

Introduktion til Statistik

Forelæsning 13: Et overblik over kursets indhold

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Efterår 2020

Overview

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

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})
- Standard afvigelse (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)

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

Specifikke distributioner:

- Binomial (*tæl antal succes 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

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

- t -fordelingen, χ^2 -fordelingen (*Chi-i-anden*) og F -fordelingen

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 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

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)

Specifikke metoder, to populationer:

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

Kapitel 4: Statistik ved simulering

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)

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

Kapitel 5: Simpel lineær regressions analyse

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 ($= \text{sign}_{\beta_1} R$) 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 $H_0 : \rho = 0$

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 var.*)
- Mindstekvadraters rette plan (*et plan da der er >2 dimensioner*)

Inferens for en multipel lineær regressionmodel

- Statistisk model: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$
- 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

Model validering af antagelser ved residual analyse

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

Kapitel 7: Inferens for andele

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 (χ^2 -test) (alle forventede antal > 5)

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 poolet varians estimat
 - Hvis planlagt på forhånd, så uden Bonferroni korrektion
 - Hvis alle sammenligninger udføres, så med Bonferroni korrektion

Afsnit 3.3 og 7.2.2: Forsøgsplanlægning

Grundlæggende koncepter for forsøgsplanlægning:

- Testens styrke er $1 - \beta$ (*hvor β 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

Chapter 1: Simple Graphics and Summary Statistics

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

Chapter 2: Discrete Distributions

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

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

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

- t -distribution, χ^2 -distribution (*Chi-square*) og F -distribution

Chapter 3: One sample confidence intervals

General concepts

- Population and a random sample
- Statistical model
- 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 interval for the mean (t -distribution)
- Confidence interval for the variance (χ^2 -distribution)

Chapter 3: One sample hypothesis testing

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

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"

Chapter 4: Statistics by simulation

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

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 ($= \text{sign}(\beta_1) R$) 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

Multiple lineær regressionsmodel

- 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 surface (*a plane surface since there are >2 dimensions*)

Inferences for a the multiple linear regression model

- Statistical model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$
- Confidence interval estimation and test for the β '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 $x_{j,i}$

Chapter 7: Inferences for Proportions

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 (χ^2 -test) (all expected > 5)

Chapter 8: One-way Analysis of Variance

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

Section 3.3 and 7.2.2: Design of experiments

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
(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