

Kursus 02403 Introduktion til matematisk statistik

Forelæsning 13: Et overblik over kursets indhold

Jan Kloppenborg Møller

DTU Compute, Dynamiske systemer
Building 303B, Room 016
Danish Technical University
2800 Lyngby – Denmark
e-mail: jkmo@dtu.dk

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 1: Simple plots og deskriptive statistik

- Teknikker til at “se” på data! (deskriptiv statistik)
- Opsummerende statistikker
 - Gennemsnittet: \bar{x}
 - Empirisk standard afvigelse: s
 - Empirisk varians: s^2
 - Median, øvre- og nedre kvartiler, fraktiler, korrelationer
- Simple plots
 - Scatter plot (*xy plot*)
 - Histogram (*empirisk tæthed*)
 - Kumulativ fordeling (*empirisk fordeling*)
 - Boxplots, søjlediagram, cirkeldiagram (lagkagediagram)

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote2: Diskrete fordelinger

- Grundlæggende koncepter:
 - Stokastisk variabel (*udfaldet af et 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 fordelinger:
 - Binomial (*Kast med een mønt*)
 - Hypergeometrisk (*trækning uden tilbagelægning*)
 - Poisson (*antal hændelser i interval*)

eNote 2: Kontinuerte fordelinger

- 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

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable**
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 2+4:Funktioner af stokastiske variable

- Grundlæggende koncepter:

- Error propagation ($V(f(X_1, \dots, X_n))$) - Taylor
- Simulering: Fordelings/tæthedsfunktion for $Y = f(X_1, \dots, X_n)$

- Samplings fordelinger:

- X_1, \dots, X_n iid $N(\mu, \sigma^2)$
- Normal (\bar{X} når σ^2 kendt)
- χ^2 - $\sum X_i^2$ eller S^2
- t - standardliseret (med observeret standardafvigelse) gennemsnit
- F - Forholdet mellem χ^2 fordelinger (S_y^2/S_x^2)

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample**
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 3: Konfidensintervaller for én gruppe/stikprøve

- Grundlæggende koncepter
 - Estimation
 - Signifikans niveau α
 - Konfidensintervaller (*fanger rigtige værdi $1 - \alpha$ af gangene*)
 - Population og tilfældig stikprøve
 - Stikprøvefordelinger (t og 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

eNote 3: Hypotese tests for én gruppe/stikprøve

- Grundlæggende koncepter:
 - Hypoteser
 - p -værdi (*sandsynlighed for teststørrelsen eller mere ekstremt, 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 $1 - \beta$
- Specifikke metoder, én gruppe:
 - t -test for middelværdiniveau
 - Stikprøvestørrelse for ønsket styrke
 - Normal qq-plot

eNote 3: Statistik for to grupper/stikprøver

- 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"

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation**
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 4: Statistik ved simulation

- 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:
 - En gruppe/stikprøve og to grupper/stikprøver data
 - Middelværdier (μ s) og/eller andre features (θ s)

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse**
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 5: Simpel lineær regressions analyse

- To variable: x og y
- Beregn mindstekvadraters estimat af rette linje
- Inferens med simpel lineær regressionsmodel
 - Statistisk model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
 - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
 - 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 $H_0 : \rho = 0$

eNote 6: Multipel lineær regressions analyse

- 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: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \varepsilon_i$
 - $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
 - 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

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele**
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

eNote 7: Inferens for andele

- 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 ($np \geq 15$ og $(1 - p)n \geq 15$)
- Hypoteser for to andele
- Analyse af antalstabeller (χ^2 -test) (Alle forventede antal > 5)

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)**
- 9 Some further perspectives
- 10 Evaluering

eNote 8: Envejs variansanalyse (envejs ANOVA)

- 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/uden Bonferroni korrektion

eNote 8: Tovejs variansanalyse (tovejs ANOVA)

- 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: parvise t -test med/uden Bonferroni korrektion

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives**
- 10 Evaluering

Some further perspectives

- Likelihood theory (estimations teknik)
- General Linear Models (GLM) - generalisering af multiple linear regression og variansanalyse
- Generalized Linear Models - GLM men andre fordelinger end den Gaussiske
- Korrelations strukturer
- Stokastiske dynamiske systemer
- og meget mere ...

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering

Evaluering

Husk evaluering!

Overview

- 1 eNote 1: Simple plots og deskriptive statistikker
- 2 eNote2: Fordelinger
- 3 eNote 2+4: Funktioner af stokastiske variable
- 4 eNote 3: One and two sample
- 5 eNote 4: Statistik ved simulation
- 6 eNote 5-6: Lineær regressions analyse
- 7 eNote 7: Inferens for andele
- 8 eNote 8: Variansanalyse (ANOVA)
- 9 Some further perspectives
- 10 Evaluering