

TBD

TBA

Daniel Johnsen, Gonde Winkelmann, Linda Knudsen
Morten Thygesen, Niklas Andersen, Salma Hirsi

Datavidenskab & Machine Learning, cs-23-dvml-2-02, March 30, 2023

P2 projekt



Copyright © Aalborg University 2015

Here you can write something about which tools and software you have used for typesetting the document, running simulations and creating figures. If you do not know what to write, either leave this page blank or have a look at the colophon in some of your books.



Elektronik og IT
Aalborg Universitet
<http://www.aau.dk>

AALBORG UNIVERSITET

STUDENTERRAPPORT

Titel:

TBD

Abstract:

Her er resuméet

Tema:

Fra data til videnskab

Projektperiode:

Forårssemester 2023

Projektgruppe:

cs-23-dvml-2-02

Deltager(e):

Daniel Elsborg Johnsen
Gonde Leon Winkelmann
Linda Knudsen
Morten Tejlmand Thygesen
Niklas Kjærgaard Andersen
Salma Ahmed Hassan Hirsi

Vejleder(e):

Søren Byg Vilsen

Oplagstal: 0

Sidetal: 14

Afleveringsdato:




30. marts 2023

Rapportens indhold er frit tilgængeligt, men offentliggørelse (med kildeangivelse) må kun ske efter aftale med forfatterne.

Contents

Preface	v
1 Introduction	1
1.1 Examples	1
1.2 How Does Sections, Subsections, and Subsections Look?	1
1.2.1 This is a Subsection	1
2 Problemanalyse	3
2.1 Teori: Statistik	3
2.1.1 Population og stikprøve	3
2.1.2 Middelværdi	4
2.1.3 Varians	4
2.1.4 Standardafvigelse	4
2.1.5 Fordelingsteori	5
2.1.6 Estimationsteori	7
2.2 Teori: PRNG	8
2.2.1 Fordelings Test (χ^2)	9
2.2.2 Spektral Test	9
2.2.3 LCG	10
2.2.4 R's Pseudo-Random Number Generator	10
3 Chapter 2 name	11
4 Conclusion	12
Bibliography	13
A Appendix A name	14

Todo list

 Is it possible to add a subsubparagraph?	2
 I think that a summary of this exciting chapter should be added.	2
 I think this word is misspelled	11
Figure: We need a figure right here!	11

Preface

Here is the preface. You should put your signatures at the end of the preface.

Aalborg University, March 30, 2023

Author 1

<username1@XX.aau.dk>

Author 2

<username2@XX.aau.dk>

Author 3

<username3@XX.aau.dk>

Chapter 1

Introduction

Here is the introduction. The next chapter is chapter 3.
a new paragraph

1.1 Examples

You can also have examples in your document such as in example 1.1.

Example 1.1 (An Example of an Example)

Here is an example with some math

$$0 = \exp(i\pi) + 1 . \tag{1.1}$$

You can adjust the colour and the line width in the `macros.tex` file.

1.2 How Does Sections, Subsections, and Subsections Look?

Well, like this

1.2.1 This is a Subsection

and this

This is a Subsubsection

and this.

A Paragraph You can also use paragraph titles which look like this.

A Subparagraph Moreover, you can also use subparagraph titles which look like this. They have a small indentation as opposed to the paragraph titles.

I think that a summary of this exciting chapter should be added.

Is it possible to add a subsub-paragraph?

Chapter 2

Problemanalyse

2.1 Teori: Statistik

Statistik er en videnskabelig metode, der systematisk og empirisk beskriver data ved hjælp af forskellige matematiske operationer og metoder. En grundig statistisk analyse kræver indsamling og organisering af data, efterfulgt af analyse ved hjælp af forskellige metoder og statistiske tests for at udvælge information. Typen af data varierer alt efter, hvad der undersøges, og kan derfor omfatte numeriske data såvel som kvalitative data, der skal systematiseres, før de kan analyseres statistisk. I dette kommende afsnit vil der blive introduceret forskellige statistiske begreber, såsom stikprøve og middelværdi, for at danne en overordnet forståelse af teorien. Dette er vigtigt for at kunne udføre diverse statistiske tests senere hen. Derfor vil teorien og metoden blive undersøgt i det kommende afsnit.

2.1.1 Population og stikprøve

I statistisk sammenhæng forstås "population" som en gruppe af enheder eller individer, som skal undersøges og udvindes information fra. Enhver gruppe af ting, der skal undersøges statistisk, kaldes en population. Nogle eksempler på populationer, der kan undersøges, er mænd født efter 1990, Startups i Danmark eller generelt en gruppe af enheder med lignende karakteristika. En "stikprøve" er en tilfældigt udvalgt prøve, der udtages fra en større population. Stikprøven bør have en statistisk signifikant størrelse i forhold til den samlede population, så den kan give et mere eller mindre pålideligt indtryk af hele mængdens egenskaber. Stikprøver bruges ofte, fordi det enten ikke er muligt eller er for krævende at undersøge hele populationen. Et eksempel på en stikprøve er en meningsmåling af befolkningen, hvor der undersøges en mindre del af populationen, og observationerne tilskrives populationen som helhed. Når man estimerer observationer fra en stikprøve til en samlet population, vil der altid være en vis grad af statistisk usikkerhed. Dog kan man

reducere usikkerheden ved at tage højde for faktorer som stikprøvens repræsentativitet for den samlede population, muligt bias i stikprøven, som kan påvirke og gøre resultaterne mindre generaliserbare, og størrelsen på ens stikprøve, hvor større stikprøver er mere repræsentative for den samlede population.

2.1.2 Middelværdi

For populationer (μ) og stikprøver (\bar{x}) beregnes middelværdien ved at tage summen af alle observationerne og dividere det med antallet af observationer.

$$\bar{x}, \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

Hvor n er antallet af observationer og x_i er den i 'te observation i populationen eller stikprøve.

2.1.3 Varians

Variansen for populationer beregnes ved at summere den kvadrerede forskel af hver observation og middelværdien og derefter divideres med antallet af observationer fra populationen:

$$\sigma = \sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$$

Hvor x_i er den i 'te observation, n er antallet af observationer og μ er middelværdien af populationen.

Variansen for stikprøven beregnes på tilsvarende måde som populationen, dog benyttes antallet af observationer fra stikprøven:

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}$$

Hvor x_i er den i 'te observation af stikprøven, n er antallet af observationer i stikprøven, \bar{x} er stikprøvens middelværdi.

2.1.4 Standardafvigelse

Standardafvigelse er et mål for spredningen af data i en population eller stikprøve. Det er defineret som kvadratroden af variansen. Formlen for standardafvigelse i en population er:

$$\sigma = \sqrt{\sigma^2}$$

Hvor σ er standardafvigelsen og σ^2 er variansen i populationen.

I en stikprøve beregnes standardafvigelsen på samme måde, men variansen er beregnet ud fra stikprøven i stedet for hele populationen. Formlen for standardafvigelse i en stikprøve er:

$$s = \sqrt{\sigma^2}$$

Hvor s er standardafvigelsen og s^2 er variansen i stikprøven.

2.1.5 Fordelingsteori

Normalfordeling

Normalfordelingen er en statistisk fordeling, som har en symmetrisk fordeling med en klokkeformet kurve. Den beskrives af følgende formel:

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, -\infty < x < \infty$$

Hvor μ angiver midterpunktet og σ angiver spredningen, som er to kendte værdier. Som vist på figuren, er 68%, 95% og 99% af observationerne henholdsvis indenfor $\sigma \pm$, $\pm 2\sigma$ og $\pm 3\sigma$ fra middelværdien μ . Z-scoren er et statistisk mål, som angiver, hvor mange standardafvigelser en observation ligger fra middelværdien. Dette beregnes ved at benytte følgende formel:

$$Z = \frac{X - \mu}{\sigma}$$

Værdierne i Z-scoren kan bruges i en standard normalfordeling, hvor $\mu = 0$ og standardafvigelsen er $\sigma = 1$. En positiv Z-score angiver, at observationen er over middelværdien, mens en negativ Z-score angiver, at observationen er under middelværdien i forhold til standardafvigelsen.

Uniformfordeling

Uniformfordeling er en statistisk fordeling, hvor alle værdier har samme sandsynlighed for at optræde. En uniform fordeling kan repræsenteres grafisk som en lige linje, der går fra den ene ende af intervallet $[A, B]$ til den anden, hvor højden er $1/(B - A)$.

En uniform fordeling har følgende formel:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B, \\ 0 & \text{ellers} \end{cases} \quad (2.1)$$

Her er $f(x)$ sandsynligheden for, at en tilfældig variabel x har en værdi i intervallet $[A, B]$. For at kunne arbejde med uniform fordeling er det nyttigt at have kendskab

til hvordan middelværdi og spredning beregnes. Middelværdien for en uniform fordeling er gennemsnittet af de to endepunkter i intervallet og kan beregnes ved hjælp af følgende formel:

$$\mu = \frac{A + B}{2}$$

Når det kommer til spredningen for en uniform fordeling kan den findes ved at bruge følgende formel:

$$\sigma^2 = \frac{(B - A)^2}{12}$$

Skævfordeling

Skævfordeling er en statistisk fordeling, hvor dataene er skævt fordelt i forhold til middelværdien. Dette betyder, at der er flere observationer i den ene ende af fordelingen end i den anden. Dette koncept kan forstås som enten højreskæv eller venstreskæv. En højreskæv fordeling betyder, at den mest hyppige værdi forekommer på venstre side af middelværdien, mens halen strækker sig mod højre. Dette betyder, at der vil være flere ekstreme værdier på den høje side af fordelingen, og middelværdien vil være større end medianen. For venstreskæve fordelinger gælder det modsatte.

Poissonfordeling

Poissonfordelingen er en statistisk fordeling, der beskriver sandsynligheden for, at en bestemt begivenhed indtræffer et bestemt antal gange inden for en given tidsperiode eller i et givet område. Fordelingen antager, at hændelserne opstår uafhængigt af hinanden, og at sandsynligheden for, at en hændelse opstår, er den samme i hele tidsintervallet. Poissonfordelingen har den egenskab, at middelværdien og variansen er lig med λ . Poissonfordelingen er givet ved:

$$p(x; \lambda t) = \frac{(e^{-\lambda t} \lambda t^x)}{x!}, x = 0, 1, 2, \dots,$$

Her er x er forekommer inden for et givent tidsinterval, λ er forventningen om antallet af forekomster i det samme tidsinterval, og $x!$ er det faktiske antal hændelser af x .

Store tals lov

Store tals lov bruges til at forstå og analysere de forskellige typer af fordelinger som nævnt ovenover; normalfordeling, uniform fordeling skæv fordeling og poissonfordeling. Det grundlæggende princip er, at gennemsnittet af et stort antal

uafhængige fordelte stokastiske variable vil konvergere mod den teoretiske forventede værdi. Hvilket kan bruges til at estimere populationens middelværdi ud fra et stort prøveudtræk.

CLT

CLT(Central Limit Theorem) er en sætning som lyder på, hvis der bliver taget en tilfældig stikprøve n , fra en population, vil middelværdierne ligne en normalfordeling. Så jo større n der er jo mere vil den ligne normalfordelingen. Generelt set er det også kendt at over 30 så vil den ligne en normalfordeling. Jo større stikprøven bliver en 30 gør ikke den store forskel i forhold til at ligne normalfordelingen, men man kan være mere præcis med det den fortæller noget om. I CLT siger også at det er ligemeget hvordan populationen fordeler sig, så vil den stadig efterligne en normalfordeling.

CDF OG PMF

Cumulative Distribution Function (CDF), er i statistik, når man tager sandsynligheden af, at et givet x vil være $\leq x$. For eksempel, hvis man kaster en terning, er der $\frac{1}{6}$ chance for at man vil lande på 1, så der er 16.667% chance for at du slår et. Der er også 16.667% chance for at slå 2, men med CDF siger den, at det er den kumulative værdi, så chancen for at du slår 2 eller derunder, hvilket giver en CDF-værdi på 33.33% chance for at slå det. Probability Mass Function (PMF) minder om CDF, men den siger i stedet for, at det skal være det samme eller mindre end x , så skal det være præcis x . Den fokuserer på en specifik værdi, så i stedet for, når man slår 2, vil chancen være det samme som for at slå, altså 16.667%. Den fungerer ikke kumulativ.

2.1.6 Estimationsteori

Statisk inferens

Statistisk inferens anvendes til at drage konklusioner om en populations parametre baseret på en stikprøve taget fra populationen. Inden for begrebet "estimering" benyttes disse stikprøver til at lave to typer estimer, nemlig "punkttestimat" og "intervalestimat". Et punkttestimat anvendes til at estimere en ukendt parameter ud fra en population, for eksempel middelværdien. Et intervalestimat anvendes til at estimere de sande værdi-intervaller for middelværdien for en population. Projektet omfatter konfidensintervaller, da det giver en sandsynligheds værdi af, om det estimerede interval dækker den sande populationsparameter.

Konfidensinterval

Når der estimeres en populationsparameter i statistik ved brug af et punktestimat, er der altid en vis grad af usikkerhed. Denne usikkerhed repræsenterer mængden af variation, der er associeret med vores stikprøve, også kaldet fejlmargen. Derfor er det vigtigt at arbejde med en formel, der tager højde for denne fejlmargen, hvilket konfidensintervaller gør. Et konfidensinterval giver en sandsynlighed for, at intervalestimat har fanget den populationsparameter, der søges. Det tal kaldes dækningsgraden, og er typisk valgt til at være tættest på 1, eksempelvis enten 0.95 eller 0.99. Der er en risiko baseret på det valgte dækningsgrad. Dette kaldes signifikansniveauet α og beskriver risikoen for ikke at ramme populationsmiddelværdien med det valgte estimering. Når man skal skabe et konfidensinterval for populationsmiddelværdien, er der to mulige formler.

Hvis populationsvariansen er kendt og normalfordelt, skal man bruge følgende formel:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Hvor \bar{x} repræsenterer punktestimatet og $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ repræsenterer fejlmargen.

Hvis populationsvariansen er ukendt og har en t-fordeling, skal man bruge følgende formel:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Hvor \bar{x} repræsenterer punktestimatet, $t_{\alpha/2}$ repræsenterer t-værdien og $\frac{s}{\sqrt{n}}$ repræsenterer standardafvigelsen for stikprøven.

2.2 Teori: PRNG

Tilfældighed refererer til egenskaberne ved et system eller en proces, der mangler forudsigelighed eller orden, og producerer resultater, der ikke kan forudsiges med sikkerhed. Tilfældighed er afgørende i mange felter, såsom statistik, kryptografi og spil, hvor det bruges til at sikre lighed, sikkerhed og forudsigelighed.

Tilfældige tal er tal, der genereres uden nogen forudsigelig mønster og følger en sandsynlighedsfordeling. De anvendes ofte i computersimuleringer, kryptografi og statistisk analyse. Tilfældige tal kan genereres ved hjælp af forskellige teknikker, såsom fysiske processer som radioaktivt henfald, atmosfærisk støj eller matematiske algoritmer, der simulerer tilfældighed.

Pseudo Random Number Generator (PRNG) er metoder som bliver brugt til at skabe tilfældige tal. En computer vil altid være deterministisk og vil derfor aldrig

kunne generere tilfældige tal. Dertil kan der bruges PRNG, hvor der er udarbejdet forskellige metoder, hvorpå man kan genere tal, der så vidt muligt kan give tal, som ligner de er tilfældige. Dette ændrer ikke på at tallene, som algoritmen ender med at producere, ikke er deterministiske, algoritmen får det til at ligne at de er tilfældige, derfor pseudo.

2.2.1 Fordelings Test (χ^2)

For at kunne undersøge, hvorvidt et datasæt genereret af PRNG er uafhængig, sammenlignes det med de forventede værdier, en uniform fordeling. **Antagelserne for en uniform fordeling er, at minimum og maksimum er forhåndsbestemt, samt at alle værdier har samme sandsynlighed for at optræde.** Ud fra antagelserne, kan der opstilles en nulhypotese (H_0), hvilket er en prædikeret hypotese, opstillet på baggrund af antagelserne for fordelingen. Dermed vil nulhypotesen, for at teste en PRNG model, være at der ikke er en statistisk signifikant forskel, mellem frekvensen af værdierne som PRNG har genereret og den teoretiske fordeling af tal.

Signifikansniveauet, angivet med α , er sandsynligheden for at en nulhypotese afvises, selvom nulhypotesen er sand. Inden for statistik bruges der forskellige significansniveauer til at bekræfte ens nulhypotese. Signifikansniveauet vælges på forhånd og ligger ofte omkring 1 – 5%, altså 0.01 – 0.05. Sammen med p-værdien, som er resultatet af testen, afgøres det, hvorvidt nulhypotesen bekræftes eller forkastes, at resultaterne er statistisk signifikante. P-værdien angiver, hvad sandsynligheden er for, at et lignende resultat forekommer, selvom nulhypotesen er sand. Siden stikprøven trækkes tilfældigt fra en population, er der altid en sandsynlighed for at den observerede situation kun opstod på grund af stikprøvefejl, siden en stikprøve muligvis ikke er repræsentativ for hele populationen. For at bestemme antallet af frihedsgrader i en χ^2 test skal dataene opdeles i kategorier, således at fordelingen kan sammenlignes med den forventede fordeling. Antallet af frihedsgrader er lig med antallet af kategorier minus 1.

2.2.2 Spektral Test

Spektral testen er en metode til at kontrollere, hvorvidt en given talrække af pseudo-tilfældige tal er stokastisk uafhængig af hinanden. Ideen bagved metoden er at tage henholdsvis i-tal, og kombinerer dem til i-tupler, som fortolkes som en vektor i i-dimensioner. På baggrund af vektorens fordeling i værdiområdet kan der konkluderes hvor godt fordelingen svarer til, hvad der teoretisk forventes af ligefordelte tilfældige værdier.

Når Spektral testen bruges for at undersøge resultaterne på en PRNG som ikke generer tilstrækkelig tilfældigt tal, kan der opstå planer som tallene følger. I ek-

semplet nedenfor blev algoritmen RANDU testet, og resultatet viser at tallene tydeligt falder i 15 to-dimensionelle planer.

2.2.3 LCG

Der vil i projektet bruges metoden Linear Congruential Generator (LCG), som er en af mange metoder til at skabe pseudo tilfældige tal på.

$$X_{i+1} = (aX_i + c) \mod m$$

Metoden består af en startværdi, X_0 , som overholder $0 < X_0 < m$. Dernæst bliver a , også kaldet multiplikatoren, gange på X_i ledet, hvorefter der bliver adderet en konstant c . Dette bliver sat modulus m , hvor resultatet af det tidligere led bliver divideret med m , hvilket giver de tilfældige tal, samt en ny multiplikator, på baggrund af resten af divisionen. Derudover skal m overholde $0 < m$, a skal $0 < a < m$, og c skal overholde $0 < c < m$. Når metoden LCG bliver brugt, er der chance for, at metoden går i et loop, hvorved den vil give en sekvens af pseudo-tilfældige tal efterfulgt af det samme tal. Dette kan undgås ved at udvælge værdier for a og m , der er indbyrdes primisk, hvilket betyder, at de eneste tal, som går op i dem, er 1 og -1 . Når tallene bliver generet efter ovenstående formel, vil der alt efter a , c og m 's værdier, opstå et tidspunkt, hvor tallene bliver gengivet. Længden fra det oprindelige tal til gengivelsen kaldes perioden.

2.2.4 R's Pseudo-Random Number Generator

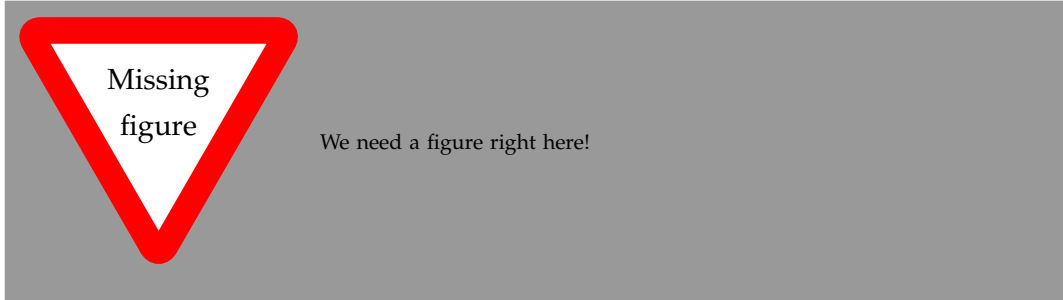
En af de standard PRNG i programmeringssproget R, kaldes for Mersenne - Twister. Mersenne - Twister har en periode på $2^{19937} - 1$ og en ligefordeling i 623 dimensioner. Problemerne med LCG's lave periode og dårlige fordeling af punkterne, var en grund til at mere komplekse PRNG var nødvendig. Mersenne - Twister blev udviklet efter algoritmer som LCG ikke var tilstrækkelig tilfældige længere. For at generer tal der er højt uafhængige og uforudsigelige bruger algoritmen bit-shifting og XOR-operationer. (uddybes/tilrettes så snart der bliver lavet spektraltest for LCG).

Chapter 3

Chapter 2 name

Here is chapter 2. If you want to leearn more about $\text{\LaTeX}2_{\epsilon}$, have a look at [1], [3] and [2].

I think this word is misspelled



Chapter 4

Conclusion

There are probably still some bugs in the theme. If you should find one, then please submit it on <https://github.com/jkjaer/aauLatexTemplates>.

Bibliography

- [1] Lars Madsen. *Introduktion til LaTeX*. <http://www.imf.au.dk/system/latex/bog/>. 2010.
- [2] Frank Mittelbach. *The LATEX companion*. 2. ed. Addison-Wesley, 2005.
- [3] Tobias Oetiker. *The Not So Short A Introduction to LaTeX2e*. <http://tobi.oetiker.ch/lshort/lshort.pdf>. 2010.

Appendix A

Appendix A name

Here is the first appendix