A study on the Universe

Development of a predictive model

Biggie Blackie, Dickie Slickie, Bjarne, Johnny sins skrr
Energy Technology, TEPE4-1005, 2018-06

P2 Projekt

STUDENT REPORT
AALBORG UNIVERSITY

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.

Institut for Datalogi
Aalborg Universitet
http://www.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Rapportens titel

Tema:
Fra data til videnskab

Projektperiode:
Forårssemestret 2023

Projektgruppe:
XXX

Deltager(e):
Forfatter 1
Forfatter 2
Forfatter 3

Vejleder(e):
Søren Byg Vilsen

Oplagstal: 1

Sidetal: 9

Afleveringsdato:
12. marts 2023

Abstract:

Her er resuméet

Contents

Preface

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

Aalborg University, March 12, 2023

Author 1
<username1@XX.aau.dk>

Author 2
<username2@XX.aau.dk>

Author 3
<username3@XX.aau.dk>

Chapter 1

Chapter 2 name

1.1 Sandsynlighed

Noget indledende tekst om sandsynlighed

1.1.1 Udfaldsrum og hændelser

Mængden af alle mulige udfald fra et statisk eksperiment kaldes for *udfaldsrummet*, S, ethvert udfald kalde et element eller medlem af udfaldsrummet. Det vil sige, at udfaldsrummet for kast med en mønt indeholder de to elementer; plat og krone.

En delmængde af udfaldsrummet kaldes en *hændelse*, A, dette noteres som $A \subseteq S$. Ved kast med en terning kan en hændelse, A være, at terningen viser et lige antal øjne. Denne hændelse noteres $A = \{2, 4, 6\}$.

Sandsynligheden for en hændelse

Sandsynligheden for at en hændelse forekommer er andelen af gange, hvor den givne hændelse, A, sker ved gentagelse af eksperiment, dette noteres som $P(A)$. Sandsynligheden for en hændelse er et tal mellem 0 og 1, $0 \leq P(A) \leq 1$. Summen af alle sandsynlighederne for udfaldsrummet er 1, $P(S) = 1$, og derved indikerer en sandsynlighed tæt på 1 en sandsynlighed for denne hændelse.

Hvis der er lige stor sandsynlighed for alle udfald i udfaldsrummet kalde sandsynlighedsfordelingen uniform. Dette er tilfældet ved kast af mønt såvel som kast med terning.

1.1.2 Stokastiske variable

En stokastisk variabel, X, tildeler variable i udfaldsrummet en talværdi. Dette kunne være antallet af plat ved 5 kast med mønt eller summen af øjne ved flere kast med to terninger. Den stokastiske varibel fordeler sig efter en sandsynlighedfunktion, $f(X)$.

Diskrete stokastiske variable

Ved diskrete stokastiske variable antager X kun hele værdier. Det giver eksempelvis ikke mening at tale om en halv plat eller halve øjne på terninger, hvorfor begge disse er eksempler på diskrete stokastiske variable. Sandsynlighedsfunktionen for en diskret stokatisk variabel beskriver sandligheden for at variablen antager en given værdi, $f(x) = P(X = x)$. Dette kaldes også for *massefunktionen*. Herudfra kan variablens fordelingsfunktion, $F(x)$, bestemmes. Denne anvendes til at bestemme sandsynligheden for, at den stokastiske variabel antager en værdi lig med eller mindre end x:

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for} -\infty < x < \infty.$$

Middel Middelværdien kaldes også for den forventede værdi, $E(X)$. Denne beregnes ved brug af sandsynlighedsfunktionen og er et vægtet gennemsnit, hvor hver mulig værdi for X indgår med sin sandsynlighed, $f(x)$. Den forventede værdi beskriver det gennemsnitlige resultat ved mange gentagelser af samme eksperiment.

$$\mu = E(X) = \sum_{x} x \cdot f(x).$$

Varians og standardafvigelse Middelværdien alene kan ikke give en tilstrækkelig beskrivelse af data fordelingen. Man må også have beregreber, der beskriver dataets variabiliet. Den bedste størrelse til netop dette formål er varians, σ^2. Variansen er den gennemsnitlige kvadrerede afstand til middelværdien og er givet ved formlen:

$$\sigma^2 = E[(X - \mu)^2] = \sum_{x} (x - \mu)^2 \cdot f(x).$$

Kvadratroden af variansen kaldes standardafvigelsen.

Kontinuerte stokastiske variable

Kontinuerte stokastiske variable antager værdier på en kontinuert skala. Det vil sige, at variabelværdier ikke er begrænset til heltal.

For kontinuerte stokastiske variable kaldes sandsynlighedfunktionen for *tætheds-funktionen*, denne er defineret for alle $x \in \mathbb{R}$, derudover gælder det, at $f(x) \geq 0$ for alle $x \in \mathbb{R}$, og slutteligt at $\int_{-\infty}^{\infty} f(x)\ dx = 1$.

Sandsynligheden for at den stokastiske kontinuerte variabel X ligger mellem værdierne a og b er beregnes således:

$$P(a < X < b) = \int_a^b f(x)\ dx.$$

Ved integretion af tæthedsfunktionen fås fordelingsfunktionen. Denne er altså givet ved:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\ dy, \quad x \in \mathbb{R}.$$

Denne definition er analog med definitionen af fordelingsfunktionen for diskrete stokastiske variable. Her anvendes integration i stedet for summation.

Middel For beregning af middelværdien for kontinuerte stokastiske variable gælder samme princip, som til beregning af middelværdien for diskrete stokastiske variable. Summation udskiftes ligeledes med integration i dette tilfælde.

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\ dx.$$

Varians og standardafvigelse Variansen for kontinuerte stokastiske variable fortæller det samme, som ved diskrete stokastiske variable, men der anvendes igen integration fremfor summation til at beregne variansen. Formlen ser således ud:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\ dx.$$

1.1.3 Stikprøver

En population betegner hele den gruppe, som ønskes beskrevet med statistik. Udtages kun en delmængde af denne gruppe, tales i stedet om en stikprøve. En stikprøve indeholder et antal elementer, som uafhængigt og tilfældigt er udvalgt fra populationen. Dette gøres på denne måde for at undgå bias og betyder, at sandsynligheden ved udtagningen af et element ikke påvirkes af værdien af de foregående observationer. Stikprøverne kan bruges til at drage konklusioner om populationen, eksempelvis kan middelværdien i stikprøven bruges til at sige noget om middelværdien i populationen.

Stikprøvens middelværdi I stikprøven kan middelværdien beregnes som gennemsnittet af alle stikprøvens elementer:

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i.$$

Størrelsen $\bar{X}$ er en stokastisk variabel, og det gælder, at $E(\bar{X} = \mu)$.

Standardafvigelse for $\bar{X}$ kaldes *standardfejlen* og er givet ved: $\frac{\sigma}{\sqrt{n}}$, hvor σ er populationens standardafvigelse og n er stikprøvestørrelsen. Denne definition fører til, at standardfejlen bliver mindre desto større stikprøve.

1.1.4 Normalfordelingen

Normalfordelingen er en kontinuert distribution med tæthedsfunktion:

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

Den klokkeformede graf, som denne funktion giver, er altså et afgørende karakteristika for denne fordeling.

Normalfordelingen er afhængig af to parametre, som er middelværdien μ og standardafvigelsen σ. *Standardnormalfordelingen* er et særtilfælde af normalfordelingen for middelværdien er 0 og standardafvigelsen er 1. Hvis den stokastiske varibel Y følger en normalfordeling med parametrene μ og σ bruges notationen: $Y \sim norm(\mu, \sigma)$.

Z-score

Enhver normalfordelt stokastisk variabel kan standardiseres ved brug denne formel:

$$Z = \frac{Y - \mu}{\sigma},$$

hvor Y er den stokastiske variabel, μ er populationens middelværdi, og σ er populationens standardafvigelse.

Denne standardisering betyder, gør Z til en stokastisk variabel, der følger standardnormalfordelingen, og at z repræsenterer antallet af standardafvigelser, y ligger fra μ.

Central limit theorem

Når tilfældige stikprøver udtages fra en population med en hvilken som helst fordeling, vil det ses, at fordelingen af stikprøvens middelværdi approksimerer normalfordelingen, når stikprøvestørrelsen går mod uendeligt. Dette noteres som $\bar{X} \approx norm\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ Afhængigt af fordelingen i populationen, holder central limit theorem som regel for stikprøver med 30 eller flere observationer.

Da $\bar{X}$ derved er en normalfordelt stokastisk variabel, kan denne ligeledes standardiseres ved at trække middelværdien fra og dele med standardafvigelse, her standardfejlen, som er givet ved $\frac{\sigma}{\sqrt{n}}$:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

1.1.5 t-fordeling

t-score

1.1.6 Chi i anden-fordelingen

Chi i anden-fordelingen er et særtilfælde af gammafordelingen, som kun afhænger af antallet af frihedsgrader, v, idet $\mu = v$ og $\sigma^2 = 2v$. Fordelingsfunktionen er givet ved:

$$f(x; v) \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} \cdot x^{v/2-1} \cdot e^{-x/2}, & x > 0, \\ 0, & \text{ellers.} \end{cases}$$

Det ses heraf, at gammafunktionen indgår i foreskriften, og denne har følgende definition:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \cdot e^{-x}\, dx, \quad \text{for } \alpha > 0.$$

Appendix A

Appendix A name

Here is the first appendix