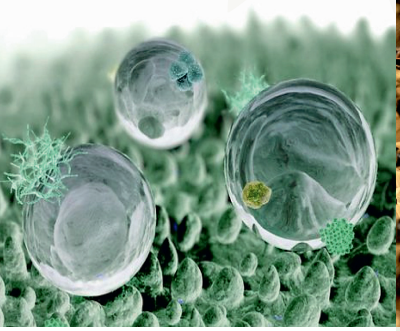
Jens Ledet Jensen

STATISTIK
viden fra data



Aarhus Universitetsforlag

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Statistik
viden fra data

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Statistik
viden fra data

JENS LEDET JENSEN

Aarhus Universitets*forlag*

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Statistik – viden fra data
Omslag: Jørgen Sparre
© Jens Ledet Jensen og Aarhus Universitetsforlag
Bogen er sat med Kp-Fonts af Lars Madsen

ISBN 978 87 7124 250 8

Aarhus Universitetsforlag
www.unipress.dk

Aarhus
Langelandsgade 177
8200 Aarhus N

København
Tuborgvej 164
2400 København NV

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Indhold

Forord **v**

1 Den basale statistiske tankegang **1**
Mendels ærteeksperiment

1.1 Uddybning: Sandsynligheder og stokastiske variable 4
1.2 Uddybning: Binomialfordelingen 7
1.3 Proteinstruktur 9

2 Poissonmodellen og konfidensinterval **11**
Erlangs telefoncentral

2.1 Uddybning: Approksimation til poissonsandsynligheder . . . 18
2.2 Uddybning: Antal betydende cifre 21
2.3 Testkatalog: Poissonfordelte data 21

3 Binomialmodellen og *p*-værdi **23**
Pige- og drengefødsler

3.1 Uddybning: Middelværdi og varians 28
3.2 Testkatalog: Binomialfordelte data 35

4 Normalfordelingen **37**
Diffusion og brownsk bevægelse

4.1 Uddybning: Middelværdi og varians af kontinuert stokastisk variabel 43
4.2 Uddybning: Generel normalfordeling 45

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

5 Goodness of fit test **49**
Lamberts fejlkurve
5.1 Uddybning: $\chi^2(f)$-fordelingen 56
5.2 Uddybning: Multinomialfordelingen 56
5.3 Historien om de knapt så "fede" unge 61
5.4 Testkatalog: Multinomialfordelte data 67

6 Ét normalfordelt observationssæt **69**
Cavendishs måling af jordens massetæthed
6.1 Fraktilsammenligning 72
6.2 Skøn over middelværdi og varians 75
6.3 Test og konfidensinterval for middelværdien 78
6.4 Uddybning: t-fordelingen 80
6.5 Konfidensinterval for variansen 82
6.6 Testkatalog: Ét normalfordelt datasæt 84

7 To normalfordelte observationssæt med samme varians **87**
Placeboeffekten
7.1 Placeboeffekten 88
7.2 Uddybning: Konfidensinterval for forskel i middelværdien . . 92
7.3 Uddybning: Konfidensinterval for fælles varians 94
7.4 Uddybning: Parret sammenligning 95

8 To normalfordelte observationssæt med forskellig varians **99**
Måling af lysets hastighed
8.1 Målte Michelson og Newcomb den samme lyshastighed? . . . 104
8.2 Uddybning: Konfidensinterval for forskel i middelværdien . . 106
8.3 Uddybning: Udbredelse af fejl 109
8.4 Kombination af to middelværdiskøn 114
8.5 Testkatalog: To normalfordelte datasæt 117

9 Lineær regression **121**
Hubbles lov
9.1 Estimation og modelkontrol 124
9.1.1 Uddybning: Udledning af bedste rette linje 127
9.1.2 Uddybning: Fordeling af estimater 128
9.2 Test og konfidensintervaller 129
9.3 Regression med kendt skæring 132
9.4 Linjens værdi i t_* 133
9.5 Misbrug af den lineære sammenhæng 134

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

9.6 Multipel regression . . . 135
9.7 Testkatalog: Lineær regression . . . 135

10 Generel lineær model 139
Diamonds are forever

10.1 Data . . . 139
10.2 Faktorer . . . 143
10.3 Generel middelværdimodel . . . 145
10.3.1 Estimater . . . 146
10.4 Har jeg råd til en diamant med et højere karattal? . . . 147
10.4.1 Generel lineær model: Hypotese og test . . . 149
10.5 Statistik-programpakke . . . 154
10.6 Output fra programpakke . . . 155
10.6.1 Parameterestimater . . . 158
10.7 Tosidet variansanalyse . . . 160
10.8 Afsluttende bemærkninger . . . 163
10.9 Testkatalog: Generel lineær model . . . 164

Appendikser 165

A Udregninger i R 167

B Fordelinger i R, MATLAB og Excel 179

C Tabeller 181
C.1 Standard normalfordelingsfunktion . . . 181
C.2 Fraktiler i standard normalfordeling . . . 184
C.3 Fraktiler i χ^2-fordeling . . . 185
C.4 Fraktiler i t-fordeling . . . 186
C.5 Fraktiler i F-fordeling . . . 187

D Opgaver 191

Indeks 239

Oversigt over testkataloger

Poissonfordelte data . . . 21
Binomialfordelte data . . . 35
Multinomialfordelte data . . . 67
Ét normalfordelt datasæt . . . 84
To normalfordelte datasæt . . . 117
Lineær regression . . . 135
Generel lineær model . . . 164

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Forord

Denne bog er beregnet til et syv-ugers introducerende kursus i statistik. Kurset består af forelæsninger og øvelser, hvor opgaverne i bogen regnes. Det forudsættes, at deltagerne har haft et indledende matematikkursus. Bogen beskriver en række simple statistiske modeller og inferens i disse. Hver model introduceres gennem et datasæt og en lille baggrundshistorie. De fleste af kapitlerne afsluttes med et afsnit med titlen *Testkatalog*, hvor de test, der er indført i kapitlet, gengives på tabelform. Bogen blev første gang trykt i 2007 og har gennemgået flere mindre revisioner.

Bogen afviger på ét punkt fra normal dansk typografi: decimaladskillelsen er punktum i stedet for komma. Endvidere anvendes ■ til at markere afslutningen af eksempler.

Gennem hele processen med at skrive denne bog har Lars Madsen været en stor støtte gennem sin kompetente og kontante hjælp vedrørende alle LaTeX-aspekter, såvel som ved sit skarpe blik for udseende og stil.

Bogen er tilegnet mine døtre Elise og Thea.

Til Indhold

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

Kapitel

1 Den basale statistiske tankegang

– Mendels ærteeksperiment

Synopsis 1.1.
Den basale statistiske tankegang præsenteres gennem et eksempel. Undervejs introduceres sandsynligheder og binomialfordelingen.

Gregor Mendel (1822–1884) (portræt i Figur 1.1) var en østrigsk præst, der i en stor del af sit liv virkede som lærer og i den sidste del af sit liv fungerede som abbed på Sankt Thomas klosteret i Brunn (Tjekkiet). Derudover dyrkede han ærteplanter ! I dag anses Mendel som grundlæggeren af genetik, selvom hans arbejde ikke blev anerkendt af samtiden. Mendel blev født den 22. juli 1822, og hans forældre var bønder. I 1843 begyndte han at studere ved Augustinerordenen i Brunn og blev præst i 1847. Han arbejdede derefter som lærer afbrudt i perioder af lærerstudier ved Wiens Universitet i matematik og biologi. I 1868 blev han abbed. I sit arbejde med ærteplanter fandt han blandt

Figur 1.1: Gregor Mendel (1822–1884)

Til Indhold

Dette materiale er ophavsretligt beskyttet og må ikke videregives.

andet, at egenskaber kan være styret af to alleler af et gen, hvor den ene allel kan være dominant, og den anden kan være recessiv. I artiklen "*Versuche über Pflanzenhybriden*", Verhandlungen des naturforschenden Vereines in Brünn, 1865, studerede Mendel syv egenskaber ved ærteplanten, se Tabel 1.1.

Tabel 1.1: Egenskaber ved ærteplanter.

Karaktertræk	Dominant	Recessiv
Form af modent frø	Glat	Rynket
Farve af frøhvide	Gul	Grøn
Farve af frøskal	Grå	Hvid
Form af moden bælg	Udspilet	Indsnævret
Farve af umoden bælg	Grøn	Gul
Position af blomster	Midtplaceret	Topplaceret
Længde af stængel	Lang	Kort

Specifikt betragtede Mendel farven på den umodne ærtebælg. Hvis farven styres af de to alleller A og a, hvor A er dominant og a er recessiv, vil genotyperne AA og Aa give grønne bælge, og genotypen aa vil give gule bælge. Hvis man krydser grønne med grønne i mange generationer, vil man få en population, der næsten udelukkende består af genotypen AA. Krydser man nu grønne AA-er med gule aa-er, vil man få en population af grønne Aa-er. Hvis denne forståelse er rigtig, mente Mendel, at en krydsning af Aa-er med sig selv skulle give cirka lige mange af de fire genotyper AA, Aa, aA og aa og dermed en udspaltning i forholdet 1:3 af gule i forhold til grønne. I et af sine forsøg fik Mendel 580 nye planter, der fordelte sig som:

Gule	Grønne	Total
152	428	580

(1.1)

Det spørgsmål, vi skal overveje, er, om data bekræfter Mendels hypotese om en 1:3 udspaltning.

Hvis vi gentager forsøget og krydser 580 andre Aa med Aa, forventer vi ikke at få præcist det samme resultat som i (1.1). Vi kalder sådanne udsving for *tilfældige variationer*. Vi kan tydeliggøre dette ved at sammenligne med møntkast. Vi forventer at få cirka lige så mange krone som plat. Dette betyder dog ikke, at hvis vi kaster mønten 10 gange, så får vi krone 5 gange og plat 5 gange. Vi ved godt, at på grund af tilfældigheder vil vi i 10 kast nogle gange få 5 krone, andre gange 4 krone, atter andre gange få 6 krone, og så videre. Faktisk er det muligt, at vi ved 10 kast slet ikke får nogen krone. Hvis vi skal

Til Indhold

Dette materiale er ophavsretligt beskyttet og må ikke videregives.