Introduktion til Statistik

Forelæsning 13: Et overblik over kursets indhold

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2017

Overview

Kapitel 1: Simple plots og deskriptiv statistik

Engelsk

Teknikker til at "se" på data! (deskriptiv statistik)

Opsummerende størrelser for stikprøve

- Gennemsnittet ($\bar{x}$)
- 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
- Empririsk korrelation (r) *(mellem to stikprøver)*

Simple plots

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

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: σ
- Varians: σ^2

Specifikke distributioner:

- Binomial *(terningekast)*
- Hypergeometrisk *(trækning uden tilbagelægning)*
- Poisson *(antal hændelser i interval)*

Kapitel 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 variabler

Specifikke fordelinger:

- Normal
- Log-Normal
- Uniform
- Eksponential
- t
- χ^2 *(Chi-i-anden)*
- F

Kapitel 3: Konfidensintervaller for én gruppe/stikprøve

Engelsk

Grundlæggende koncepter

- Population og tilfældig stikprøve
- Estimation *(f.eks. $\hat{\mu}$ er estimat af μ)*
- Signifikansniveau α
- Konfidensintervaller *(fanger rigtige prm. $1-\alpha$ 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

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

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"

Grundlæggende koncepter for forsøgsplanlægning:

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

Specifikke metoder, forsøgsplanlægning:

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

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

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 β_0 og β_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)*

ρ, R og R^2

- ρ 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 $H_0 : \rho = 0$

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: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_2 x_{p,i} + \varepsilon_i$
- 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

Model validering ved residual analyse

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

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

Kapitel 8: Tovejs variansanalyse (tovejs ANOVA)

Engelsk

k UAFHÆNGIGE grupper og blokdesign der giver to faktorer

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

Specifikke metoder, tovejs variansanalyse:

- 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

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 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)

Chapter 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

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: 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)*

Chapter 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

Chapter 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 vaiance (χ^2))*
- Central Limit Theorem

Specific methods, one sample

- Confidence interval for the mean (t-distribution)
- Confidence interval for the variance (χ^2-distribution)

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

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"

General concepts for design of experiments:

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

Specific methods, design of experiments:

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

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

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 β_0 and β_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)*

ρ, R and R^2

- ρ 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

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 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

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

Chapter 8: Two-way Analysis of Variance

Dansk

k INDEPENDENT treatments and block design give two factors

- Test if mean for at least one group is different from the others
- Model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Specific methods, two-way analysis of variance:

- ANOVA-table: $SST = SS(Tr) + SS(Bl) + SSE$
  - SST, $SS(Tr)$ and $SS(Bl)$ calculated as in 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

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:
  - Large sample vs. small sample methods
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables (χ^2-test) (All expected > 5)