

02402: Introduktion til Statistik

Forelæsning 13: Overblik over kursets indhold

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Introduktion til Statistik

De væsentligste koncepter i kurset:

- Måder at opsummere data – plots, nøgletal osv.
- Stokastiske variable
 - Forskellige fordelinger
- Statistiske modeller
 - Estimer og esimatorer
 - Konfidensintervaller
 - Hypotesetests
- Validering af modeller og CLT
- R som et værktøj til alt dette.

Tag introstat videre med i andre kurser og projekter, og byg også ovenpå.

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 1: Simple plots og deskriptiv statistik

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

Opsummerende størrelser for stikprøve

- Gennemsnittet (\bar{x})
- Standardafvigelse (s)
- Empirisk varians/stikprøvevariens (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

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

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)**
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 2: Diskrete fordelinger

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: $\mu = E(X)$
- Standardafvigelse: σ
- Varians: σ^2

Specifikke fordelinger:

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

Kapitel 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 (lineære funktioner)

Specifikke fordelinger:

- Normal
- Log-Normal
- Uniform
- Eksponential

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)**
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 3: Konfidensintervaller for én gruppe/stikprøve

Grundlæggende koncepter

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

Specifikke metoder, én gruppe/stikprøve

- Konfidensinterval for middelværdi (t -fordeling)
- Konfidensinterval for varians (χ^2 -fordeling)

Kapitel 3: Hypotesetests for én gruppe/stikprøve

Grundlæggende koncepter:

- Hypoteser (H_0 vs. H_1)
- p -værdi (*hvis H_0 er sand, sandsynlighed for mere ekstrem værdi af teststørrelsen.*)
- 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

Kapitel 3: Statistik for to populationer (2 stikprøver)

Specifikke metoder, to stikprøver:

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

Styrkeplanlægning

- One-sample CI: Stikprøvestørrelse n for ønsket præcision af konfidensintervaller.
- One-sample hypotesetest: Stikprøvestørrelse n for ønsket styrke (power).
- Two-sample hypotesetest: Stikprøvestørrelse n for ønsket styrke
- Kan også "vendes rundt".

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)**
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 4: Statistik ved simulering

Simulering:

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

Specifikke setups: (4 versioner af konfidensintervaller)

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

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)**
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 5: Simpel lineær regressions analyse

To variable: x og y

- Beregn mindstekvadratets 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 β_0 og β_1
- $1 - \alpha$ konfidensinterval for linjen (*stor sikkerhed for at den rigtige regressionslinje ligger indenfor*)
- $1 - \alpha$ prædiktionsinterval for punkter (*stor sikkerhed for at nye observationer er indenfor*)

ρ , R og R^2

- ρ er korrelationen ($= \text{sign}(\beta_1)R$) 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

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)**
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 6: Multipel lineær regressions analyse

Multipel lineær regressionsmodel

- Flere variabler: Y, x_1, x_2, \dots
(y afhængig/respons var. og x'er er forklarende/uafhængige variable)
- Mindstekvadratets metode igen *(men mere kompliceret da der er ≥ 2 forklarende var.)*

Inferens for en multipel lineær regressionmodel

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

Modelvalidering af antagelser ved residualanalyse

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

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)**
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)

Kapitel 7: Inferens for andele

Statistik for andele:

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

Specifikke metoder:

- Estimation og konfidensintervaller for andele
 - Korrektionsmetoder ved små stikprøver
- Hypoteser for én andel (p)
- Hypoteser for to andele
- Analyse af antalstabeller (χ^2 -test) (alle forventede antal > 5)

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)**

Kapitel 8: Envejs variansanalyse (envejs ANOVA)

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 poollet variansestimant
 - Hvis planlagt på forhånd, så uden Bonferroni korrektion
 - Hvis alle sammenligninger udføres, så med Bonferroni korrektion

Kapitel 8: Tovejs variansanalyse (two-way ANOVA)

k behandlinger (faktor 1), l blokke (faktor 2), $k \times l$ observationer

- Kan undersøge for effekt af blok og effekt af behandling
- Model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Specifikke metoder, tovejs variansanalyse:

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

Introduktion til Statistik

De væsentligste koncepter i kurset:

- Måder at opsummere data – plots, nøgletal osv.
- Stokastiske variable
 - Forskellige fordelinger
- Statistiske modeller
 - Estimer og esimatorer
 - Konfidensintervaller
 - Hypotesetests
- Validering af modeller og CLT
- R som et værktøj til alt dette.

Tag introstat videre med i andre kurser og projekter, og byg også ovenpå.

Overview

- 1 Kapitel 1: Simple plots og deskriptiv statistik (Uge 1)
- 2 Kapitel 2: Sandsynligheder og fordelinger (Uge 2-3)
- 3 Kapitel 3: Statistik, én og to stikprøver (uge 4-6)
- 4 Kapitel 4: Statistik ved simulation (Uge 7)
- 5 Kapitel 5: Simpel lineær regression (Uge 8)
- 6 Kapitel 6: Multipel lineær regression (Uge 9)
- 7 Kapitel 7: Inferens for andele (Uge 10)
- 8 Kapitel 8: Variansanalyse (ANOVA), (uge 11-12)