

Kursus 02402/02323 Introduktion til statistik

Forelæsning 13: Et overblik over kursets indhold

Peder Bacher

DTU Compute, Dynamiske Systemer
Building 303B, Room 017
Danish Technical University
2800 Lyngby – Denmark
e-mail: pbac@dtu.dk

Overview

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Simpel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 1: Simple plots og deskriptiv statistik

Engelsk

- Teknikker til at “se” på data! (deskriptiv statistik)
- Opsummerende størrelser for stikprøve
 - Gennemsnittet (\bar{x})
 - Empirisk standard afvigelse (s)
 - Empirisk varians (s^2)
 - Fraktiler og percentiler (*f.eks. 15% af data ligger under 0.15 fraktil*)
 - 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, cirklediagram (lagkagediagram)

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger**
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetest for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote2: Diskrete fordelinger

Engelsk

- Grundlæggende koncepter:
 - Stokastisk variabel (*den får værdi afhængigt 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: σ
 - Varians: σ^2
- Specifikke distributioner:
 - Binomial (*terningekast*)
 - Hypergeometrisk (*trækning uden tilbagelægning*)
 - Poisson (*antal hændelser i interval*)

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger**
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 2: Kontinuerte fordelinger

Engelsk

- Grundlæggende koncepter:
 - Tæthedsfunktion: $f(x)$ (*pdf*)
 - Fordelingsfunktion: $F(x) = P(X \leq x)$ (*cdf*)
 - Middelværdi (μ) og varians (σ^2)
 - Regneregler for stokastiske variable
- Specifikke fordelinger:
 - Normal
 - Log-Normal
 - Uniform
 - Exponential
 - t
 - χ^2 (*Chi-i-anden*)
 - F

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve**
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 3: Konfidensintervaller for én gruppe/stikprøve

Engelsk

- Grundlæggende koncepter
 - Population og tilfældig stikprøve
 - Estimation (*f.eks. $\hat{\mu}$ er estimat af μ*)
 - Signifikans niveau α
 - Konfidensintervaller (*fanger rigtige prm. $1 - \alpha$ af gangene*)
 - Stikprøvefordelinger (*stikprøvegennemsnit (t) og empirisk varians (χ^2)*)
 - Centrale grænseværdisætning
- Specifikke metoder, én gruppe/stikprøve:
 - Konfidensintervaller for middelværdi (t -fordeling) og varians (χ^2 fordeling)
 - Forsøgsplanlægning: beregn stikprøvestørrelsen n for den ønskede præcision

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetest for én gruppe/stikprøve**
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 3: Hypotesetests for én gruppe/stikprøve

Engelsk

- Grundlæggende koncepter:
 - Hypoteser (H_0 vs. H_1)
 - p -værdi (*sandsynlighed for observeret værdi eller mere ekstremt 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*)
 $P(\text{Type I}) = \alpha$
 - Type II fejl: (*i virkeligheden effekt, men H_0 afvises ikke*)
 $P(\text{Type II}) = \beta$
 - Testens styrke er β
- Specifikke metoder, én gruppe:
 - t -test for middelværdiniveau
 - Stikprøvestørrelse for ønsket styrke
 - Modelkontrol med normal qq-plot

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver**
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 3: Statistik for to grupper/stikprøver

Engelsk

- Specifikke metoder, to grupper:
 - Test og konfidensintervaller for forskel i middelværdi (t -test)
 - Forsøgsplanlægning: Beregn sample størrelsen for den ønskede styrke
- Specifikke metoder, to PARREDE grupper:
 - "Tag differencen for hver måling" \Rightarrow "statistik for én gruppe"

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation**
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 4: Statistik ved simulation

Engelsk

- Introduktion til simulering
(Beregn statistik mange gange)
- Fejlforplantning (error propagation rules)
(F.eks. igennem ikke-lineær funktion)
- Bootstrapping:
 - Parametrisk *(Simuler mange udfald af stokastisk var.)*
 - Ikke-parametrisk *(Træk direkte fra data)*
 - Konfidensintervaller (og derfor også hypotesetest)
- Specifikke setups: (4 versioner af konfidensintervaller)
 - En gruppe/stikprøve og to grupper/stikprøver data
 - Parametrisk vs. ikke-parametrisk

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Simpel lineær regressions analyse**
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 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 af konfidensintervaller og tests for β_0 og β_1
 - Konfidensintervaller for linjen (*95% gange ligger linjen indenfor*)
 - Prædiktionsintervaller for punkter (*95% af nye punkter ligger indenfor*)
- ρ , R og R^2
 - ρ er korrelationen ($= \text{sign}_{\beta_1} R$) beskriver 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 $H_0 : \rho = 0$

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse**
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 6: Multipel lineær regressions analyse

Engelsk

- Flere variabler: y, x_1, x_2, \dots
(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: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \varepsilon_i$
 - Estimation af konfidensintervaller og tests for β 'er
 - Konfidensintervaller for modellen *(For det forventede plan)*
 - Prædiktionsintervaller for nye punkter
- R^2 er andelen af den totale variationen som er forklaret af modellen

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetestes for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)**
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 8: Envejs variansanalyse (envejs ANOVA)

Engelsk

- k UAFHÆNGIGE grupper
- Specifikke metoder, envejs variansanalyse:
 - Test der sammenligner middelværdien af grupperne
 - 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

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)**
- 12 eNote 7: Inferens for andele

eNote 8: Tovejs variansanalyse (tovejs ANOVA)

Engelsk

- Blokdesign giver to faktorer
- ANOVA-tabel: $SST = SS(Tr) + SS(Bl) + SSE$
 - SST , $SS(Tr)$ og $SS(Bl)$ beregnes som ved envejs ANOVA
 - $SSE = SST - SS(Tr) - SS(Bl)$
- 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

Oversigt

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetestes for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Sempel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele**

eNote 7: Inferens for andele

Engelsk

- Andel: $p = \frac{x}{n}$ (x succeser ud af n observationer)
- Specifikke metoder, én, to og $k > 2$ grupper
 - Binær/kategorisk respons
- Estimation og konfidensintervaller for andele
 - Metoder til store stikprøver vs. til små stikprøver
- Hypoteser for én andel (p)
- Hypoteser for to andele
- Analyse af antalstabeller (χ^2 -test) (Alle forventede antal > 5)

Overview

- 1 eNote 1: Simple plots og deskriptive statistik
- 2 eNote 2: Diskrete fordelinger
- 3 eNote 2: Kontinuerte fordelinger
- 4 eNote 3: Konfidensintervaller for én gruppe/stikprøve
- 5 eNote 3: Hypotesetests for én gruppe/stikprøve
- 6 eNote 3: Statistik for to grupper/stikprøver
- 7 eNote 4: Statistik ved simulation
- 8 eNote 5: Simpel lineær regressions analyse
- 9 eNote 6: Multipel lineær regressions analyse
- 10 eNote 8: Envejs variansanalyse (envejs ANOVA)
- 11 eNote 8: Tovejs variansanalyse (ANOVA)
- 12 eNote 7: Inferens for andele

eNote 1: Simple Graphics and Summary Statistics

Dansk

- Look at data as it is! (descriptive statistics)
- 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

eNote 2: Discrete Distributions

Dansk

- General concepts:
 - Random variable (*Gets its value dependent on 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: σ
 - Variance: σ^2
- Specific distributions:
 - The binomial distribution (*Dice roll*)
 - The hypergeometric distribution (*Draw without replacement*)
 - The Poisson distribution (*Number of events in interval*)

eNote 2: Continuous Distributions

Dansk

- General concepts:
 - Density function: $f(x)$ (*pdf*)
 - Distribution: $F(x) = P(X \leq x)$ (*cdf*)
 - Mean (μ) and variance (σ^2)
 - Calculation rules for random variables
- Specific distributions:
 - Normal
 - Log-Normal
 - Uniform
 - Exponential
 - t
 - χ^2 (*Chi-square*)
 - F

eNote 3: One sample confidence intervals

Dansk

- General concepts
 - Population and a random sample
 - Estimation (*e.g. $\hat{\mu}$ is estimate of μ*)
 - Significance level α
 - Confidence intervals (*Catches true value $1 - \alpha$ times*)
 - Sampling distributions (*sample mean (t) and sample variance (χ^2)*)
 - Central Limit Theorem
- Specific methods, one sample:
 - Confidence intervals for the mean (t -distribution) and variance (χ^2 distribution)
 - Design of experiments: calculating the sample size n for wanted precision

eNote 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*)
 $P(\text{Type I}) = \alpha$
 - Type II error: (*In reality an effect, but H_0 is not rejected*)
 $P(\text{Type II}) = \beta$
 - Power of a test is β
- Specific methods, one sample:
 - t -test for mean difference
 - Sample size for wanted power
 - Model validation with normal qq-plot

eNote 3: Two Samples

Dansk

- Specific methods, two samples:
 - Test and confidence interval for the mean difference (t -test)
 - Planning: calculating the sample size for wanted power
- Specific methods, two PAIRED samples:
 - "Take difference" \Rightarrow "One sample"

eNote 4, Statistics by simulation

Dansk

- Introduction to simulation
(Calculate the statistic many times)
- Error propagation rules
(e.g. through a non-linear function)
- Bootstrapping:
 - Parametric *(Simulate many outcomes of random var.)*
 - Non-parametric *(Draw values directly from data)*
 - Confidence intervals (and hence also hypothesis testing)
- Specific situations: (4 versions of confidence intervals)
 - One-sample and Two-sample data
 - Parametric vs. non-parametric

eNote 5: Simple linear Regression Analysis

Dansk

- Two quantitative variables: x and y
- Calculating least squares line
- Inferences for a simple linear regression model
 - Statistical model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
 - Interval estimation and test for β_0 and β_1 .
 - Confidence interval for the line (*95% times the line will be inside*)
 - Prediction interval for punkter (*95% times new points will be inside*)
- ρ , R og R^2
 - ρ is the correlation ($= \text{sign}_{\beta_1} R$) describes 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

eNote 6: Multiple linear Regression Analysis

Dansk

- Many quantitative variables: y, x_1, x_2, \dots
(*y is the dependent/response var. and x's are explanatory/independent var.*)
- Calculating least squares plane (*A plane since there are >2 dimensions*)
- Inferences for a the multiple linear regression model
 - Statistical model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \varepsilon_i$
 - Confidence interval estimation and test for the β 's
 - Confidence interval for the expected fit (*fitted line*)
 - Prediction interval for new points
- R^2 expresses the proportion of the total variation explained by the linear fit

eNote 8: One-way Analysis of Variance

Dansk

- Specific methods, k INDEPENDENT samples
- One-way analysis of variance
 - Test for comparing the means of the groups
 - 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

eNote 8: Two-way Analysis of Variance

Dansk

- Block design - two-way analysis of variance
- ANOVA-tabel: $SST = SS(Tr) + SS(Bl) + SSE$
 - SST , $SS(Tr)$ and $SS(Bl)$ calculated as one-way ANOVA
 - $SSE = SST - SS(Tr) - SS(Bl)$
- 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

eNote 7: Inferences for Proportions

Dansk

- Proportion: $p = \frac{x}{n}$ (*x successes out of n observations*)
- Specific methods, one, two and $k > 2$ samples
 - Binary/categorical response
- Estimation and confidence interval of proportions
 - Large sample vs. small sample methods
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables (χ^2 -test) (All expected > 5)