# Course 02402 Introduction to Statistics

## Lecture 13: An overview of course content

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark



## Overview







## Overview







## Kapitel 1: Simple plots og deskriptiv statistik

Engelsk

Tag en *stikprøve*: Brug deskriptiv statistik til at "se" på den!

### Opsummerende størrelser for stikprøve

- Gennemsnittet ($\bar{x}$)
- Standardafvigelse ($s$)
- Empirisk varians ($s^2$)
- Fraktiler og percentiler *(f.eks. 15% af data ligger under 0.15 fraktilen)*
- Median, øvre- og nedre kvartiler
- Empirisk korrelation ($r$) *(mellem to stikprøver)*

### Simple plots

- Scatter plot *(xy plot)*
- Histogram *(empirisk tæthed)*
- Kumulativ fordeling *(empirisk fordeling)*
- Boxplots, søjlediagram, cirkeldiagram (lagkagediagram)

## Overview







## Kapitel 2: Diskrete fordelinger

Engelsk

Grundlæggende koncepter:

- Stokastisk variabel *(værdi afhængig af udfald af endnu ikke udført eksperiment)*
- Tæthedsfunktion: $f(x) = P(X = x)$ *(pdf)*
- Fordelingsfunktion: $F(x) = P(X \leq x)$ *(cdf)*
- Middelværdi: $\mu = E(X)$
- Standard afvigelse: $\sigma$
- Varians: $\sigma^2$

Specifikke distributioner:

- Binomial *(tæl antal succes ud af n trækninger)*
- Hypergeometrisk *(trækning uden tilbagelægning)*
- Poisson *(antal hændelser i interval)*





## Overview







## Kapitel 2: Kontinuerte fordelinger

Engelsk

Grundlæggende koncepter:

- Tæthedsfunktion: $f(x)$ *(pdf)*
- Fordelingsfunktion: $F(x) = P(X \leq x)$ *(cdf)*
- Middelværdi ($\mu$) og varians ($\sigma^2$)
- Regneregler for stokastiske variabler (lineære funktioner)

Specifikke fordelinger:

- Normal
- Log-Normal
- Uniform
- Eksponential

Funktioner af normalfordeling (afsn. 2.10) (introduceres først i de næste uger):

- $t$-fordelingen, $\chi^2$-fordelingen *(Chi-i-anden)* og $F$-fordelingen

## Overview







## Kapitel 3: Konfidensintervaller for én gruppe/stikprøve

Engelsk

### Grundlæggende koncepter

- Population og tilfældig stikprøve
- Statistisk model
- Estimation *(f.eks. $\hat{\mu}$ er estimat af $\mu$)*
- Signifikansniveau $\alpha$
- Konfidensintervaller *(fanger rigtige prm. $1-\alpha$ af gangene)*
- Stikprøvefordelinger *(stikprøvegennemsnit ($t$) og empirisk varians ($\chi^2$))*
- Centrale grænseværdisætning

### Specifikke metoder, én gruppe/stikprøve

- Konfidensinterval for middelværdi ($t$-fordeling)
- Konfidensinterval for varians ($\chi^2$-fordeling)





## Overview







## Kapitel 3: Hypotesetests for én gruppe/stikprøve

Engelsk

### Grundlæggende koncepter:

- Hypoteser *($H_0$ vs. $H_1$)*
- $p$-værdi *(Sandsynlighed for observeret eller mere ekstrem værdi af teststørrelsen, hvis $H_0$ er sand, e.g. $P(T > t_{\text{obs}})$)*
- Type I fejl *(I virkeligheden ingen effekt, men $H_0$ afvises)*
  - $P(\text{Type I}) = \alpha$ *(Sandsynligheden for at begå type I fejl)*
- Type II fejl *(I virkeligheden effekt, men $H_0$ afvises ikke)*
  - $P(\text{Type II}) = \beta$ *(Sandsynligheden for type II fejl)*
- Modelkontrol

### Specifikke metoder, én gruppe:

- $t$-test for middelværdiniveau
- Modelkontrol med normal qq-plot

## Overview







## Kapitel 3: Statistik for to populationer (2 stikprøver)

Engelsk

Specifikke metoder, to populationer:

- Konfidensinterval for forskel i middelværdi
- Test for forskel i middelværdi ($t$-test)
- To PARREDE grupper: "Tag differencen" $\Rightarrow$ "Én gruppe"





## Overview







## Kapitel 4: Statistik ved simulering

Engelsk

Simulering:

- Træk tilfældige værdier og beregn statistik mange gange
- Fejlforplantning (error propagation rules)
  *(F.eks. igennem ikke-lineær funktion)*
- Bootstrapping af konfidensintervaller:
  - Parametrisk *(Simuler mange udfald af stokastisk var.)*
  - Ikke-parametrisk *(Træk direkte fra data)*

Specifikke setups: (4 versioner af konfidensintervaller)

- Èn gruppe/stikprøve og to grupper/stikprøver data
- Parametrisk vs. ikke-parametrisk

## Overview







## Kapitel 5: Simpel lineær regressions analyse

Engelsk

### To variable: $x$ og $y$

- Beregn mindstekvadraters estimat af ret linje

### Inferens med simpel lineær regressionsmodel

- Statistisk model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, konfidensintervaller og tests for $\beta_0$ og $\beta_1$
- $1-\alpha$ konfidensinterval for linjen *(stor sikkerhed for den rigtige linje ligger indenfor)*
- $1-\alpha$ prædiktionsinterval for punkter *(stor sikkerhed for at nye punkter er indenfor)*

### $\rho$, $R$ og $R^2$

- $\rho$ er korrelationen ($= sign_{\beta_1} R$) er graden af lineær sammenhæng mellem $x$ og $y$
- $R^2$ er andelen af den totale variation som er forklaret af modellen
- Afvises $H_0 : \beta_1 = 0$ så afvises også $H_0 : \rho = 0$





## Overview







## Kapitel 6: Multipel lineær regressions analyse

Engelsk

### Multipel lineær regressionsmodel

- Flere variabler: $Y$, $x_1$, $x_2$, ...
  *($y$ afhængig/respons var. og $x$'er er forklarende/uafhængige var.)*
- Mindstekvadraters rette plan *(et plan da der er >2 dimensioner)*

### Inferens for en multipel lineær regressionmodel

- Statistisk model: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$
- Estimation af konfidensintervaller og tests for $\beta$'er
- Konfidensintervaller for modellen (middelplanet)
- Prædiktionsintervaller for nye punkter
- $R^2$ er andelen af den totale variationen som er forklaret af modellen

### Model validering af antagelser ved residual analyse

- Normalfordeling? q-q plots af residualer
- Uafhængighed? Plot residualer mod prædikterede værdier $\hat{y}_i$ og inputs $x_{j,i}$

## Overview







## Kapitel 7: Inferens for andele

Engelsk

### Statistik for andele:

- Andel: $p = \frac{x}{n}$ *($x$ successer ud af $n$ observationer)*
- Specifikke metoder, én, to og $k > 2$ grupper
  - Binær/kategorisk respons

### Specifikke metoder:

- Estimation og konfidensintervaller for andele
  - Metoder korrektion ved små stikprøver
- Hypoteser for én andel ($p$)
- Hypoteser for to andele
- Analyse af antalstabeller ($\chi^2$-test) (alle forventede antal $> 5$)





## Overview







## Kapitel 8: Envejs variansanalyse (envejs ANOVA)

Engelsk

### $k$ UAFHÆNGIGE grupper

- Test om middelværdi for mindst en gruppe er forskellig fra de andre gruppers middelværdi
- Model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

### Specifikke metoder, envejs variansanalyse:

- ANOVA-tabel: $SST = SS(Tr) + SSE$
- $F$-test
- Post hoc test(s): Parvise $t$-test med poolet varians estimat
  - Hvis planlagt på forhånd, så uden Bonferroni korrektion
  - Hvis alle sammenligninger udføres, så med Bonferroni korrektion

## Overview







## Afsnit 3.3 og 7.2.2: Forsøgsplanlægning

Engelsk

Grundlæggende koncepter for forsøgsplanlægning:

- Testens styrke er $1-\beta$ *(hvor $\beta$ er sandsynligheden for at begå Type II fejl)*

Specifikke metoder, forsøgsplanlægning
(middelværdi, både one og two sample setup):

- Stikprøvestørrelse $n$ for ønsket præcision af konfidensintervaller
- Stikprøvestørrelse $n$ for ønsket styrke af tests

Specifikke metoder, forsøgsplanlægning
(andel, one sample setup):

- Stikprøvestørrelse $n$ for ønsket præcision af konfidensintervaller





## Overview







## Chapter 1: Simple Graphics and Summary Statistics

Dansk

Take a *sample*: Use descriptive statistics to "look" at it!

Summary statistics

- Sample mean: $\bar{x}$
- Sample standard deviation: $s$
- Sample variance: $s^2$
- Quantiles and percentiles *(e.g. 15% of data is below 0.15 quantile)*
- Median, upper- and lower quartiles
- Sample correlation ($r$) *(between two samples)*

Simple graphics

- Scatter plot *(xy plot)*
- Histogram *(empirical density)*
- Cumulative distribution *(empirical distribution)*
- Boxplots, Bar charts, Pie charts

## Overview







## Chapter 2: Discrete Distributions

Dansk

General concepts:

- Random variable *(value is outcome of yet not carried out experiment)*
- Density function: $f(x) = P(X = x)$ *(pdf)*
- Distribution function: $F(x) = P(X \leq x)$ *(cdf)*
- Mean: $\mu = E(X)$
- Standard deviation: $\sigma$
- Variance: $\sigma^2$

Specific distributions:

- The binomial distribution *(dice roll)*
- The hypergeometric distribution *(draw without replacement)*
- The Poisson distribution *(number of events in interval)*





## Overview







## Chapter 2: Continuous Distributions

Dansk

General concepts:

- Density function: $f(x)$ *(pdf)*
- Distribution: $F(x) = P(X \leq x)$ *(cdf)*
- Mean ($\mu$) and variance ($\sigma^2$)
- Calculation rules for random variables

Specific distributions:

- Normal
- Log-Normal
- Uniform
- Exponential

Funktions of normaldist. (Sec. 2.10) (introduced in the coming weeks):

- $t$-distribution, $\chi^2$-distribution *(Chi-square)* og $F$-distribution

## Overview







## Chapter 3: One sample confidence intervals

Dansk

### General concepts

- Population and a random sample
- Statistical model
- Estimation *(e.g. $\hat{\mu}$ is estimate of $\mu$)*
- Significance level $\alpha$
- Confidence intervals *(Catches true value $1-\alpha$ times)*
- Sampling distributions *(sample mean ($t$) and sample vaiance ($\chi^2$))*
- Central Limit Theorem

### Specific methods, one sample

- Confidence interval for the mean ($t$-distribution)
- Confidence interval for the variance ($\chi^2$-distribution)





## Overview







## Chapter 3: One sample hypothesis testing

Dansk

### General concepts:

- Hypotheses *($H_0$ vs. $H_1$)*
- $p$-value *(Probability for observing the test value or more extreme, if $H_0$ is true, e.g. $P(T > t_{\text{obs}})$)*
- Type I error *(No effect in reality, but $H_0$ is rejected)*
  - $P(\text{Type I}) = \alpha$ *(The probability for a Type I error)*
- Type II error: *(In reality an effect, but $H_0$ is not rejected)*
  - $P(\text{Type II}) = \beta$ *(The probability for a Type II error)*
- Model validation

### Specific methods, one sample:

- $t$-test for the mean
- Model validation with normal q-q plot

## Overview







## Chapter 3: Two Samples

Dansk

Specific methods, two samples:

- Confidence interval for the mean difference
- Test for the mean difference ($t$-test)
- Two PAIRED samples: "Take difference" $\Rightarrow$ "One sample"





## Overview







## Chapter 4: Statistics by simulation

Dansk

Simulation:

- Draw random values and calculate the statistic many times
- Error propagation rules
  *(e.g. through a non-linear function)*
- Bootstrapping of confidence intervals:
  - Parametric *(Simulate many outcomes of random var.)*
  - Non-parametric *(Draw values directly from data)*

Specific situations: (4 versions of confidence intervals)

- One-sample and Two-sample data
- Parametric vs. non-parametric

## Overview







## Chapter 5: Simple linear Regression Analysis

Dansk

### Two quantitative variables: $x$ and $y$

- Calculate the least squares line

### Inferences for a simple linear regression model

- Statistical model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, confidence intervals and tests for $\beta_0$ and $\beta_1$.
- $1-\alpha$ confidence interval for the line *(high certainty that the real line will be inside)*
- $1-\alpha$ prediction interval for punkter *(high certainty that new points will be inside)*

### $\rho$, $R$ and $R^2$

- $\rho$ is the correlation $(= sign_{\beta_1} R)$ is the strength of linear relation between $x$ and $y$
- $R^2$ is the fraction of the total variation explained by the model
- If $H_0 : \beta_1 = 0$ is rejected, then $H_0 : \rho = 0$ is also rejected





## Overview







## Chapter 6: Multiple linear Regression Analysis

Dansk

### Multipel lineær regressionsmodel

- Many quantitative variables: $y$, $x_1$, $x_2$, ...
  *($y$ is the dependent/response var. and $x$'s are explanatory/independent var.)*
- Calculating least squares surface *(a plane surface since there are >2 dimensions)*

### Inferences for a the multiple linear regression model

- Statistical model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_2 x_{p,i} + \varepsilon_i$
- Confidence interval estimation and test for the $\beta$'s
- Confidence interval for the model (the mean surface)
- Prediction interval for new points
- $R^2$ expresses the proportion of the total variation explained by the linear fit

### Model validation of assumptions with residual analysis

- Normal distribution? q-q plots of residuals
- Independence? Plot residuals against predicted values $\hat{y}_i$ and inputs $x_{j,i}$

## Overview







## Chapter 7: Inferences for Proportions

Dansk

Statistics for proportions:

- Proportion: $p = \frac{x}{n}$ *($x$ successes out of $n$ observations)*
- Specific methods: one, two and $k > 2$ samples:
  - Binary/categorical response

Specific methods:

- Estimation and confidence interval of proportions
  - Methods for correction for small samples
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables ($\chi^2$-test) (all expected $> 5$)





## Overview







## Chapter 8: One-way Analysis of Variance

Dansk

$k$ INDEPENDENT samples (groups)

- Test if the mean of at least one of the groups is different from the mean of the other groups
- Model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Specific methods, one-way analysis of variance:

- ANOVA-table: $SST = SS(Tr) + SSE$
- $F$-test
- Post hoc test(s): pairwise $t$-test with pooled variance estimate
  - If planned on beforehand, then without Bonferroni correction
  - If all samples are compared, then with Bonferroni correction

## Overview







## Section 3.3 and 7.2.2: Design of experiments

Dansk

General concepts for design of experiments:

- Power of a test is $1-\beta$ *(where $\beta$ is the probability of making a Type II error)*

Specific methods, design of experiments
(mean, both one and two sample setup):

- Sample size $n$ for wanted precision of confidence intervals
- Sample size $n$ for wanted power of tests

Specific methods, design of experiments
(proportion, one-sample setup):

- Sample size $n$ for wanted precision of confidence intervals