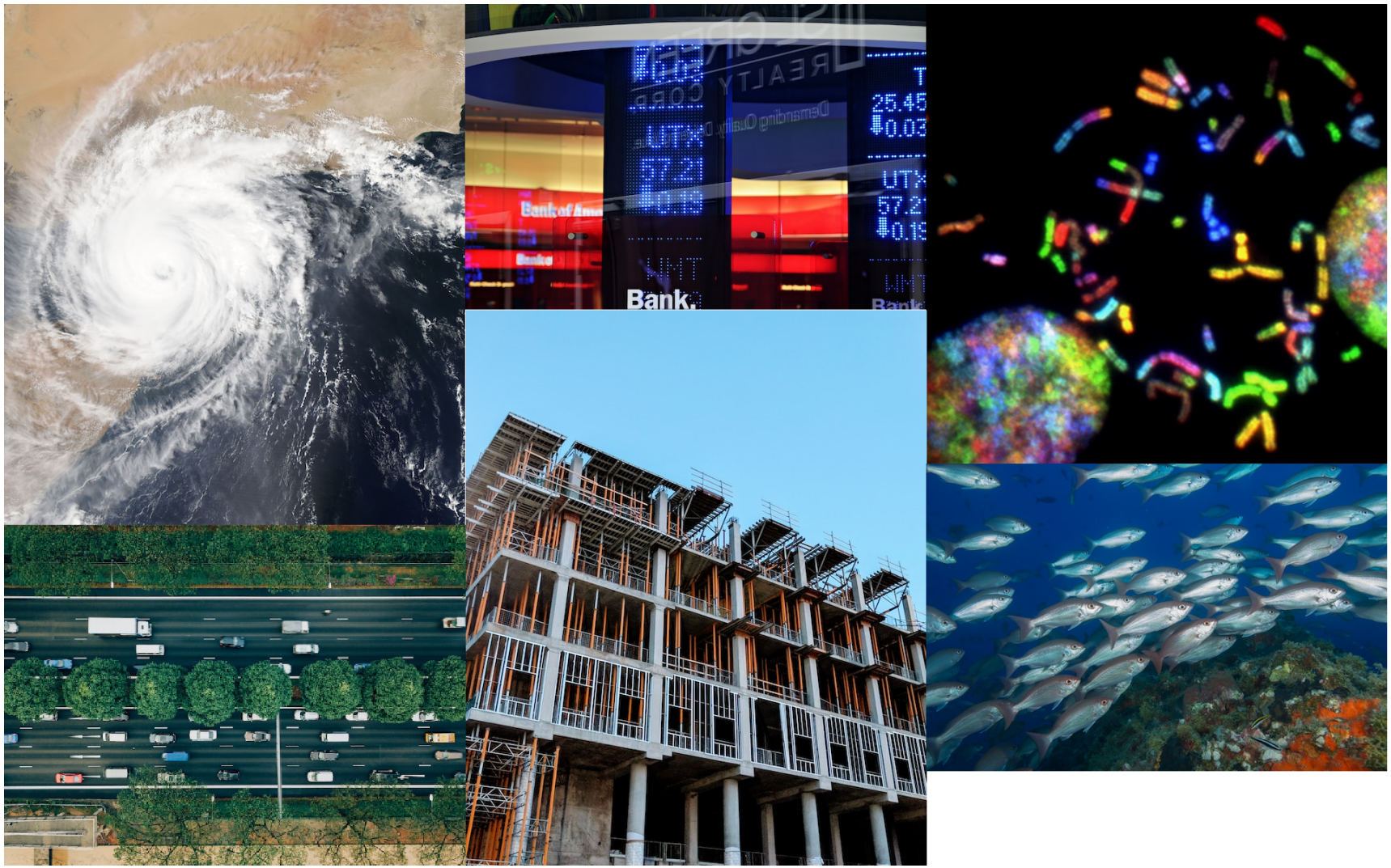
Forskellige aspekter

Statistik anvendes i mange områder, f.eks.:

- Data Science
- Kliniske studier og epidemiologi (eksempler fra bogen: James Lind 1747 og John Snow 1854)
- Produktion, planlægning og kvalitetskontrol
- Analyse af laboratoriedata og eksperimenter
- Machine learning og AI

Der er mange spændende forskningsområder inden for både anvendt og teoretisk statistik.

Anvendelse



Intro case-historier:
IBM big data, Novo Nordisk small data, Skive fjord

- Præsentation af Senior Scientist Hanne Refsgaard, Novo Nordisk A/S
- *IBM Social Media* podcast af Henrik H. Eliassen, IBM
- *Skive Fjord* podcasts, af Jan K. Møller, DTU

I hverdagen

Statistik eller elementer fra faget forekommer mange steder i hverdagen, herunder:

- Nyheder
- Politik
- Reklamer
- Sport
- Arbejde

Statistik bruges ofte som beslutningsstøtte! Statistik kan bruges til at bestemme, hvad man skal undersøge nærmere.

Almindelige fejlslutninger og bias

Statistik kan være kontraintuitivt, og vores hjerner skal trænes i statistisk tænkning for ikke at lave en række almindelige fejlslutninger. *Selv veluddannede, professionelle statistikere begår simple fejl*.

Nogle typiske biases (systematiske skævvridninger) i statistik er:

- Udvælgelsesbias
- Overlevelsesbias
- OVB (Omitted-variable bias)

Den sidste bias er tæt knyttet til koncepterne p-hacking og konfunderende variable.

Kursets overordnede mål og afgrænsning

Kurset skal bl.a. gøre jer bedre til at:

- Behandle og analysere data hensigtsmæssigt
- Beskrive og forstå tilfældig variation og usikkerheder
- Tænke kritisk over statistiske udsagn
- Forstå mulighederne og begrænsningerne af statistik

Kurset skal også forberede jer til diverse videregående kurser.

Kursets indhold i store træk

En stor del af kurset omhandler:

1. Formulering af modeller
2. Udregning af konfidensintervaller
3. Udførsel af hypotesetest
4. afgøre om diverse forhold er *statistisk signifikante*

i forskellige kontekster og setups.

Sandsynlighedsregningen bliver vores primære værktøj.

Grundlæggende om statistik

Statistik kan generelt opdeles i to dele:

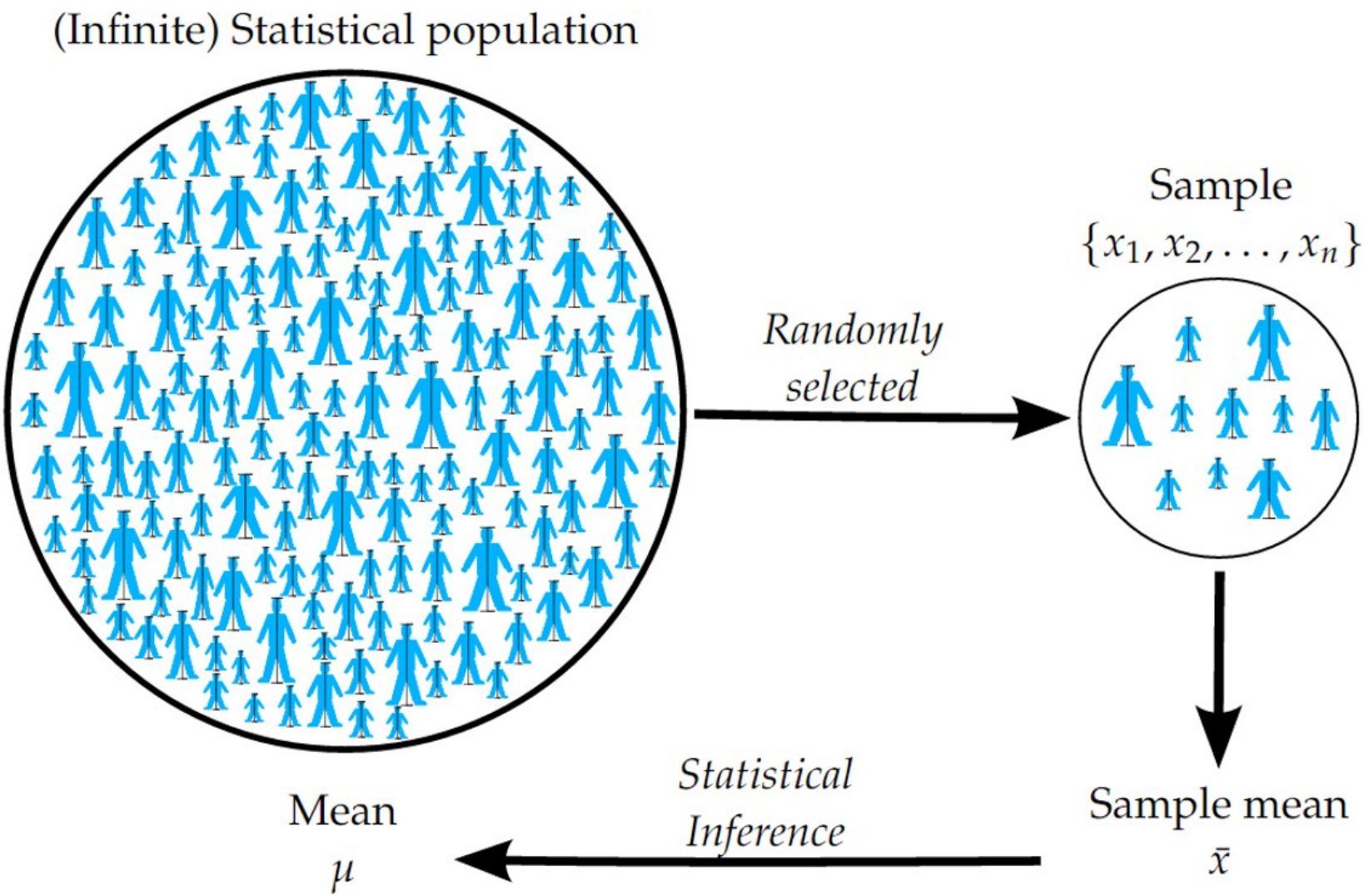
- Beskrivende statistik (deskriptiv statistik)
- Konkluderende statistik (statistisk inferens)

Statistik handler typisk om at analysere en *stikprøve*, taget ud af en *population*.

Ud fra stikprøven, udtaler vi os generelt om populationen.

Det er derfor vigtigt at stikprøven er *repræsentativ* for stikprøven. *I langt det meste af kurset vil vi bare antage, at stikprøverne er repræsentative.*

Populationen og stikprøven



Overview

1. Praktiske informationer
2. Introduktion og motivation
3. Deskriptiv Statistik
   - Middelværdi og median (centralitetsmål)
   - Varians og standardafvigelse
   - Fraktiler
   - Kovarians og korrelation
4. Software: Python

Det generelle setup

Der er en underliggende population, hvorfra der er udtaget en repræsentativ stikprøve med n observationer.

Stikprøven bliver almindeligvis repræsenteret med en vektor

$$x = (x_1, x_2, \ldots, x_n).$$

Den sorterede stikprøve er så

$$(x_{(1)}, x_{(2)}, \ldots, x_{(n)}),$$

hvor $x_{(1)}$ angiver den mindste observation og $x_{(n)}$ angiver den største observation.

Nøgletal (Summary statistics)

Nøgletal bruges til at opsummere og beskrive data.

- *Positionsmål*
  - f.eks.: gennemsnit, median og fraktiler
- *Spredningsmål*
  - f.eks.: varians og standardafvigelse
- *Sammenhængsmål*
  - f.eks.: kovarians og korrelation

Husk at skelne mellem nøgletal for populationen og stikprøven!

Gennemsnit, definition 1.4

Gennemsnittet er et nøgletal, der angiver tyngdepunktet for data.

Middelværdien af en stikprøve (Stikprøvegennemsnittet):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Vi siger, at $\bar{x}$ er et *estimat* for populationens middelværdi.

Median, Definition 1.5

Medianen er et også nøgletal, der angiver centreringen for data.

I nogle tilfælde, f.eks. hvis man har ekstreme observationer, er medianen at foretrække frem for gennemsnittet.

Medianen af en stikprøve (stikprøvemedianen):
Den midterste observation (af de sorterede data) eller gennemsnittet af de to midterste observationer (af de sorterede data) afhængigt af, om stikprøven har et lige eller ulige antal observationer.

Eksempel: Højde på studerende

- **Stikprøve:** Studerendes højde i cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Gennemsnit:**

$$\bar{x} = \frac{1}{5}(185 + 184 + 194 + 180 + 182) = 185$$

- **Median:**
  - Først sorteres data: (180, 182, 184, 185, 194).
  - Da n er ulige, vælges det midterste tal: 184.

- Hvis vi tilføjer en 235 cm høj person til stikprøven: (180, 182, 184, 185, 194, 235)
  - *Gennemsnit:* 193
  - *Median:* 184.5

Stikprøvevarians (sample variance) og -standardafvigelse (sample standard deviation), Definition 1.10

Stikprøvevariansen indikerer, hvor meget observationerne er spredt:

- Stikprøvevarians

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Stikprøvestandardafvigelse

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Eksempel med spredning: Højde på studerende

- **Stikprøve:** Studerendes højde i cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Stikprøvevarians:**

$$s^2 = \frac{1}{4}((185-185)^2 + (184-185)^2 + \cdots + (182-185)^2) = 29$$

- **Stikprøvestandardafvigelse:**

$$s = \sqrt{29} = 5.385$$

Kahoot!

Variationskoefficienten (coefficient of variation), Definition 1.12

Standardavigelsen og variansen er de primære nøgletal til at beskrive variationen i data.

Nogle gange ønsker man at sammenligne variationen mellem forskellige datasæt; da kan det være en god ide at se på et forholdsmæssigt tal:

Variationskoefficient:

$$CV = \frac{s}{\bar{x}}$$

Fraktiler (percentiles eller quantiles)

Medianen beregnes som det punkt, der deler data ind i to halvdele. Mere generelt kan vi beregne *fraktiler*. Ofte beregner man:

- $0\%, 25\%, 50\%, 75\%, 100\%$-fraktilerne

Bemærk:

- Medianen er 50%-fraktilen.
- 25%, 50%, 75%-fraktilerne kaldes hhv. *første*, *anden* og *tredje* kvartil, betegnet med hhv. $Q1$, $Q2$ og $Q3$.
- Dette giver anledning til spredningsmålet *den interkvartile variationsbredde* (*Inter Quartile Range* eller IQR): $Q3 - Q1$

Fraktiler, Definition 1.7

p-fraktilen, q_p, kan defineres ud fra følgende procedure:

1. Sorter de n observationer fra mindst til størst: $(x_{(1)}, x_{(2)}, ..., x_{(n)})$.
2. Beregn pn.
3. Hvis pn er et heltal: Tag gennemsnittet af den pn'te og den $(pn+1)$'te ordnede observation:

$$q_p = \left(x_{(np)} + x_{(np+1)}\right)/2$$

4. Hvis pn ikke er et heltal:

$$q_p = x_{(\lceil np \rceil)}$$

hvor $\lceil np \rceil$ er *ceiling*("loftet") af np, dvs. det mindste heltal større en np. Man afrunder altså np op til nærmeste heltal.

Eksempel: Højde på studerende

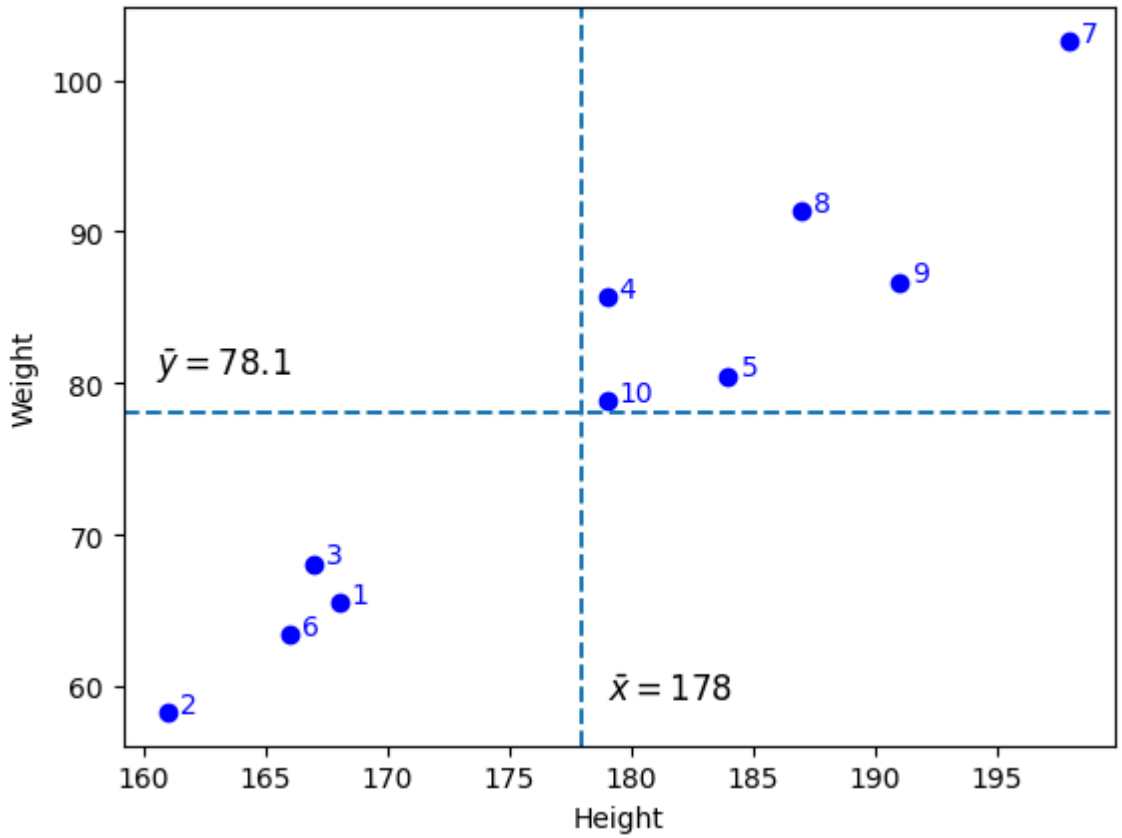
- **Sorteret stikprøve:** Studerendes højde i cm, $n = 5$.

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (180, 182, 184, 185, 194)$$

- **Nedre kvartil, Q1:**
  - Her er $p = 0.25$ og $n = 5$, hvorfor $np = 1.25$.
  - Det mindste heltal større end np er 2.
  - $Q1 = q_{0.25} = x_{(\lceil 1.25 \rceil)} = x_{(2)} = 182$.
- **Øvre kvartil, Q3:**
  - Her er $p = 0.75$ og $n = 5$, hvorfor $np = 3.75$.
  - Det mindste heltal større end np er 4.
  - $Q3 = q_{0.75} = x_{(\lceil 3.75 \rceil)} = x_{(4)} = 185$.
- **IQR:**
  - $Q3 - Q1 = 185 - 182 = 3$.

Kovarians og korrelation - Sammenhængsmål

Observation nr.:	1	2	3	4	5	6	7	8	9	10
Højde (cm) - (x_i)	168	161	167	179	184	166	198	187	191	179
Vægt (kg) - (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Stikprøvekovarians og -korrelation - Def 1.18 og 1.19

Stikprøvekovariansen er defineret ved

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Stikprøvekorrelationskoefficienten er defineret ved

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$$

hvor s_x og s_y standardfvigelserne for hhv. x og y.

Stikprøvekovarians og -korrelation

Studerende (ID)	1	2	3	4	5	6	7	8	9	10
Højde (cm) - (x_i)	168	161	167	179	184	166	198	187	191	179
Vægt (kg) - (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(x_i - \bar{x})$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_i - \bar{y})$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(x_i - \bar{x})(y_i - \bar{y})$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

$$\begin{aligned} s_{xy} &= \frac{1}{9}(126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 \\ &\quad + 119.6 + 111.7 + 0.8) \\ &= \frac{1}{9} \cdot 1493.3 \\ &= 165.9 \end{aligned}$$

$$s_x = 12.21 \quad \text{og} \quad s_y = 14.07$$

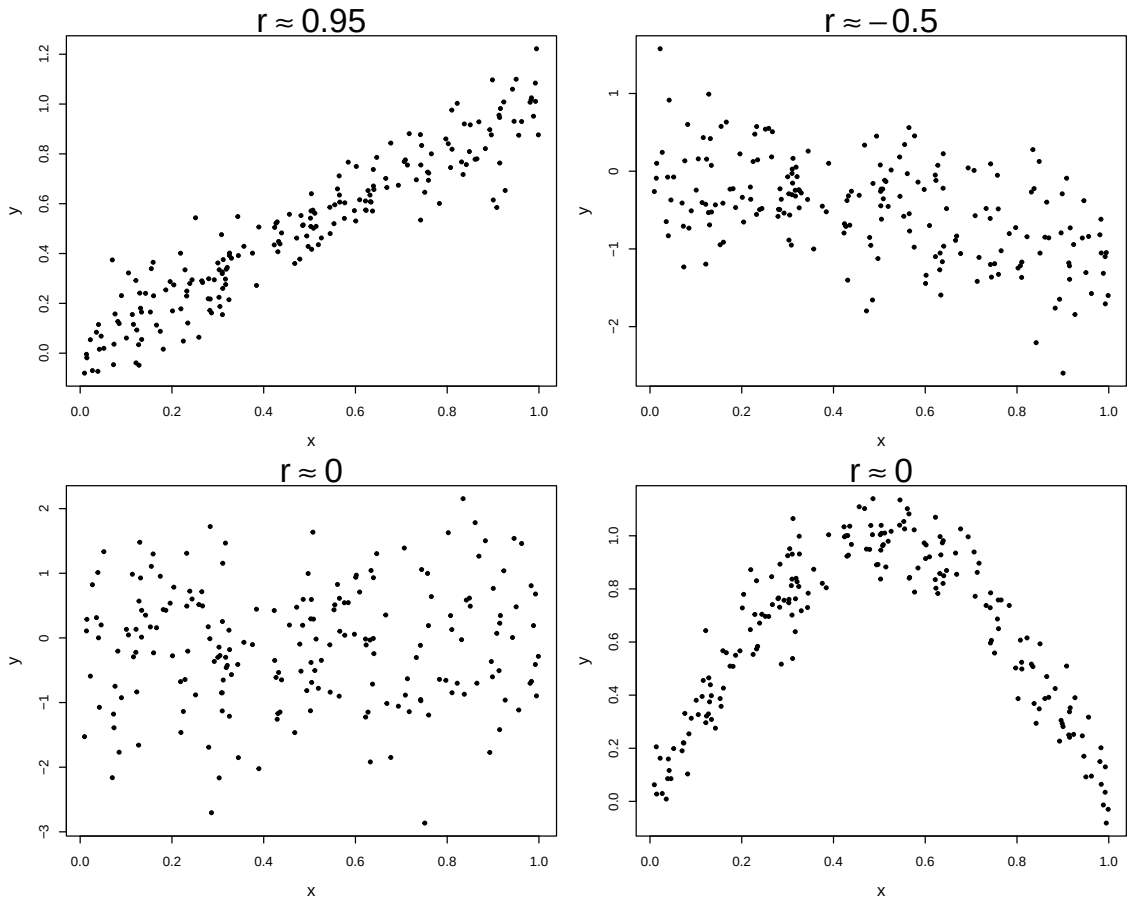
$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

Egenskaber for korrelationskoefficienten

De vigtigste egenskaber for korrelationskoefficienten er:

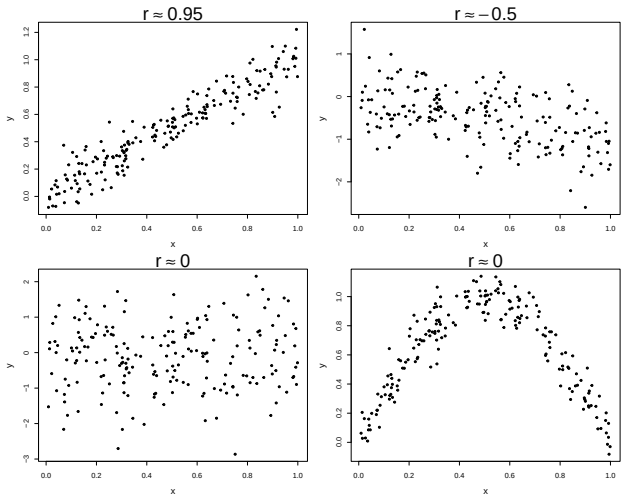
- r er altid mellem -1 og 1: $-1 \leq r \leq 1$
- r er et mål for lineær sammenhæng mellem x og y
- $r = \pm 1$ hvis og kun hvis punkterne ligger på en ret linie
- $r > 0$ hvis den generelle trend i scatterplottet er positiv
- $r < 0$ hvis den generelle trend i scatterplottet er negativ

Korrelation



Figurer/Tabeller

- Numeriske data
  - Scatterplot (xy plot)
  - Histogram
  - Kumuleret fordeling
  - Boxplot
- Tælledata
  - Søjlediagram (bar chart)
  - Cirkeldiagram (pie chart)



Visualisering af data er vigtigt!
Vi laver forskellige figurer og tabeller i Python

Overview

Software: Python

- Python: Software/programmeringssprog for statistisk analyse og datavisualisering (og mange andre ting).
- Python: Gratis, open source, virker på alle platforme.
- Mange Python "libraries" til alskens dataanalyse.
- Introduceres i bogen.
- Integreret del af kurset.
- Learning by doing.

Kahoot!

Software: Python

- I 02402 bruger vi VS Code til at editere og køre vores Python programmer
- Vi bruger primært jupyter notebooks (.ipynb)
- Vi forventer at I har installeret både Python og VS Code og at I er i stand til at oprette en jupyter notebook
- Hjælp til installering (og mere): https://pythonsupport.dtu.dk/

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 scipy.stats as stats
- import pandas as pd
- import statsmodels.api as sm
- import statsmodels.formula.api as smf
- import statsmodels.stats.power as smp

Python

- Gå til dagens Python notebook i VS Code

Opsummering fra i dag:

1. Praktiske informationer
2. Introduktion og motivation
3. Deskriptiv Statistik
   - Middelværdi og median (centralitetsmål)
   - Varians og standardafvigelse
   - Fraktiler
   - Kovarians og korrelation
4. Software: Python

Lidt mere Kahoot

(x8)

Øvelser

- Øvelser starter kl 10.15
- I finder øvelserne på kursushjemmesidens "Agenda" (øvelserne findes ligesom bogen kun på engelsk)
  - Start med at åbne VS Code og oprette en ny notebook (tjek at det virker ved at beregne 2+2)
  - Lav herefter dagens øvelser (både med og uden Python)
  - Det er også en god mulighed for at diskutere pensum og stille spørgsmål til hjælpelærene
- Alle hjælpelærere taler engelsk, mange taler også dansk
- Lokaler i stueetagen af 324 og 303A
- KID studerende i 303A foyer
- Dem der har problemer med at Python og VS Code kan gå til rum 040