

# 02402 Statistik (Polyteknisk grundlag)

## Uge 1: Introduktion, Deskriptiv statistik, Stokastiske variable

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**

(This is us!)

(lectures in **english**, Friday 8-10)

(held in June 3-week period)

NOTE: There are small differences between the courses (some parts are **not** covered in 02323).

# Agenda

## 1 Praktiske informationer

## 2 Introduktion

## 3 Deskriptiv Statistik

- Statistiske nøgletal
- Data visualisering

## 4 Stokastiske variable

- Sandsynligheder, pdf, cdf og forventningsværdier

## 5 Husk før næste uge

# 02402 Statistik (Polyteknisk grundlag)

## Praktiske informationer

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Forelæsninger

Tirsdag 8-10

- Bygning 306, Aud. 33 + stream i Aud 32.
- Online via Learn - "Video & Streaming".

Husk at opdatere siden kl 8.00 (evt et par gange) - video'en bliver først synlig når forelæsningen går i gang.

# Øvelser

Tirsdag 10-12

- 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 Cyril
  - 040. Hjælpelærer: Phd-student Thea
  - 050. Hjælpelærer: Jakob
  - 020. Hjælpelærer: Drin
  - 030. Hjælpelærer: Maliha
  - Man kan også sidde i foyerområde 003, 004, 005 og 008.

# Projekter

## Projekter

- I løbet af kurset har vi to frivillige afleveringer ("projekter"):
  - Projekt 1: afleveres i uge 7
  - Projekt 2: afleveres i uge 10
- Man modtager feedback på sine afleveringer.
- For hvert projekt vælges ét af fire emner. Man må gerne (men behøver ikke) vælge samme emne til begge projekter.
- Se <https://02402.compute.dtu.dk/projects>

# Eksamensplan

## Eksamensplan

- Lørdag den 20. december 2024. Bemærk særlig dato!
- 4 timers multiple choice-prøve (30 spørgsmål).
- Alle skriftlige hjælpemidler + lommeregner (NYT)
- Re-eksamen i maj (bemærk særlig dato)

# Generel ugeseddel

- Før undervisningen: Læs de relevante afsnit bogen. Forsøg at løse ugens opgaver.
- Forelæsninger: Gennemgang af ugens pensum
- Øvelser: Få hjælp til opgaver. Brug TA'erne alt hvad du kan
- Efter undervisningen: test-quizzler på hjemmesiden

# Hvor finder jeg information?

- Kursushjemmeside: 02402.compute.dtu.dk
  - Generel info (læs grundigt!)
  - Agenda med undervisningsplan, slides (+ Python filer), øvelser, quiz
  - Bog (vent med at printe til opdateringer er fuldført)
  - Information om projekter
  - Tidligere års eksamenssæt, optagede forelæsninger
- DTU Learn
  - Announcements
  - Projekter - formulering og aflevering
  - Video & Streaming - livestream af forelæsninger
- Ed Discussion
  - <https://edstem.org/eu/join/YJ2hDS>
  - Forum for spørgsmål og diskussioner



# Python

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

Vi starter med at bruge Python fra næste uge (kursusuge 2).

Der kommer en announcement ud senere på ugen.



Visual Studio Code



Dem der har lyst kan få adgang til [www.DataCamp.com](http://www.DataCamp.com), hvor man (bl.a.) kan få en introduction til Python, med hands on øvelser.

Brugen af DataCamp er frivillig.

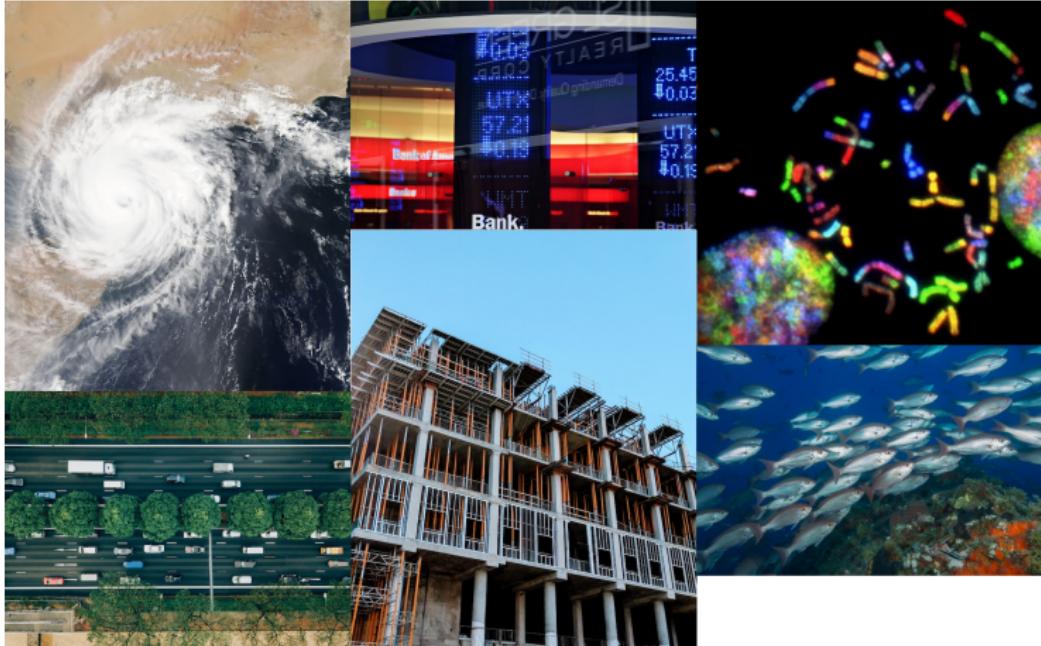
Information om adgang/oprettelse kommer i announcement.

# 02402 Statistik (Polyteknisk grundlag)

## Introduktion

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Statistik anvendes alle vejne!



# Grundlæggende om statistik

Statistik har at gøre med at indsamle og analysere **data**

Statistik kan generelt opdeles i to dele:

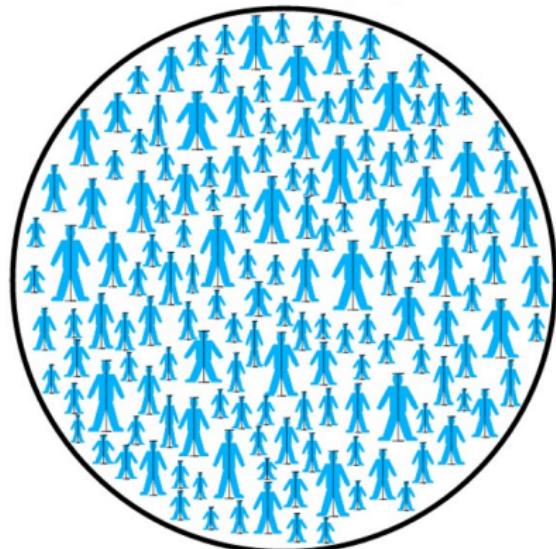
- Deskriptiv statistik
- Konkluderende statistik (Statistisk inferens)

For at lave statistisk inferens har vi brug for **teoretiske modeller**.

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

# Populationen og stikprøven

(Infinite) Statistical population

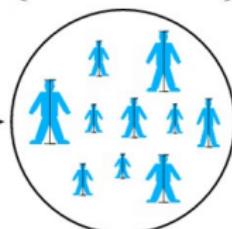


Mean  
 $\mu$

Statistical  
Inference

Sample  
 $\{x_1, x_2, \dots, x_n\}$

Randomly  
selected



Sample mean  
 $\bar{x}$

# Grundlæggende begreber

## ☰ Definition 1.1 Sample and population

- An *observational unit* is the single entity about which information is sought (e.g. a person)
- An *observational variable* is a property which can be measured on the observational unit (e.g. the height of a person)
- The *statistical population* consists of the value of the observational variable for all observational units (e.g. the heights of all people in Denmark)
- The *sample* is a subset of the statistical population, which has been chosen to represent the population (e.g. the heights of 20 persons in Denmark).

# Populationen

Ordet "Population" bruges i statistik på en lidt anden måde end man bruger det i dagligdags tale.

En statistisk population behøver ikke være personer - det kan også være en population af genstande eller tilfælde (events).

Den statistiske population kan være konkret defineret (fx "alle studerende på DTU i hele 2025") eller den kan være mere abstrakt og evt uendelig stor (fx "alle tænkelige tilfælde af hundegalskab").

*Det er ikke altid klart defineret hvad populationen er, men det er altid en god ide at prøve at definere populationen.*

# En repræsentativ stikprøve

Stikprøven vil altid være klart defineret, da det er denne vores konkrete **data** omhandler.

Når vi laver *statistisk inferens* udtaler vi os generelt om populationen ud fra de observationer vi har i stikprøven. Vi *antager* altså at de tendenser vi ser i stikprøven kan overføres til populationen.

Det er derfor vigtigt at stikprøven er *repræsentativ* for populationen.

En stikprøve er repræsentativ når den er udtrukket fuldstændig tilfældigt - dvs alle enheder i populationen har *samme* sandsynlighed for at komme med i stikprøven.

*I langt det meste af kurset vil vi bare antage, at stikprøverne er repræsentative.*

# Et eksempel med stikprøve og population

Lad os sige vi ønsker at vide noget om trivsel blandt studerende på DTU. Vi laver derfor et spørgeskema og uddeler dette til de 54 studerende der dukker op til kurset "Diskret Matematik" torsdag morgen (dette kursus ligger på 1. semester for uddannelsen Softwareteknologi). Spørgeskemaet indeholder en række spørgsmål om både faglig og social trivsel. Ikke alle studerende har lyst til at deltage, men vi modtager besvarelser fra 42 studerende.

Spørgsmål:

Hvad er den statistiske population?

Er stikprøven repræsentativ for den population vi ønsker at undersøge?

Hvordan kunne vi have lavet en bedre stikprøve?

---

---

(Skriv dine egne noter her)

# 02402 Statistik (Polyteknisk grundlag)

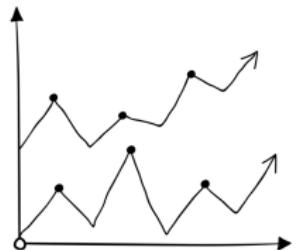
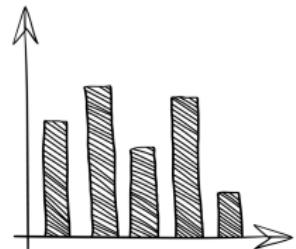
## Deskriptiv Statistik

- Statistiske nøgletal
- Data visualisering

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Deskriptiv statistik

At beskrive og sammenfatte **DATA.**



# Forskellige typer af data

- Numeriske data

- Kvantitavite værdier målt på en skala (cm, kr, ..)
- Diskrete værdier (0/1, antal)

- Kvalitative/Kategoriske data

- Grupperinger (mand/kvinde, før/efter, rød/gul/blå)
- Evt rangerede grupper (lidt/mellem/meget, "i lav grad" /"i høj grad" )

Kvalitative data vil man ofte "oversætte" til binære værdier med 0/1 (med såkaldte "dummy variable").

# Eksempel: Stikprøve på 20 observationer

En stikprøve med 20 observationer har følgende værdier:

41, 44, 37, 38, 36, 39, 38, 47, 32, 43, 36, 39, 49, 35, 40, 35, 45, 36, 44, 38

Her er stikprøvens værdier i sorteret rækkefølge:

32, 35, 35, 36, 36, 36, 36, 37, 38, 38, 38, 38, 39, 39, 40, 41, 43, 44, 44, 45, 47, 49

Hvordan kan man beskrive stikprøven?

---

---

---

(Skriv dine egne noter her)

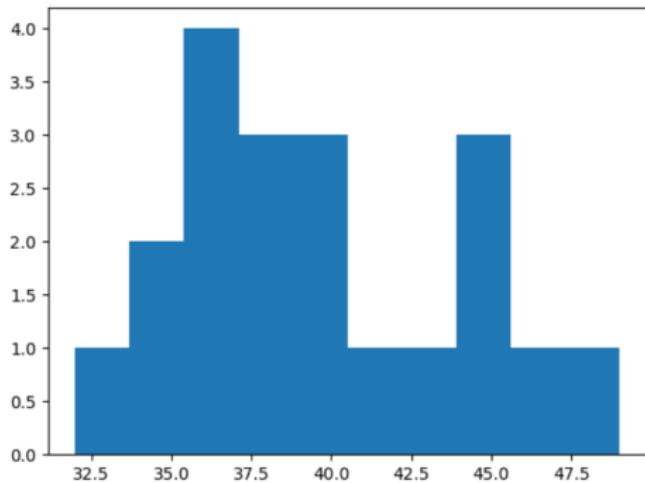
---

---

---

## Eksempel: visualisering af data

Vi kan også visualisere stikprøven. Fx med et histogram:



32, 35, 35, 36, 36, 36, 36, 37, 38, 38, 38, 38, 39, 39, 39, 40, 41, 43, 44, 44, 45, 47, 49

Hvad viser histogrammet egentlig?

**Kahoot!**  
(x1)

# 02402 Statistik (Polyteknisk grundlag)

- Statistiske nøgletal
- Data visualisering

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Nøgletal (eng: *Summary Statistics*)

Nøgletal bruges til at opsummere og beskrive data.

- *Mål for center*
  - f.eks.: Gennemsnit og Median
- *Mål for spredning*
  - f.eks.: Range, IQR, Varians, Standardafvigelse
- *Mål for sammenhænge*
  - f.eks.: Kovarians og Korrelation

# Gennemsnit (også kaldet middelværdi) (eng: *Mean* eller *Average*)

Antag vi har en stikprøve med i alt  $n$  observationer:

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, \dots, x_n\}$$

## III Definition 1.4 Sample mean

The sample mean is the sum of observations divided by the number of observations

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

Sometimes this is referred to as the *average*.

32 35 35 36 36 36 37 38 38 38 39 39 40 41 43 44 44 45 47 49

$$\bar{x} = 39.6 \quad (n = 20)$$

# Median (eng: *Median*)

## |||| Definition 1.5 Median

Order the  $n$  observations  $x_1, \dots, x_n$  from the smallest to largest:  $x_{(1)}, \dots, x_{(n)}$ . The median is defined as:

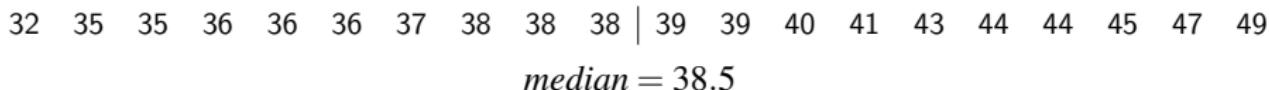
- If  $n$  is odd the median is the observation in position  $\frac{n+1}{2}$ :

$$Q_2 = x_{\left(\frac{n+1}{2}\right)}. \quad (1-2)$$

- If  $n$  is even the median is the average of the two observations in positions  $\frac{n}{2}$  and  $\frac{n+2}{2}$ :

$$Q_2 = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}. \quad (1-3)$$

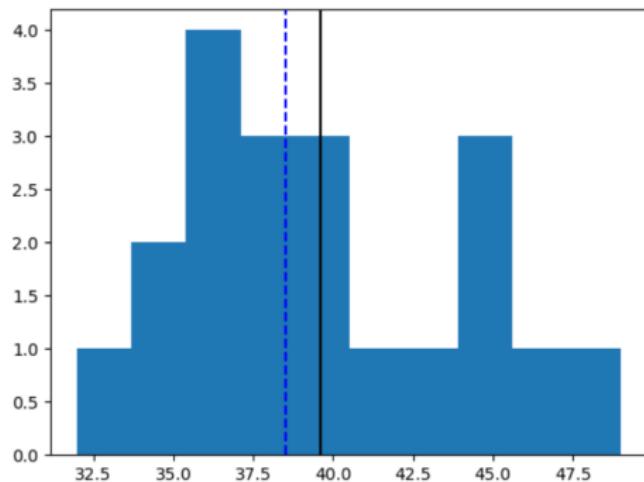
The reason why it is denoted with  $Q_2$  is explained below in Definition 1.8.



# Gennemsnit og Median

Gennemsnit og median er ikke det samme (men de vil ofte ligge tæt på hinanden).

Her ses eksempel fra tidligere samt gennemsnit (sort) og median (blå):



# Medianen - og andre *fraktiler*

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* (eng: **Quartiles**), betegnet med hhv.  $Q_1$ ,  $Q_2$  og  $Q_3$ .

32	35	35	36	36		36	37	38	38	38		39	39	40	41	43		44	44	45	47	49
----	----	----	----	----	--	----	----	----	----	----	--	----	----	----	----	----	--	----	----	----	----	----

# Fraktiler (eng: *Quantiles* eller *Percentiles*)

## ||| Definition 1.7 Quantiles and percentiles

The  $p$  quantile also called the  $100p\%$  quantile or  $100p$ 'th percentile, can be defined by the following procedure:<sup>a</sup>

1. Order the  $n$  observations from smallest to largest:  $x_{(1)}, \dots, x_{(n)}$
2. Compute  $pn$
3. If  $pn$  is an integer: average the  $pn$ 'th and  $(pn + 1)$ 'th ordered observations. Then the  $p$  quantile is

$$q_p = \left( x_{(np)} + x_{(np+1)} \right) / 2 \quad (1-4)$$

4. If  $pn$  is a non-integer: take the “next one” in the ordered list. Then the  $p$ 'th quantile is

$$q_p = x_{(\lceil np \rceil)}, \quad (1-5)$$

where  $\lceil np \rceil$  is the *ceiling* of  $np$ , that is, the smallest integer larger than  $np$

# Fraktiler: Eksempel

50% fraktilen (medianen) illustreret i eksemplet fra tidligere:

32 35 35 36 36 36 37 38 38 | 39 39 40 41 43 44 44 45 47 49

Dvs Medianen her er 38.5.

10% fraktilen illustreret:

32 35 | 35 36 36 37 38 38 38 39 39 40 41 43 44 44 45 47 49

Dvs 10% fraktilen her er 35.

Ønskes en procentdel der rammer "skævt" ift. antallet af observationer i data, benyttes nærmeste værdi. Se formel i forrige slide.

Eksempelvis er 33% fraktilen 37.

# Range og IQR

## |||| Definition 1.15 Range

The *range* of the sample is

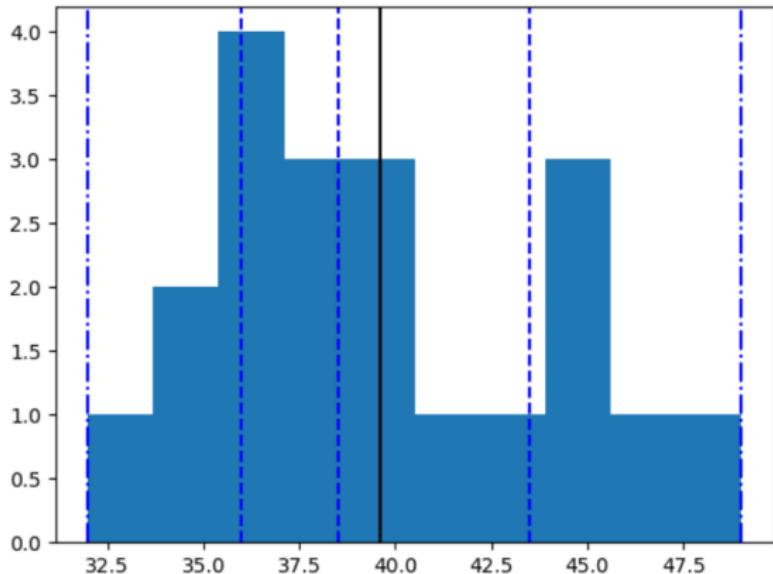
$$\text{Range} = \text{Maximum} - \text{Minimum} = Q_4 - Q_0 = x_{(n)} - x_{(1)}. \quad (1-9)$$

The Inter Quartile Range (IQR) is the middle 50% range of data defined as

$$IQR = q_{0.75} - q_{0.25} = Q_3 - Q_1. \quad (1-10)$$

OBS: Minimum og Maximum kaldes her også for Q0 og Q4

# Visualisering af min, Q1, Q2, gennemsnit, Q3 og max



# Stikprøve varians (eng: *Sample variance*)

## ||| Definition 1.10 Sample variance

The *sample variance* of a sample  $x_1, \dots, x_n$  is the sum of squared differences from the sample mean divided by  $n - 1$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1-6)$$

# Stikprøve standardafvigelse (eng: *sample standard deviation*)

## ||| Definition 1.11 Sample standard deviation

The *sample standard deviation* is the square root of the sample variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (1-7)$$

Nogle gange bruges ordet *spredning* i stedet for *standardafvigelse*.

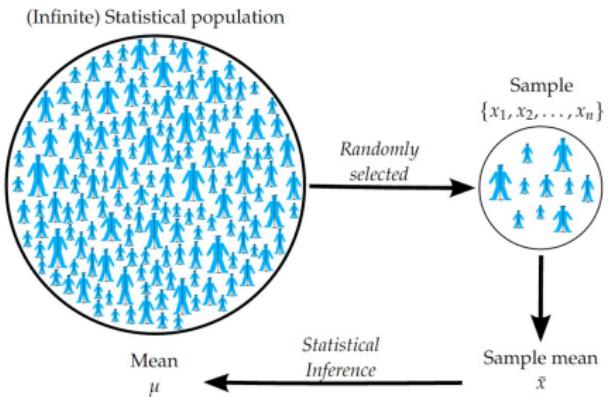
# Hvorfor $n - 1$ ?

## Hvorfor $n - 1$ ?

Når vi definerer "Stikprøve varians" og "Stikprøve standardafvigelse" som vi gør, er det for at få det bedste estimat for populationens varians og standardafvigelse.

Derfor divideres med  $(n - 1)$  i stedet for  $n$ .

Pas på, der findes forskellige definitioner derude! Dette er dog den mest almindelige definition af "Stikprøve varians" og "Stikprøve standardafvigelse".



# Variationskoefficient (eng: *coefficient of variation*)

## |||| Definition 1.12 Coefficient of variation

The *coefficient of variation* is the sample standard deviation seen relative to the sample mean

$$CV = \frac{s}{\bar{x}}. \quad (1-8)$$

CV er et *relativt* mål for variation.

# Eksempel med spredning

- **Stikprøven:**

32, 35, 35, 36, 36, 36, 37, 38, 38, 38, 39, 39, 40, 41, 43, 44, 44, 45, 47, 49

Der er i alt 20 observationer i stikprøven;  $n = 20$ .

- **Stikprøvegennemsnit:**

$$\bar{x} = \frac{1}{20}(32 + 35 + 35 + 36 + \dots + 49) = 39.59$$

- **Stikprøve varians:**

$$s^2 = \frac{1}{19}((32 - 39.59)^2 + (35 - 39.59)^2 + \dots + (49 - 39.59)^2) = 20.52$$

- **Stikprøve standardafvigelse:**

$$s = \sqrt{20.52} = 4.59$$



# Mere komplekse data

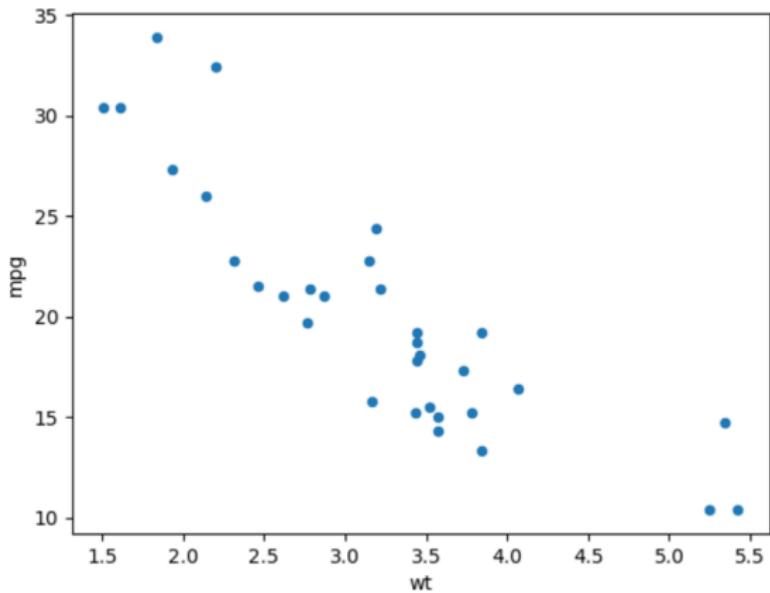
Vi kan forestille os mere komplekse data - i stedet for én række med tal, kan vi have et datasæt hvor der er flere forskellige oplysninger for hver observation:

model	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21	6	160	110	3.9	2.62
Mazda RX4 Wag	21	6	160	110	3.9	2.875
Datsun 710	22.8	4	108	93	3.85	2.32
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.44
Valiant	18.1	6	225	105	2.76	3.46

Her ses et eksempel med data for nogle forskellige biler.

# Mål for sammenhænge

Vi kan undersøge sammenhænge i data - fx mellem informationerne i kolonnerne *mpg* og *wt*:



# Stikprøve kovarians

Vi kalder nu værdierne i den ene kolonne for  $\{x_1, x_2, x_3, \dots, x_n\}$  og værdierne i den anden kolonne for  $\{y_1, y_2, y_3, \dots, y_n\}$ .

## ||| Definition 1.18 Sample covariance

The sample covariance is

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \quad (1-11)$$

# Stikprøve korrelation

## ||| Definition 1.19 Sample correlation

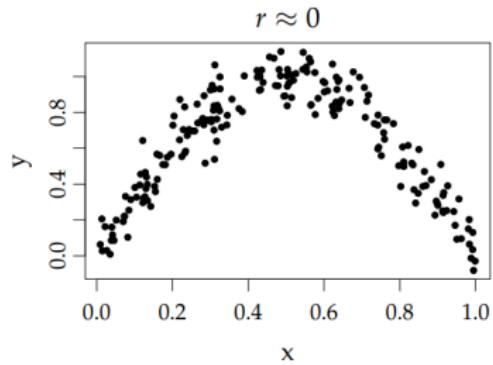
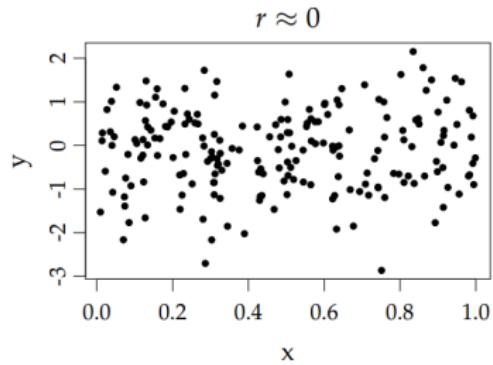
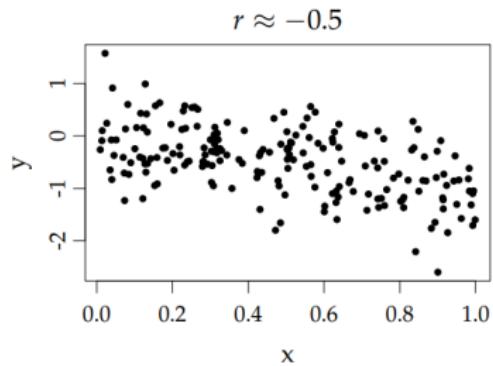
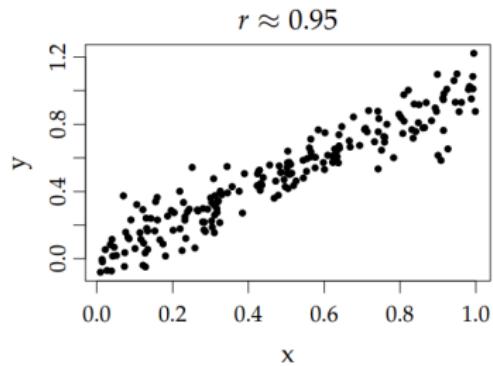
The sample correlation coefficient is

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}, \quad (1-12)$$

where  $s_x$  and  $s_y$  is the sample standard deviation for  $x$  and  $y$  respectively.

Korrelationen  $r$  er et tal mellem  $+1$  og  $-1$ .

# Stikprøve korrelation



# 02402 Statistik (Polyteknisk grundlag)

- Statistiske nøgletal
- Data visualisering

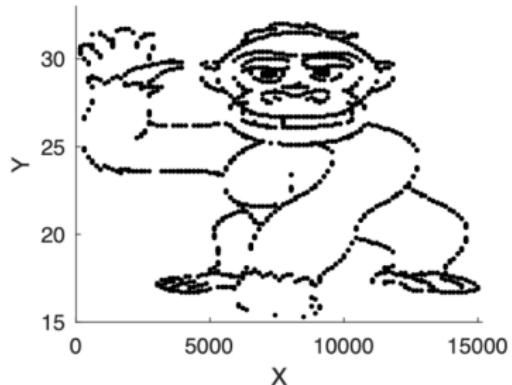
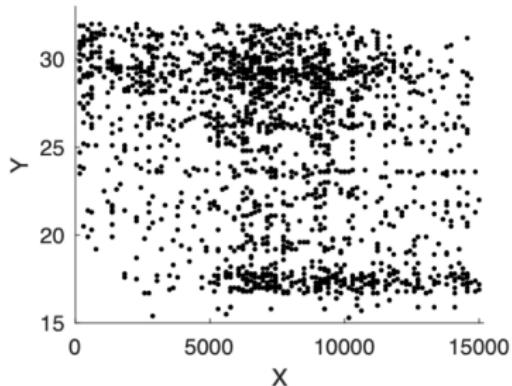
DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Data visualisering

Data visualisering er vigtigt!

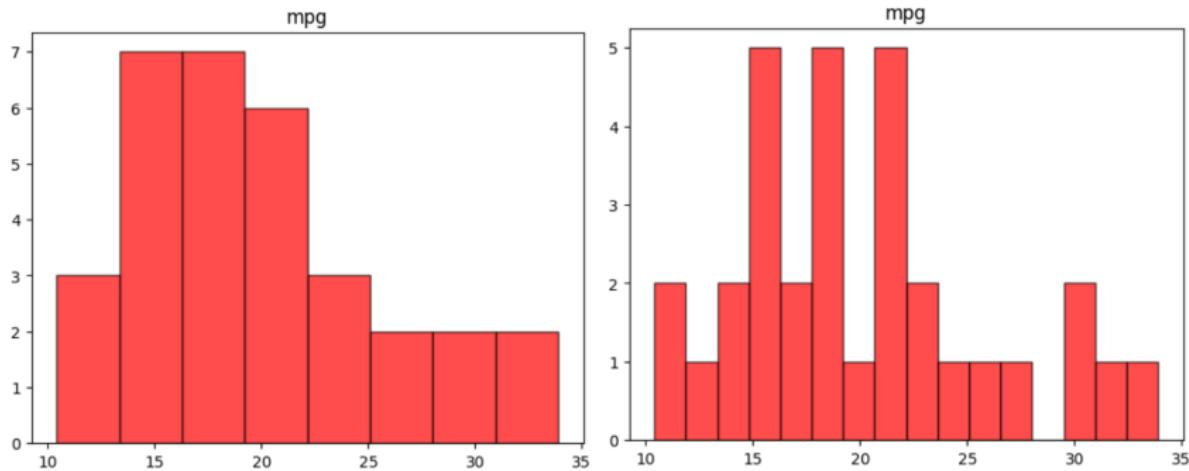
Mange slags plots:

- Histogram
- Boxplot
- Kumuleret fordeling (ecdf)
- Scatterplot (xy plot)
- Søjlediagram (bar chart)
- Cirkeldiagram (pie chart)



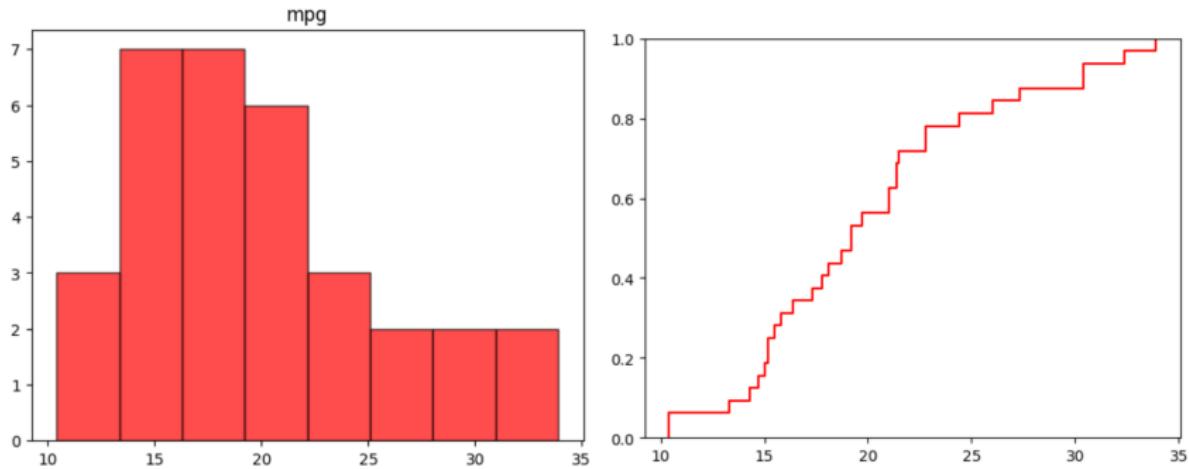
# Histogram

Her ses to histogrammer for de samme data (kolumnen mpg i datasættet for biler). Forskellen på de to histogrammer er udelukkende valget af "bin size".



# Kumuleret fordeling (ecdf)

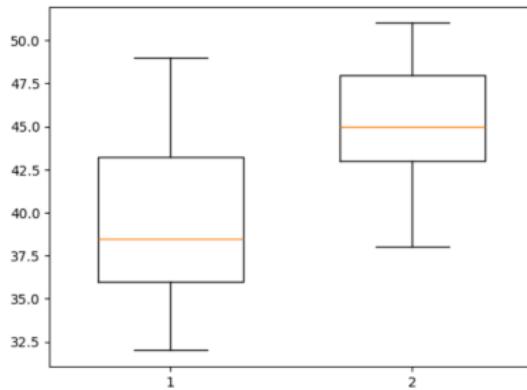
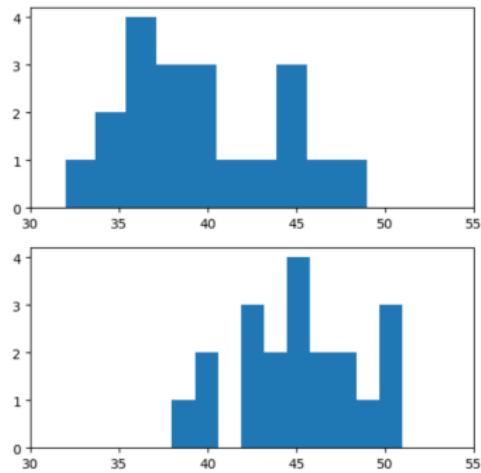
I stedet for et histogram kan man visualisere stikprøven med et kumuleret plot (også kaldet ecdf - *empirical cumulated distribution function*), hvilket har den fordel at det ikke afhænger af "bin size"



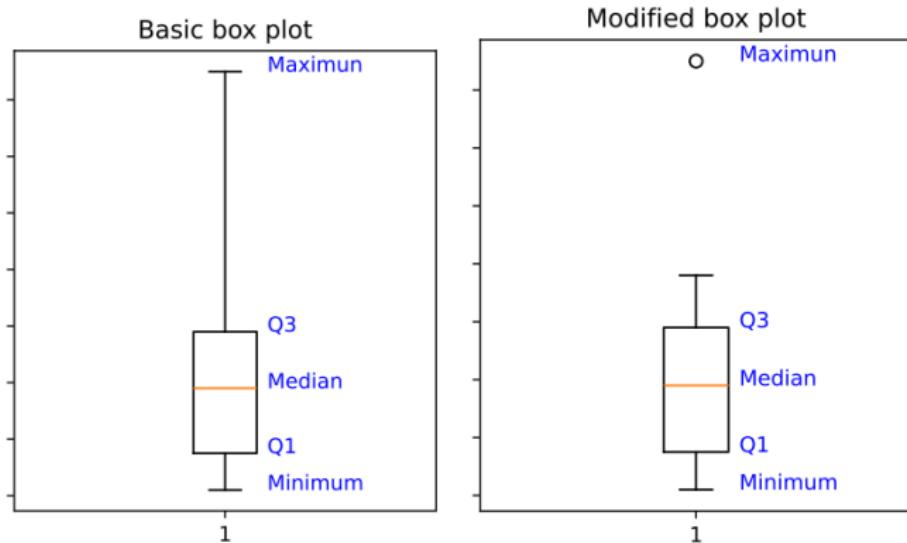
# Boxplot

Til sammenligning af flere stikprøver bruges ofte boxplot, som indeholder færre detaljer end histogrammer.

Boxplottene er (ofte) tegnet med værdierne op ad y-aksen i stedet for hen ad x-aksen (som på histogrammerne).



# Hvordan tegnes et boxplot?

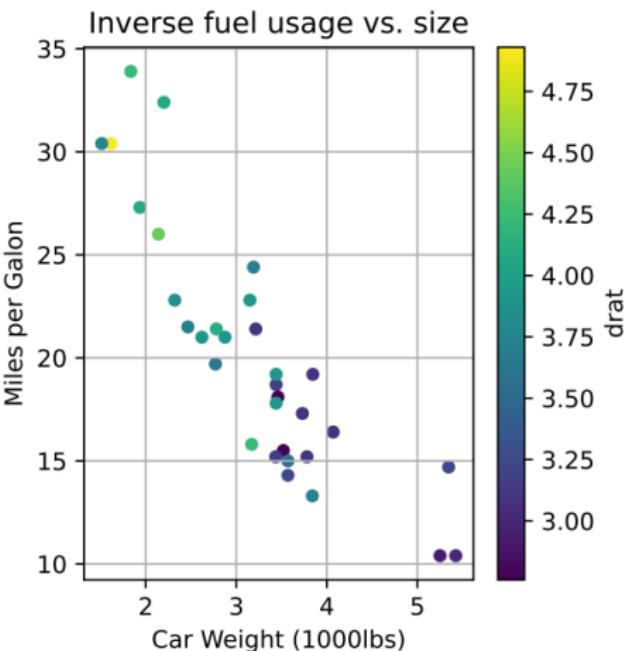
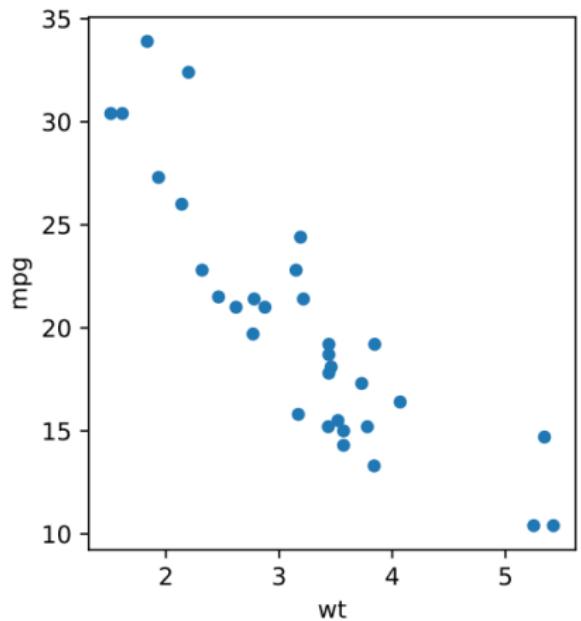


"Whiskers" kan tegnes på forskellige måder. Bogen har 2 definitioner.

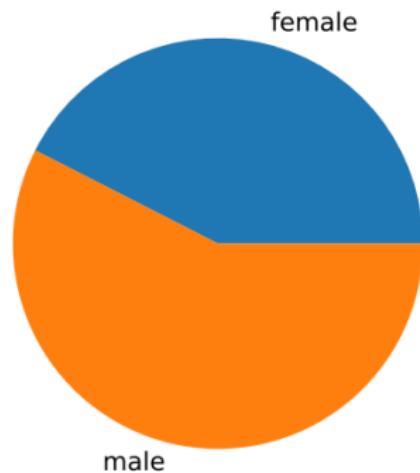
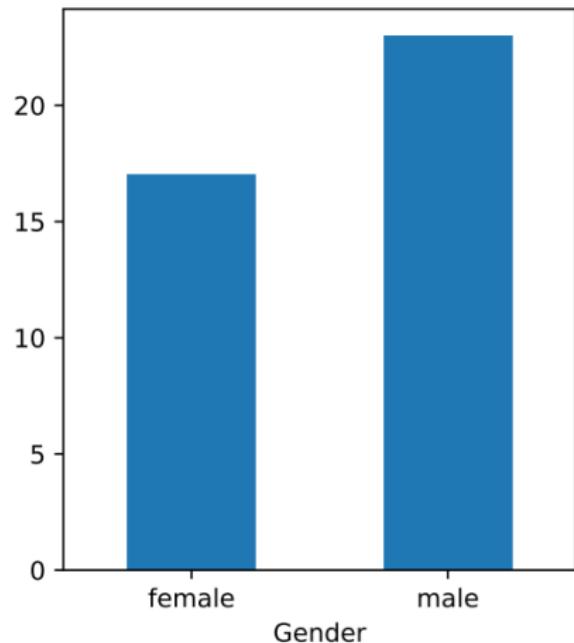
"Basic boxplot": Whiskers tegnes helt ud til maximum og minimum.

"Modified boxplot": Whiskers tegnes maximalt ud til  $1.5 \cdot IQR$  (fra Q1/Q3). Værdier udenfor dette tegnes individuelt og kaldes "outliers" eller "ekstreme" værdier.

# Scatterplot (xy plot)



# Søjle- og cirkeldiagrammer



# 02402 Statistik (Polyteknisk grundlag)

## Stokastiske variable

- Sandsynligheder, pdf, cdf og forventningsværdier

# Stokastisk variabel (eng: *Random* eller *Stochastic* variable)

En stokastisk variabel  $Y$  repræsenterer værdien af udfaldet **før** det tilhørende *eksperiment* finder sted.

- Kast med mønt
- Antallet af seksere ved kast med 5 terninger
- Benzinforsbrug af en bil
- Måling af blodsukker
- ...

"Eksperimentet" skal altså forstås meget bredt.

Vi kan også kalde det "*en data genererende process*".

# Udfaldsrum (eng: *Sample space*)

## ||| Definition 2.1

The *sample space*  $S$  is the set of all possible outcomes of an experiment.

## ||| Definition 2.3

A *random variable* is a function which assigns a numerical value to each outcome in the sample space. In this book random variables are denoted with capital letters, e.g.

$$X, Y, \dots \quad (2-2)$$

# Eksempel på en diskret stokastisk variabel

Husk: En stokastisk variabel er en funktion der *tildeler en numerisk værdi* til et udfald.

Eksempel: Kast med en mønt



Udfaldsrummet er: "Plat" eller "Krone"

Vi tildeler **numeriske værdier**: 0 eller 1

# Endnu et eksempel på en diskret stokastisk variabel

Eksempel: Antal 6'ere ved kast med fem terninger



Udfaldsrum:

---

Numeriske værdier:

---

# Eksempel på en kontinuær stokastisk variabel

Eksempel: Benzinforbrug af bil, målt i km/L



Udfaldsrum:

---

Numeriske værdier:

---

# 02402 Statistik (Polyteknisk grundlag)

- Sandsynligheder, pdf, cdf og forventningsværdier

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

# Sandsynligheder for diskrete stokastiske variable

# Sandsynligheder - diskrete tilfælde

## Hvad er sandsynligheden for et bestemt udfald?

Lad den stokastiske variabel  $Y$  beskrive udfaldet af et kast med en mønt.

Vi giver værdierne  $Y = 1$  for krone og  $Y = 0$  for plat.

**Notation:** Sandsynligheden for at få krone skrives:  $P(Y = 1)$ .

Hvad er  $P(Y = 1)$  og  $P(Y = 0)$  ?

Lad den stokastiske variabel  $Z$  beskrive udfaldet (dvs antal øjne) af et kast med én terning.

Hvad er  $P(Z = 2)$  ?

For at beskrive en stokastisk variabel, må vi kende både **udfaltsrum** og tilsvarende **sandsynligheder**.

# Tæthedsfunktionen (diskrete tilfælde) (eng: *Probability density function, pdf*)

## |||| Definition 2.6 The *pdf* of a discrete random variable

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

$$f(x) = P(X = x). \quad (2-9)$$

It assigns a probability to every possible outcome value  $x$ .

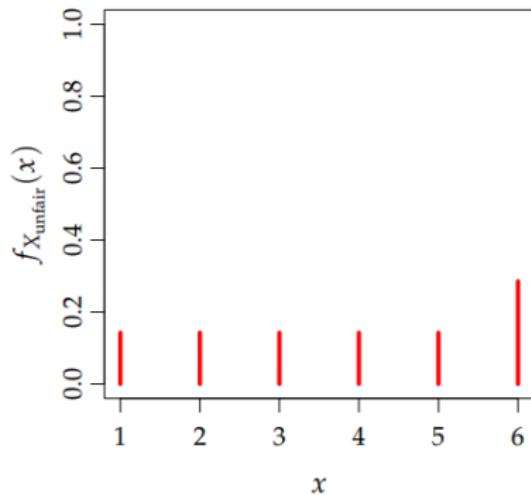
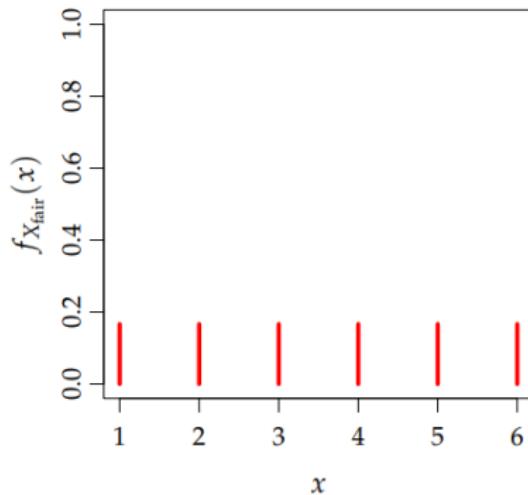
A discrete *pdf* fulfils two properties: there are no negative probabilities for any outcome value

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

and the probabilities for all outcome values sum to one

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

# Eksempel på tæthedsfunktioner



Her ses eksempler på tæthedsfunktioner for en terning. Til venstre ses en lige terning og til højre ses en ulige terning, hvor sandsynligheden for at få en 6'er er dobbelt så høj som sandsynligheden for at få en af de andre værdier.

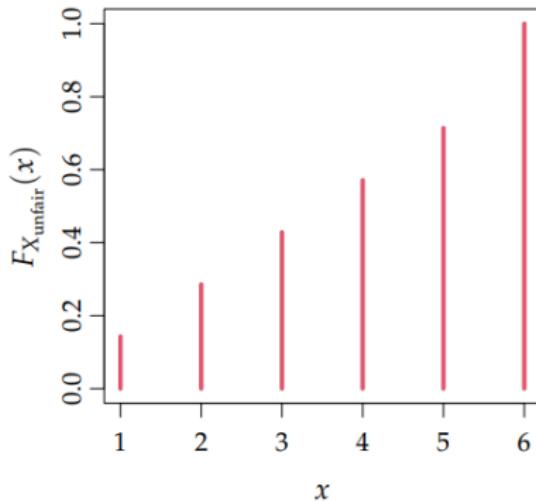
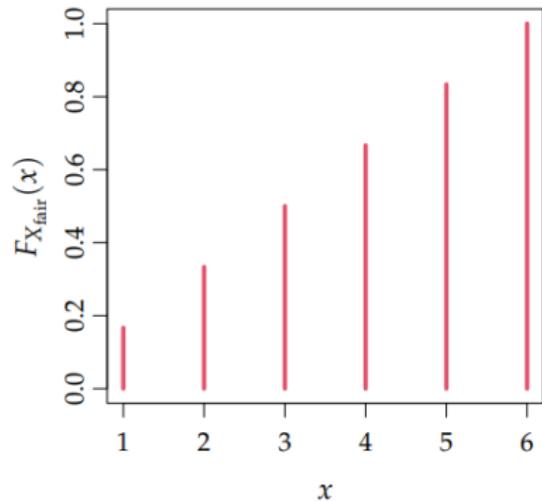
# Fordelingsfunktionen (diskrete tilfælde) (eng: *Cumulative distribution function, cdf*)

## |||| 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  $x$

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

# Eksempel på fordelingssfunktioner



Her ses eksempler på fordelingsfunktioner for en lige og en skæv terning.

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

# Eksempel med diskret stokastisk variabel

En stokastisk variabel  $X$  har følgende udfaldsrum og fordeling:

$x$	1	2	3	4	5
$f(x)$	0.1	0.2	0.4	0.2	0.1

Hvad er sandsynligheden for at  $X$  er mindre end eller lig med 3;  $P(X \leq 3)$ ?

---

Hvad er sandsynligheden for at  $X$  er 100;  $P(X = 100)$ ?

---

Tegn pdf( $x$ ) og cdf( $x$ ):

# Forventningsværdi (diskrete tilfælde) (eng: *Expectation value*)

## ||| Definition 2.13 Mean value

The mean of a discrete random variable X is

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

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

Den teoretiske forventningsværdi er det samme som det teoretiske gennemsnit.

# Varians (som forventningsværdi)

## ||| Definition 2.16 Variance

The variance of a discrete random variable X is

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

where  $x_i$  is the outcome value and  $f(x_i)$  is the *pdf* of the  $i$ th outcome value.  
The *standard deviation*  $\sigma$  is the square root of the variance.

Den teoretiske varians kan formuleres som en forventningsværdi af størrelsen  $(X - \mu)^2$ .

tf

## Eksempel med diskret stokastisk variabel

En stokastisk variabel  $Y$  har følgende udfaldsrum og fordeling:

$y$	-10	0	10
$f(y)$	0.30	0.40	0.30

Hvad er sandsynligheden for at  $Y$  ikke er nul;  $P(Y \neq 0)$ ?

---

Tegn pdf( $y$ ) og cdf( $y$ ):



Hvad er  $\mathbf{E}[Y]$  og  $\mathbf{V}[Y]$ ?

---

---

# Sandsynligheder for kontinuærte stokastiske variable

# Sandsynligheder - kontinuære tilfælde

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

Overvej eksemplet med Benzinforbrug af bil.

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

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

---

---

---

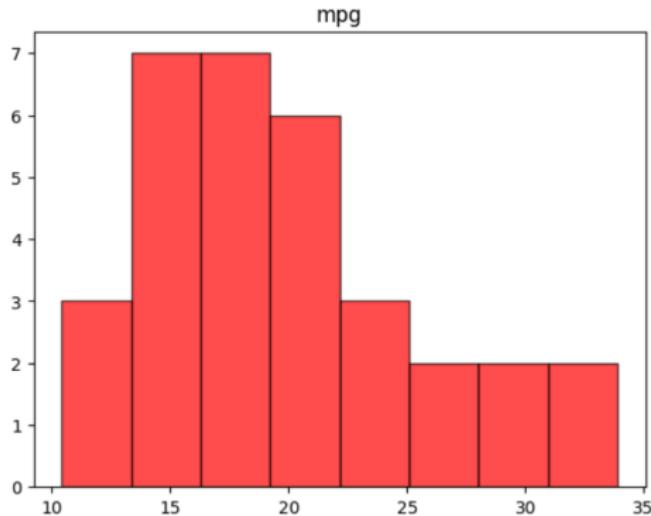
---

(Skriv dine overvejelser her)

# Sandsynligheder - kontinuære tilfælde

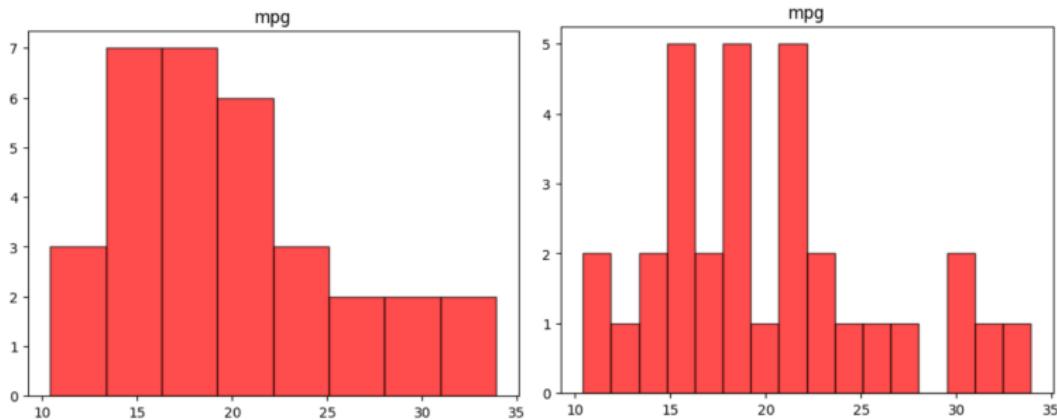
Vi har brug for mere information om den stokastiske variabel for at kunne svare på spørgsmålene.

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



# Sandsynligheder - kontinuære tilfælde

Histogrammet kan også vises med mindre "bins":



Histogrammerne illustrerer rigtig **data** (en stikprøve).

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

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

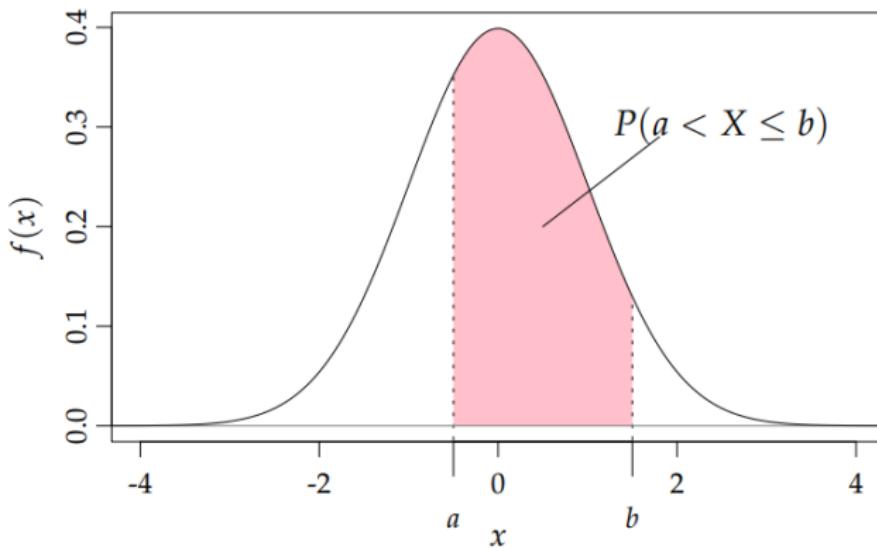
and has an area below the function of one

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

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:

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

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

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

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

## ||| Definition 2.33 Distribution

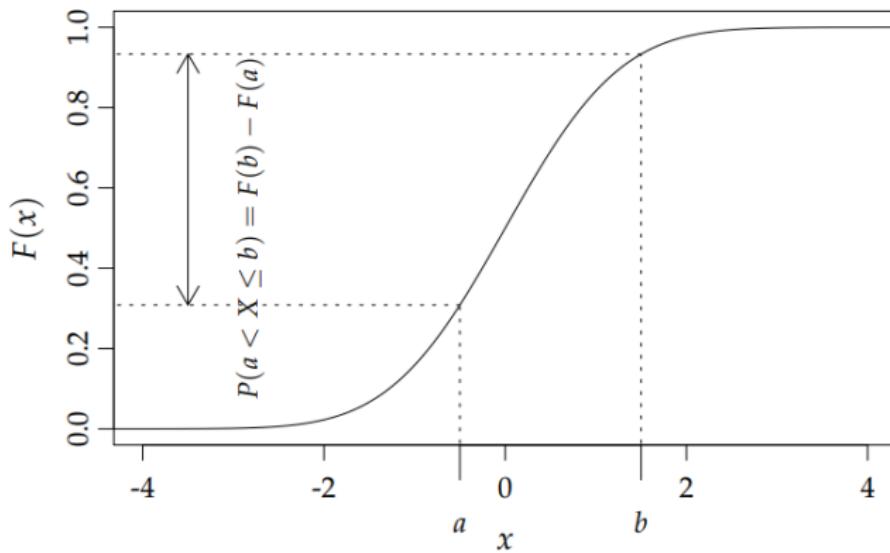
The *cdf* of a continuous variable is defined by

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

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 fordelingssfunktioner



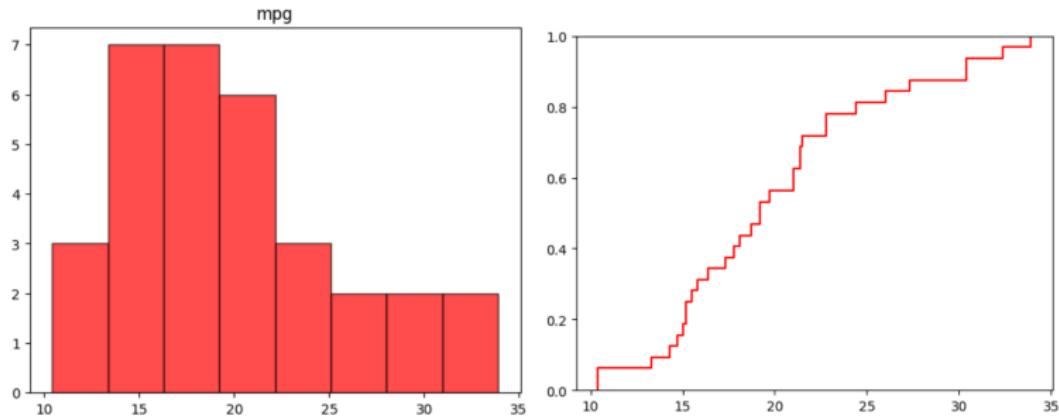
Fordelingsfunktionen er en alternativ måde at visualisere sandsyndlighedsfordelingen af udfaldene på.

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:



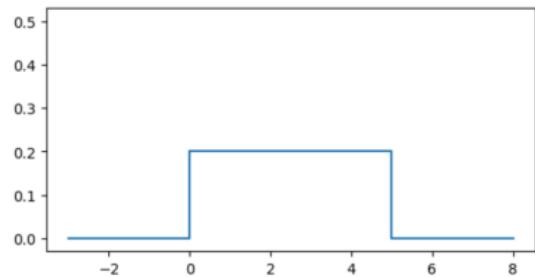
# Eksempel med kontinuær stokastisk variabel

En stokastisk variabel  $Z$  følger en tæt-hedsfordeling  $f(z)$ .

$f(z)$  er visualiseret i plottet til højre.

Hvad er sandsynligheden for at  $Z$  er mindre end eller lig med 3;  $P(Z \leq 3)$ ?

---



Hvad er sandsynligheden for at  $Z$  er 100;  $P(Z = 100)$ ?

---

Tegn cdf( $z$ ):

# Forventningsværdi (kontinuære tilfælde) (eng: *Expectation value*)

## ||| 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 med kontinuær stokastisk variabel

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

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

Tegn pdf( $x$ ):



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

---

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

---

---

# 02402 Statistik (Polyteknisk grundlag)

# Husk før næste uge

DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

Husk at have Python installeret inden næste gang!

<https://pythonsupport.dtu.dk/>

- Fra næste uge bruger vi Python !
- Vi bruger 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**  
(KID studerende vil få hjælp i morgen/onsdag)
- Hjælp til installering (og mere): <https://pythonsupport.dtu.dk/>

# Tjekliste

Efter i dag skal du kunne:

- DATA: Beregne diverse nøgletal for en stikprøve: gennemsnit, median, fraktiler (herunder kvartiler), IQR, stikprøve-variанс, stikprøve-standardafvigelse, CV, stikprøve-kovarians og -korrelation.
- DATA: Tegne og forstå/afkode histogrammer, boxplots, ecdf, scatterplot, søjle- og cirkel diagrammer.
- TEORI: Forklare hvordan man kan bruge en stokastisk variabel som en teoretisk beskrivelse af en *data genererende process* (et "eksperiment").
- TEORI: Beregne diverse teoretiske størrelser for en stokastisk variabel, herunder sandsynligheder, forventningsværdier og størrelser relateret til pdf og cdf.

# Øvelser

## Tirsdag 10-12

- 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 Cyril
  - 040. Hjælpelærer: Phd-student Thea
  - 050. Hjælpelærer: Jakob
  - 020. Hjælpelærer: Drin
  - 030. Hjælpelærer: Maliha
  - Man kan også sidde i foyerområde 003, 004, 005 og 008.