Course 02403: Introduktion til matematisk statistik

Forelæsning 1: Intro, R og beskrivende statistik

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

Oversigt

Praktisk Information

- Undervisning: Hver dag 08-12
- Generel daglig agenda:
  - FØR undervisningsmodulet: læs det annoncerede i eNoten!
  - 2x45 minutters forelæsning (dagens pensum)
  - 2 timers øvelser: Enote Excersises
- Skriftlig eksamen: Sidste dag i 3-ugers (26/6).
- OBLIGATORISK projekt (Individuelt): 1 stk - skal godkendes for at kunne gå til eksamen.

Praktisk Information

- Homepage:https://02403.compute.dtu.dk/
  - Læsemateriale: eNoter
  - Forelæsningsplan
  - Slides
  - Projekt
  - Podcast (2021/2022 forelæsning)

  Learn: `https://learn.dtu.dk`
  - Meddelelser
  - Projekt - beskrivelse OG aflevering
  - Quizzer (ikke fuldstændig alignment med kurset)

- Ekstra materiale
  - 02402.compute.dtu.dk (bliver ikke vedligeholdt under kurset og afspejler ikke pensum til dette kursus 100%).
  - Læsemateriale: eNoter
  - Podcasts af gl. forelæsninger (02402) (På dansk OG engelsk)

Area9

- Adaptive læringsplatform
- God forberedelse til eksamen
- Hjælper med at træne de rigtige ting..
- "Challange us"

Introduktion til Matematisk Statistik

Forskellen mellem "Introduktion til Matematisk Statistik"(02403) og "Introduktion til Statistik"(02402/02323) er generelt lille, men et par punkter er at vi:

- Gennemgår en (lille) del af det sandsynlighedsteoretiske grundlag for statistik.
- Gennemfører (små) beviser, baseret på relevante antagelser.
- Lægger lidt mindre vægt på konkrete eksempler og fordelinger.

Oversigt

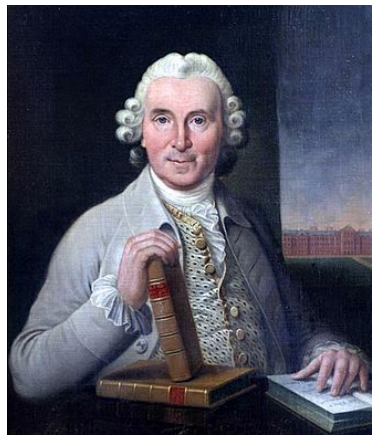
Millennium list (EDITORIAL: Looking Back on the Millennium in Medicine) [1]

- Elucidation of Human Anatomy and Physiology
- Discovery of Cells and Their Substructures
- Elucidation of the Chemistry of Life
- **Application of Statistics to Medicine**
- Development of Anesthesia
- Discovery of the Relation of Microbes to Disease
- Elucidation of Inheritance and Genetics
- Knowledge of the Immune System
- Development of Body Imaging
- Discovery of Antimicrobial Agents
- Development of Molecular Pharmacotherapy

[1] *N Engl J Med*, 342:42-49, January 6, 2000.
http://www.nejm.org/doi/full/10.1056/NEJM200001063420108

James Lind

"*One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy.*"(See also `http://en.wikipedia.org/wiki/James_Lind`).



Man kan altså undersøge fænomener man ikke forstår og derefter begynde at forstå dem!

John Snow

"*The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well.*"(See also `http://en.wikipedia.org/wiki/John_Snow_(physician)`).

Google - *Big Data*

A quote from New York Times, 5. August 2009, from the article titled "For Today's Graduate, Just One Word: Statistics" is:

"*I keep saying that the sexy job in the next 10 years will be statisticians,*"*said Hal Varian, chief economist at Google.* "*And I'm not kidding.*

IBM - *Big Data*

"*The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd,*"*said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment.* "*And that makes it easier for humans to do what they are good at - explain those anomalies.*"

Oversigt

Intro Case stories

- Senior Scientist Hanne Refsgaard, Novo Nordisk A/S
- IBM Social media by Henrik H. Eliasen, IBM
- Skive Fjord podcasts by Jan K. Møller, DTU

Oversigt

Introduktion til Statistik

- Hvordan behandles (eller analyseres) data?
- Hvad er tilfældig variation?
- Statistik er et værktøj til at træffe beslutninger:
  - Hvor mange computere har vi solgt det sidste år?
  - Styring af energisystemer med fluktuerende vedvarende energi?
  - Er maskine A mere effektiv end maskine B ?
- Statistik er et metodefag, der kan anvendes inden for de fleste fagområder, og er derfor et meget vigtigt værktøj

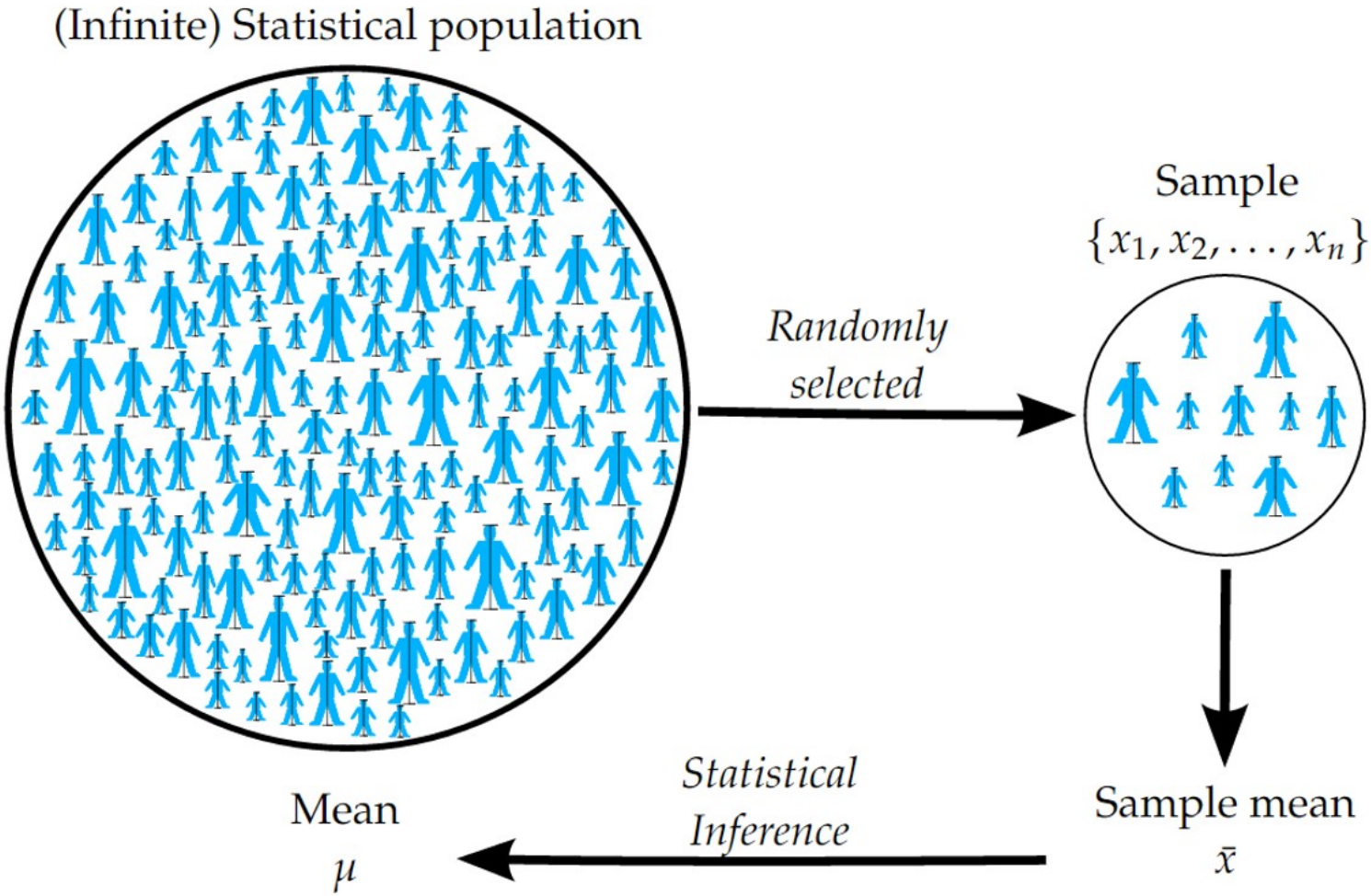
Statistik og Ingeniører

- Statistik er et vigtigt værktøj i problemløsning
- Analyse af data
- Kvalitetforbedring
- Forsøgsplanlægning
- Forudsigelse af fremtidige værdier
- .. og meget mere!

Statistik

- Moderne statistik har baggrund i sandsynlighedsregning og beskrivende statistik
- Statistik handler ofte om at analysere en *stikprøve* (sample), der er taget fra en *population* (population)
- Baseret på stikprøven, vil vi generalisere (eller udtale os) om populationen
- Det er derfor vigtigt, at stikprøven er *repræsentativ* for populationen

Statistik


(Infinite) Statistical population
Sample
$\{x_1, x_2, \ldots, x_n\}$
Randomly selected
Statistical Inference
Mean
μ
Sample mean
$\bar{x}$

Statistical software

Der eksisterer en række forskellige statistiske software programmer, i dette kursus bruger vi R:

- Open source (`www.r-project.org`)
- Stort antal bidrag ydere (så det vokser hurtigt)
- Skal bruge et interface, vi bruger R-studio (`www.rstudio.com`), også open source. Der er andre muligheder, eksempelvis Emacs Speak Statistics.
- I skal instalere R og jeg vil anbefale at i instalere R-studio
- Vi bruger R gennem kurset og I SKAL lave projektet i R.
- Et statistisk software program er en uundværlig det af statistisk analyse, MEN vi er nødt til at tænke over hvad vi putter ind i R, dvs. brug blyant og papir (og hovedet).

Eksempel 1

Udtag stikprøve til brug for udregning af nøgletal....

Oversigt

Nøgletal (summary statistics)

Vi anvender en række *nøgletal* (eller statistikker) for at opsummere og beskrive data (og stokastiske variable)

- **Gennemsnit:** tyngdepunkt eller centrering
- **Median:** tyngdepunkt eller centrering
- **Varians:** variation
- **Spredning:** variation (samme enhed som data)
- **Variations koefficient:** variationen i data (enhedsløs)
- **Kovarians:** samvariation mellem værdier
- **Korrelation:** samvariation mellem værdier (enhedsløs)
- **Fraktiler:** siger noget om fordelingen af data

Nøgletal (summary statistics)

Lad $x_1, ..., x_n$ og $y_1, ..., y_n$ være stikprøver

		Definition	R
Gennemsnit:	$\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	`mean(x)`
Varians:	s^2	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	`var(x)`
Spredning:	s	$\sqrt{s^2}$	`sd(x)`
Kovarians:	s_{xy}	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	`cov(x,y)`
Korrelation:	r	$\frac{s_{xy}}{s_x \cdot s_y}$	`cor(x,y)`

Nøgletal (summary statistics)

Lad $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ betegne den sorterede rækkefølge af $x_1, ..., x_n$.

		Definition	R
Varianskoefficient:	V	$\frac{s}{\bar{x}} \cdot 100$	
Fraktiler:	τ_p	$\frac{\left(x_{(np)}+x_{(np+1)}\right)}{2}$, $np = \lceil np \rceil$ $x_{(\lceil np \rceil)}$, $np \neq \lceil np \rceil$	`quantile(x,...)`[1]
Første kvartil	Q_1	$\tau_{0.25}$	
Median	$\tilde{x}$	$\tau_{0.50}$	`median(x)`
Tredie kvartil	Q_3	$\tau_{0.75}$	

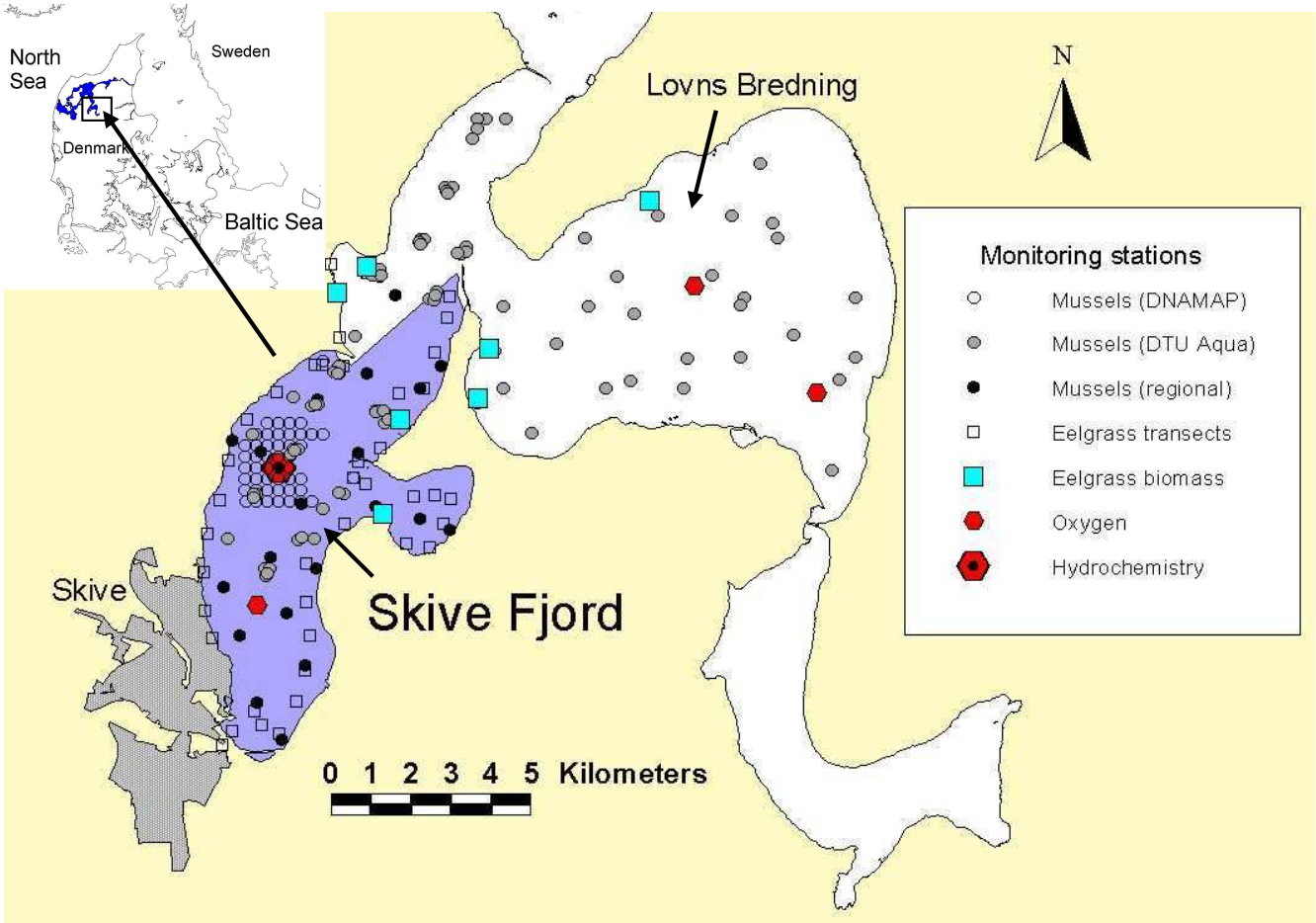
[1] brug `quantile(x, probs = p, type = 2)` for definitionen ovenfor

Oversigt

Grafisk Fremstilling

- Histogram
- Empirisk kumulativ tæthedsfunktion
- Boxplot
- Scatterplot

Eksempel: Skive fjord



Eksempel: Skive fjord, data

Data sættet indeholder en række månedlige gennemsnits observationer of forskellige variable, idag ser vi på

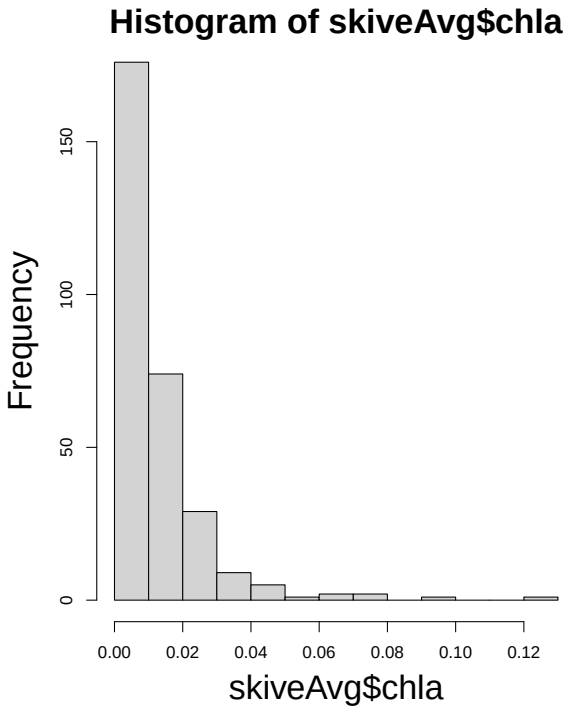
- `chla` der er er en måling af klorofyll i vandet, dvs. der er en indikator for mængden af alger i vandet.
- `temp` temperaturen i vandet

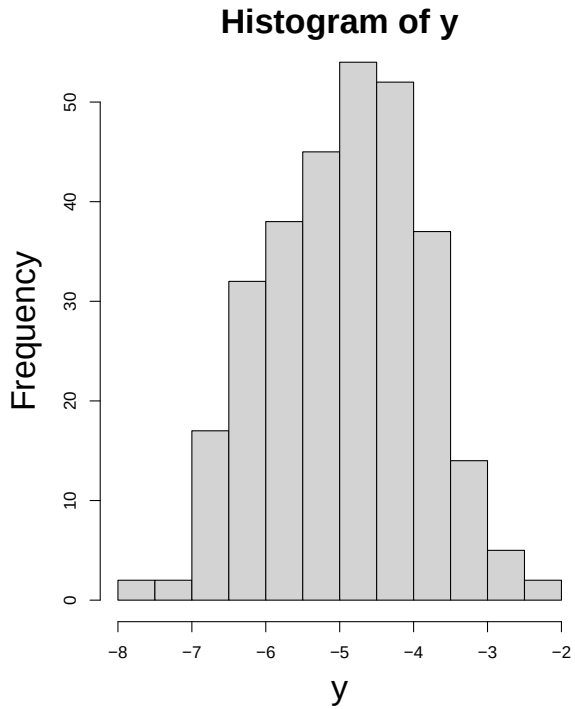
Eksempel: nogle nøgletal

```
> ## read data
> skiveAvg <- read.table("skiveAvg.csv", sep = ";", header = TRUE)
>
> ## number of observations
> dim(skiveAvg)

## [1] 300  17

> ## Some key numbers
> summary(skiveAvg[ ,c("year","month","chla","temp")])

##       year          month             chla              temp
##  Min.   :1982   Min.   : 1.00   Min.   :0.00050   Min.   :-0.74
##  1st Qu.:1988   1st Qu.: 3.75   1st Qu.:0.00356   1st Qu.: 3.75
##  Median :1994   Median : 6.50   Median :0.00792   Median : 8.36
##  Mean   :1994   Mean   : 6.50   Mean   :0.01205   Mean   : 9.53
##  3rd Qu.:2000   3rd Qu.: 9.25   3rd Qu.:0.01524   3rd Qu.:15.42
##  Max.   :2006   Max.   :12.00   Max.   :0.12092   Max.   :21.77
```

Histogram

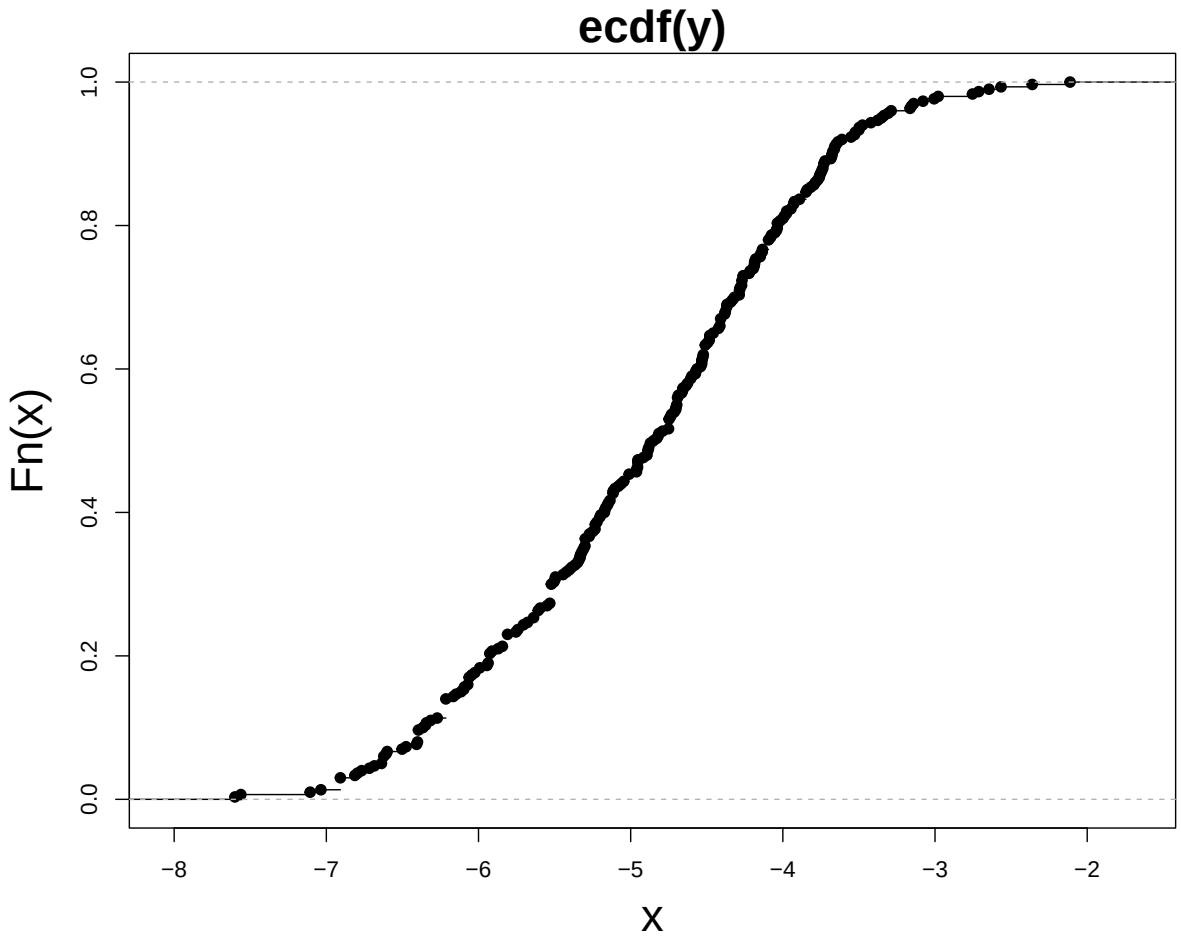
```
> ## Data from Skive fjord
> skiveAvg <- read.table("skiveAvg.csv", sep = ";", header = TRUE)
> y <- log(skiveAvg$chla)
> hist(skiveAvg$chla); hist(y)
```



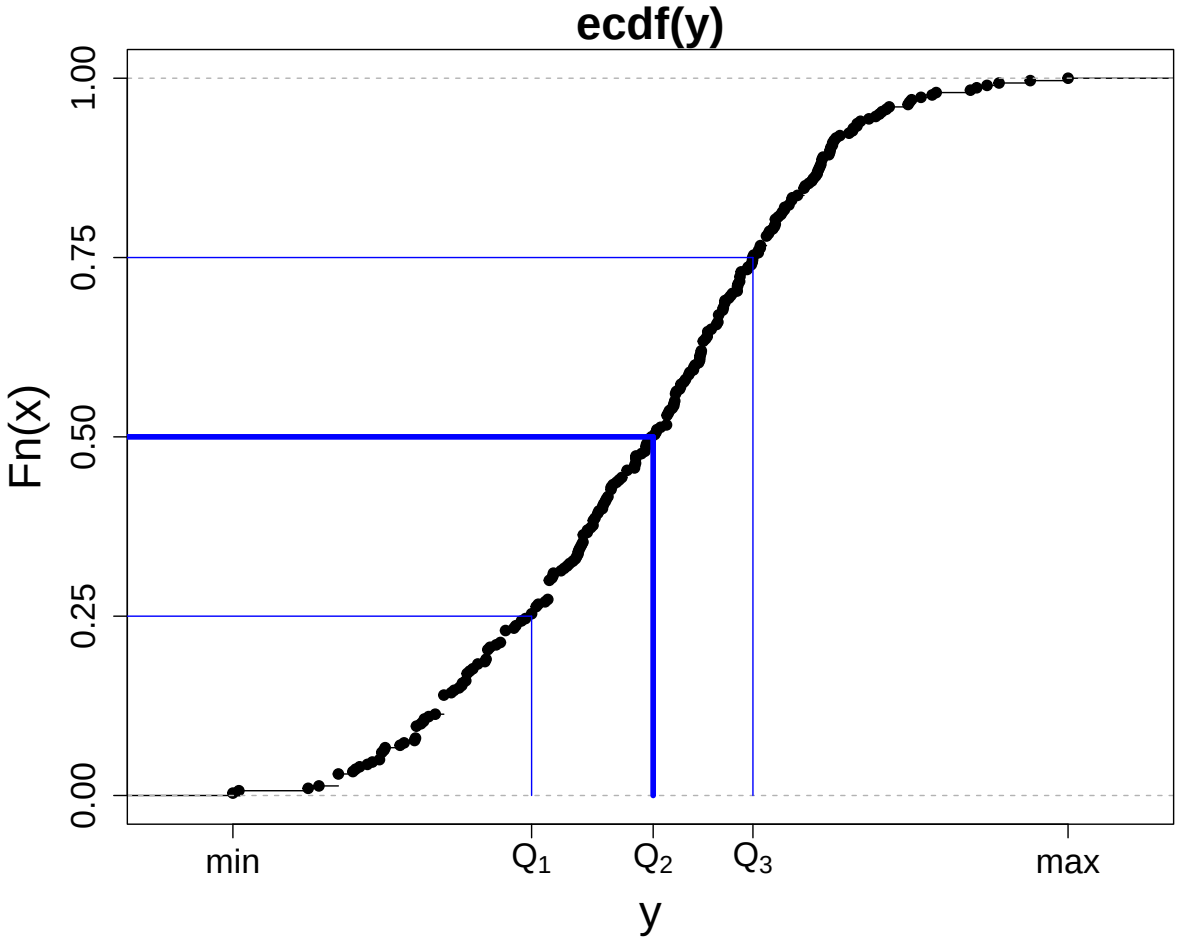


Empirisk fordelingsfunktion

```
> plot(ecdf(y))
```

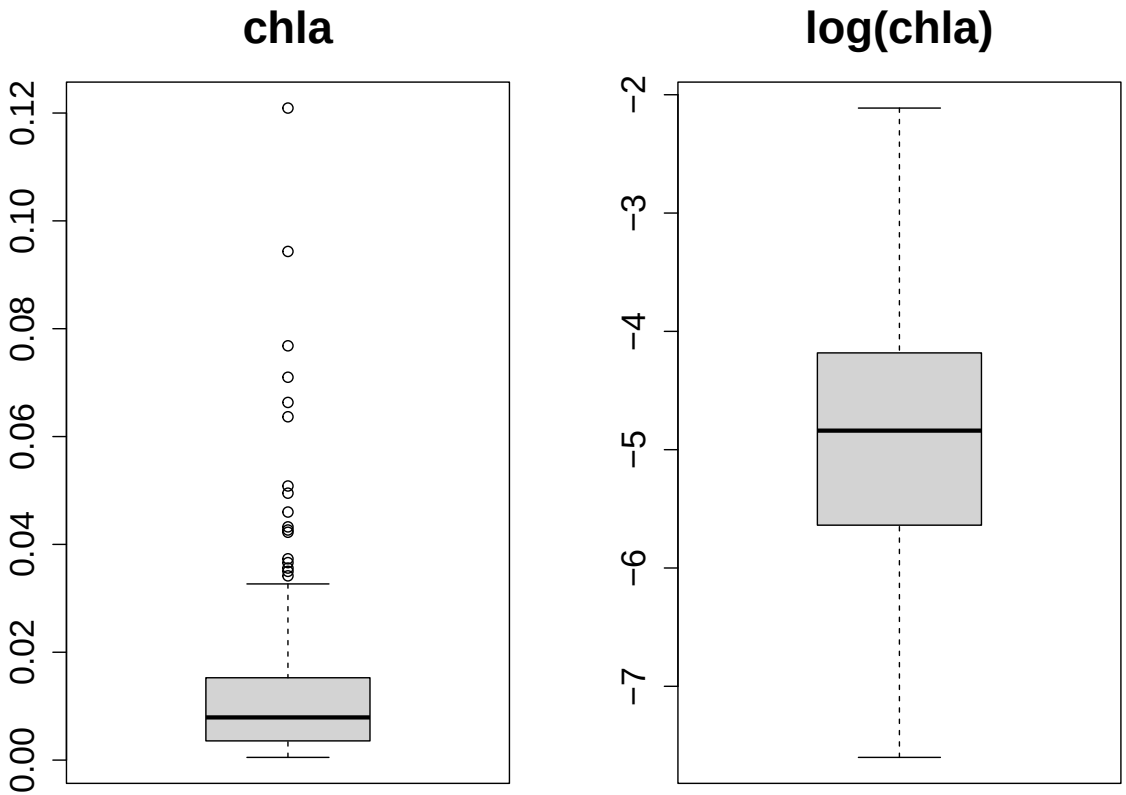


Empirisk fordelingsfunktion

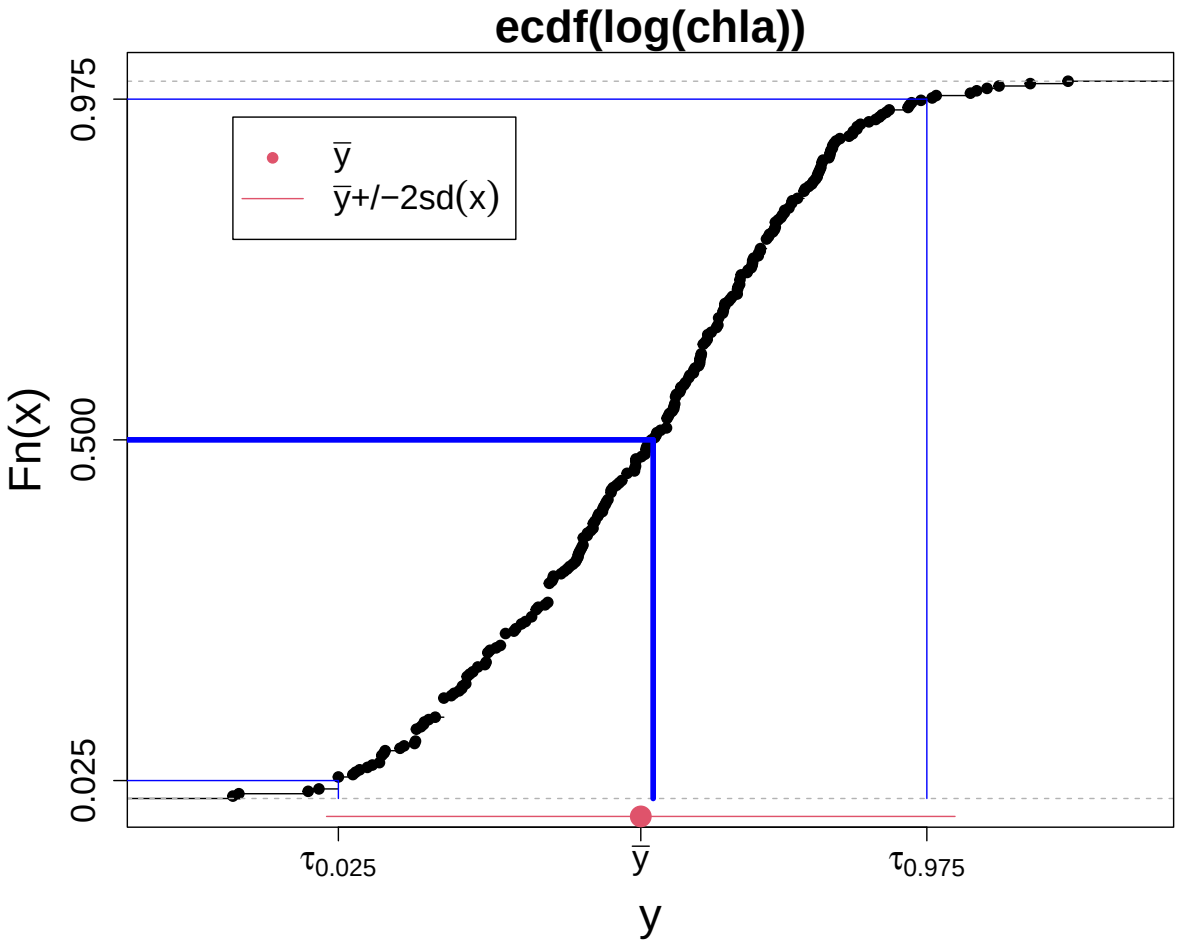


Boxplot

```
> par(mfrow=c(1,2),cex.main=2,cex.lab=2,cex.axis=1.5)
> boxplot(skiveAvg[ ,"chla"],main="chla")
> boxplot(y,main="log(chla)")
```

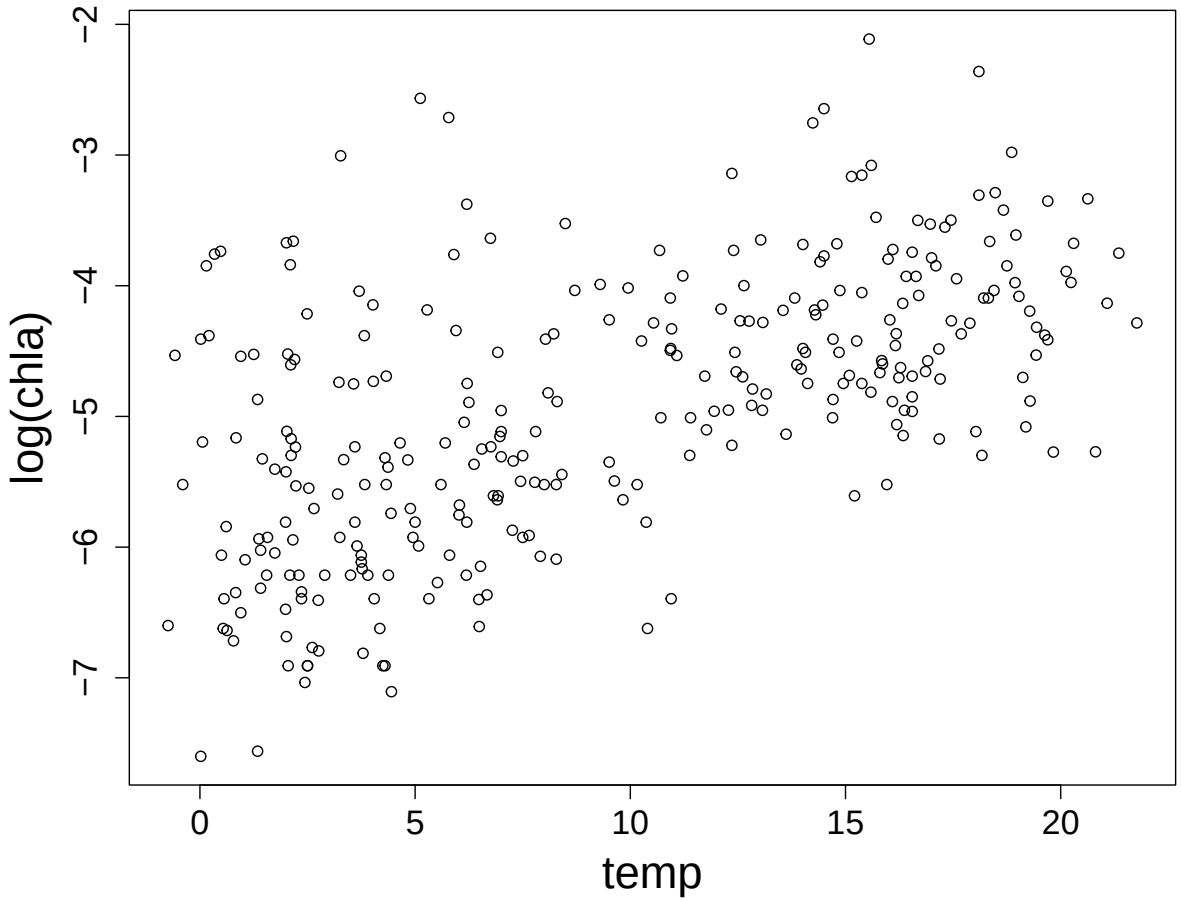


Empirisk fordelingsfunktion

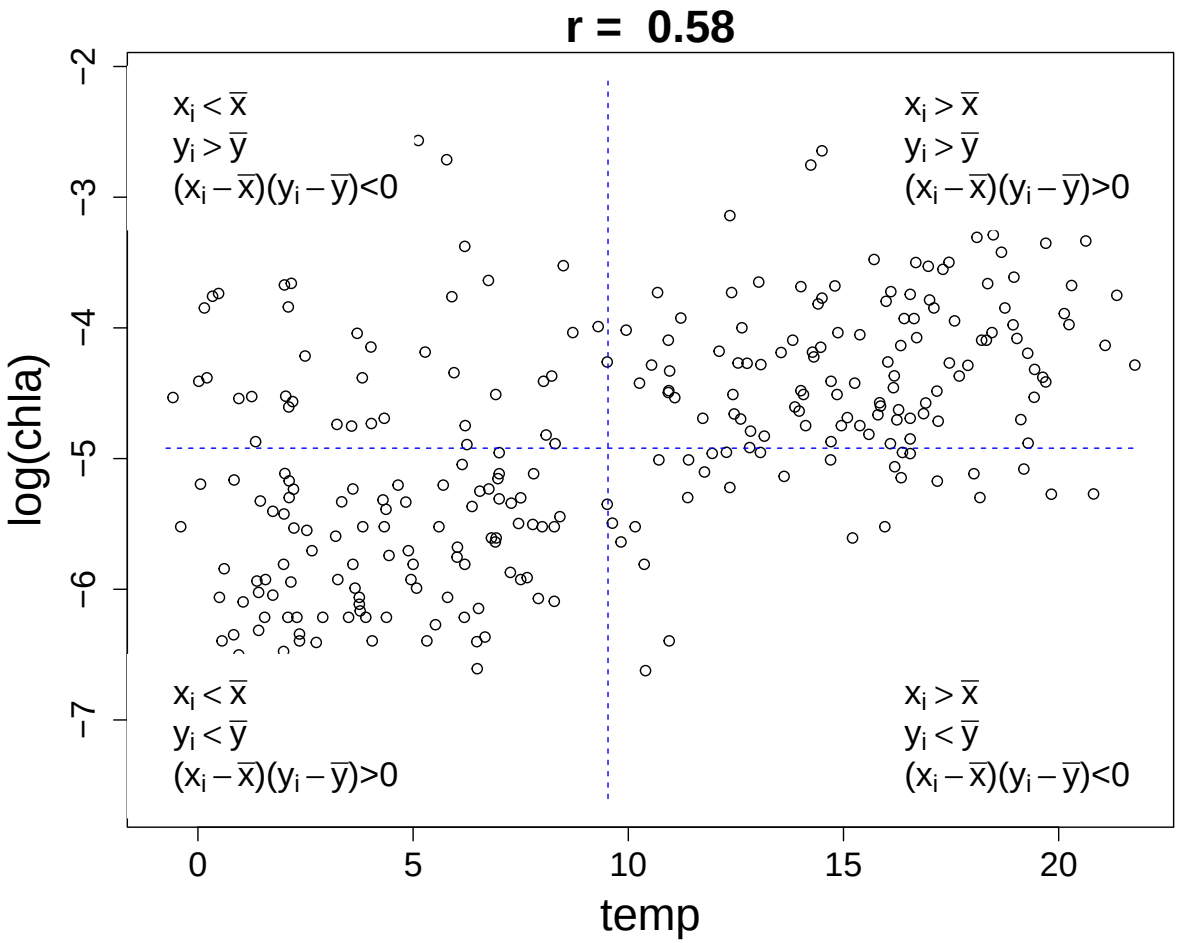


Scatterplot

```
> plot(log(chla) ~ temp, data = skiveAvg)
```



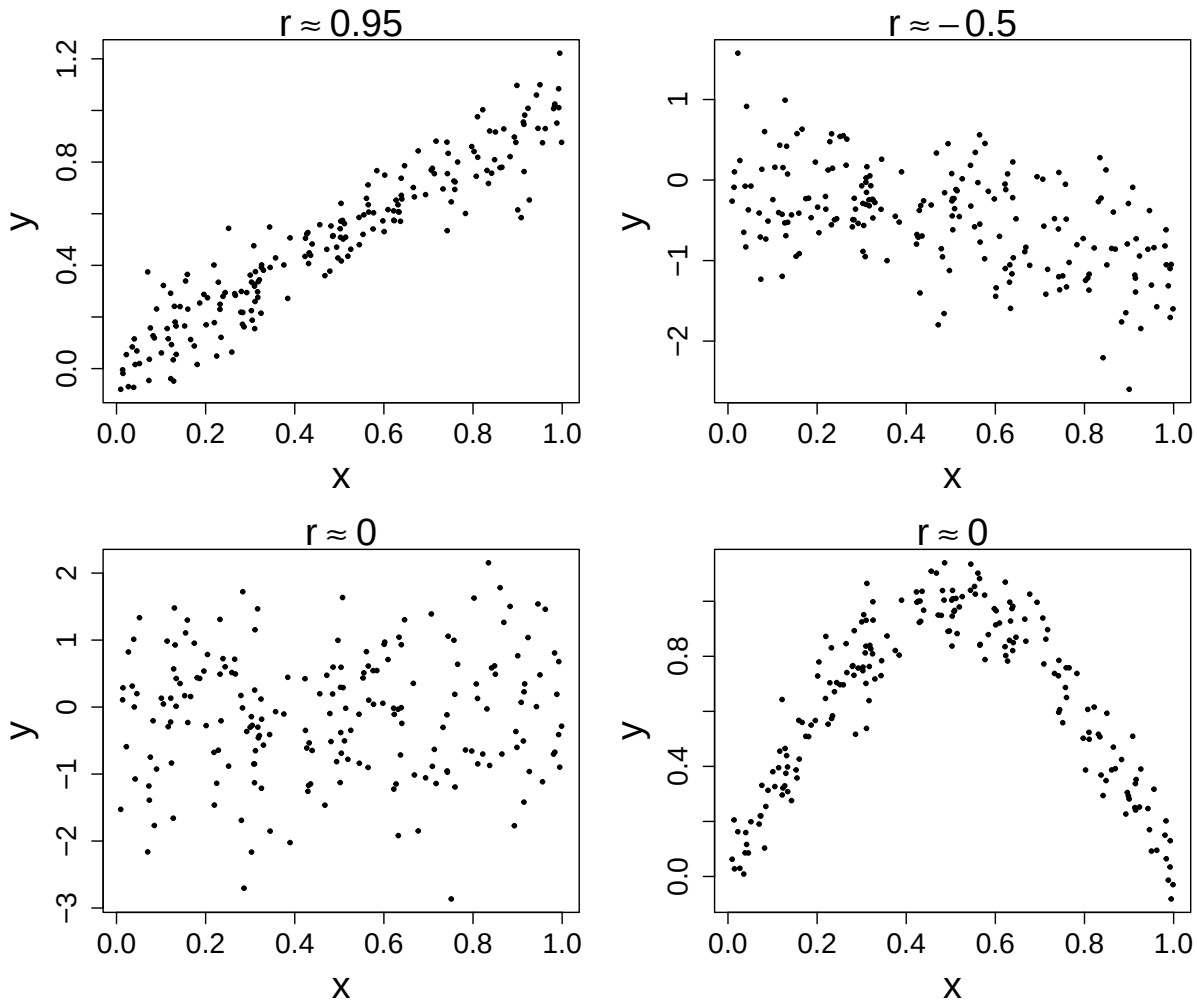
Scatterplot



Korrelation (igen) - egenskaber

- r is always between -1 and 1: $-1 \leq r \leq 1$
- r measures the degree of linear relation between x and y
- $r = \pm 1$ if and only if all points in the scatterplot are exactly on a line
- $r > 0$ if and only if the general trend in the scatterplot is positive
- $r < 0$ if and only if the general trend in the scatterplot is negative

Korrelation



Oversigt

Stokastiske variable

En stokastisk variabel (random variable) repræsenterer udfaldet af et eksperiment der endnu ikke er udført

- Et terningekast
- Antallet af seksere i 10 terningekast
- km/l for en bil
- Måling af sukkerniveau i blodprøve
- ...

eller med andre ord en beskrivelse af hvordan data realiseres.

Diskret eller kontinuert

- Vi skelner mellem diskret og kontinuert
- Diskret kan tælles:
  - Hvor mange bliver indlagt med corona i morgen
  - Antal mange fly letter den næste time
- Kontinuert:
  - Vindmåling
  - Tiden det tog at komme til DTU

Stokastisk variabel

Før eksperimentet er udført stokastisk variabel haves

$$X_i$$

noteret med stort bogstav.
Så udføres eksperimentet, og vi har da en *realisation* eller *observation*

$$x_i$$

noteret med småt bogstav.
Formelt er X_i en funktion, mens x_i er et tal.

Simuler et terningekast

Vælg et tal fra $(1, 2, 3, 4, 5, 6)$ med lige sandsynlighed for hvert udfald

Diskrete fordelinger

- For diskrete fordelinger (dvs. tælleligt udfaldrum), kan X antage værdierne $\{0, 1, 2, ...\}$
- Kan naturligvis også være endeligt dvs. X kan antage værdierne $\{0, 1, 2, ..., N\}$
- I alle tilfælde skal vi beskrive alle mulige udfald ved en model.
- I denne sammenhæng er en model en beskrivelse af sandsynligheden for hvert enkelt (mulige) udfald.

Tæthedsfunktion (Diskrete fordelinger)

En stokastisk variabel har en *tæthedsfunktion* (probability density function (pdf))

Definition

$$f(x) = P(X = x)$$

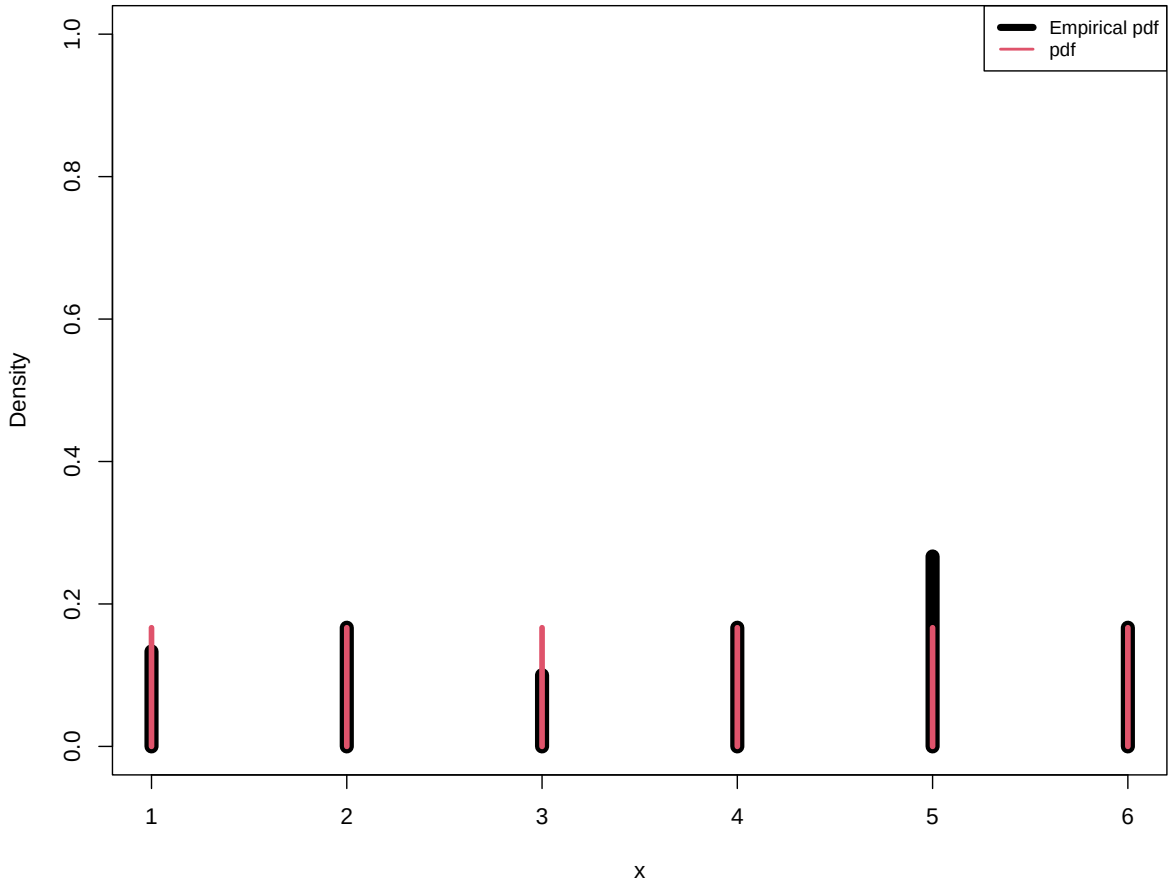
Sandsynligheden for at X bliver udfaldet x når eksperimentet udføres. Der gælder at:

$$f(x) \geq 0 \quad \text{for all } x$$

$$\sum_x f(x) = 1$$

Tæthedsfunktion

En fair ternings tæthedsfunktion



Stikprøve

Hvis vi har n observationer, så har vi en *stikprøve* (a sample)

$$\{x_1, x_2, ..., x_n\}$$

og da kan vi begynde at "se" fordelingen. Jo større n jo mere information.

Oversigt

Middelværdi (mean) og forventningsværdi (expectation)

Stokastisk variabels middelværdi

$$\mu = E(X) = \sum_{\text{alle x}} x f(x)$$

- Det "rigtige gennemsnit"
- Fortæller hvor "midten" af X er

Middelværdi eksempel

Middelværdi af en terning

$$\begin{aligned}\mu = E(X) &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5\end{aligned}$$

Jo flere observationer, jo tættere kommer man på den rigtige middelværdi[2]

$$\lim_{n\to\infty} \hat{\mu} = \mu$$

- hvor $\hat{\mu}$ er det observerede gennemsnit
- Prøv det i R

[2] Givet at μ eksisterer

Varians

Definition

$$\sigma^2 = \mathsf{Var}(X) = \sum_{\text{alle x}} (x-\mu)^2 f(x)$$

- Et mål for spredningen
- Den "rigtige spredning" af X (modsat empirisk varians (sample variance))

Varians eksempel

Varians af terningekast

$$\sigma^2 = E[(X-\mu)^2] = (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} + (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} \approx 2.92$$

Nøgletal

	Empirisk	Diskret stokastisk variable
Middelværdi	$\bar{x} = \sum x_i \frac{1}{n}$	$\mu = \sum x_i f(x_i)$
Varians	$s^2 = \sum (x_i - \bar{x})^2 \frac{1}{n-1}$	$\sigma^2 = \sum (x_i - \mu)^2 f(x_i)$
Kovarians	$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$	
Median	$x_{(\lceil n/2 \rceil)}$ [1]	"$F^{-1}(0.5)$" [2]
Fraktil	Q_τ [1]	"$F^{-1}(\tau)$"

[1] se slide 26 for præcis definition

[2] Mere præcist: x s.t. $P(X \leq x) \geq 0.5$ og $P(X \geq x) \geq 0.5$

Næste forelæsning:

- Stokastiske variable, Sandsynligheder, diskrete og kuntinuerte fordelinger fordelinger - kapitel 2 i eNoten

Oversigt