

Opgavesættet indeholder fire opgaver med i alt ti delopgaver. Delopgaverne vægtes éns, bortset fra delopgave 3(a), som vægtes dobbelt. Delopgavernes samlede vægt er angivet ved hver opgave.

Held og lykke!

Opgave 1

Denne opgave vægter med 2/11

Hvordan kan strømmen af cyklister på en cykelsti modelleres? Byrådsmedlem Bent J. ville gerne finde ud af det og registrerede derfor med et stopur tiden, der gik mellem forbisusende cyklister nedad Langelandsgade ved busstoppestedet "Universitet" en lørdag morgen mellem kl. 10:00 og 10:10. I alt kom der 49 cyklister forbi, og den gennemsnitlige tid mellem to cyklister var 12.2 sekunder. I litteraturen anføres ofte, at eksponentialfordelingen er en god model til disse ventetider mellem cyklister. For at undersøge dette nærmere har Bent grupperet tiderne som vist i følgende tabel. Tabellen indeholder også de forventede antal under eksponentialfordelingen med middelværdi 12.2.

Tabel 1: Antal ventetider som falder i et tidsinterval (sekunder)

Interval	0 – 2	2 – 4	4 – 6	6 – 10	10 – 15	15 – 25	> 25
Observeret antal	6	9	7	9	6	7	5
Forventet antal	7.4	6.3	5.3	8.4	7.3	8.0	6.3

- (a) Gør rede for, at det forventede antal ventetider af en længde mellem 2 og 4 sekunder er 6.3, under modellen hvor tiderne følger en eksponentialfordeling.
- (b) Lav et test for nulhypotesen, at data stammer fra en eksponentialfordeling. Beregn p -værdien, og angiv konklusionen.

Opgave 2

Denne opgave vægter med 2/11

Bageriet Sundbrød producerer rugbrødet "Motion Mette" i 1000 g pakker. Pakkernes vægt antages at være normalfordelt med middelværdi 1000 g og en standardafvigelse på 20 g. Firmaet er sikker på, at middelværdien overholdes, men er bekymret for, at der er for meget variation i vægten – dette kunne give anledning til reklamationer. Derfor tages jævnligt tilfældige stikprøver på 100 pakker, og den empiriske varians s^2 bestemmes for stikprøven. Hvis s^2 overstiger 500 g^2 tvivler bageriet på, at standardafvigelsen ligger på 20 g.

- (a) Opskriv modellen til pakkernes vægt formelt, og formulér nul- og alternativhypotesen i bageriets problemstilling. Angiv teststørrelsens fordeling under nulhypotesen.
- (b) Beregn sandsynligheden for type 1 fejl i firmaets test.

Opgave 3

Denne opgave vægter med 2/11

I et eksperiment lavet på en skole skulle undersøges, om lysfarven påvirker plantevækst. Atten bønner blev sået i pletter med en bønne per plette, og pletterne tilfældigt delt i to lige store grupper. Den ene gruppe blev udsat for rødt lys, og den anden for grønt lys. To uger efter spiring blev skudhøjden målt.

Datasættet `beans.csv`, som udleveredes med digital eksamen, indeholder variablen `growth` med skudhøjden i cm, og variablen `color` med lysfarven.

- (a) Undersøg ved hjælp af et permutationstest, om der er forskel mellem de to grupper. Brug forskellen mellem medianerne som teststørrelse. Husk at angive den faglige konklusion.

Opgave 4

Denne opgave vægter med 5/11

Når der skal bygges motorveje eller landeveje i Florida, inviterer Department of Transportation (DOT) forskellige byggeentreprenører til at afgive tilbud. Entreprenørerne skal sende deres tilbud i en forseglet konvolut, og dem der byder den laveste pris vinder anlægskontrakten. Ingeniører som er ansat i DOT beregner altid et estimat på omkostningerne i forvejen, og i de fleste tilfælde passer den fint med den endelige pris. I 1970'erne og '80'erne kunne der dog registreres flere tilfælde, hvor byggeentreprenørerne havde manipuleret prisen opad ved hemmelige pris-aftaler (price-fixing).

Formålet med denne opgave er at finde en model, der forudsiger den endelige pris ud fra manipulationsforhold og ingeniørernes estimat.

Det udleverede datasæt^a `FloridaRoads.csv` er indsamlet af Floridas justitsministerium i 1980'erne og indeholder 235 observationer med fem variabler:

- `cost` : pris i 1000 US dollars,
- `logcost`: logaritmen til `cost`,
- `D0Testimate`: prisen som DOTs ingeniører har estimeret (i 1000 US\$),
- `logD0Testimate`: logaritmen til `D0Testimate`,
- `bid`: en faktorvariabel med niveauer `competitive` og `fixed`. Niveauet `fixed` betyder at prisen var manipuleret. I alt var 50 af de 235 observationer manipuleret, mens 185 overholder reglerne.

Lad Y_{gj} betegne den endelige pris, og x_{gj} estimatet som ingeniørerne beregnede, hvor $g = 1, 2$ betegner gruppen ifølge `bid` (1: `competitive`, 2: `fixed`), og $j = 1, \dots, n_g$ nummererer observationen indenfor gruppen, $n_1 = 185$ og $n_2 = 50$. Som udgangspunkt betragter vi to modeller, henholdsvis for de oprindelige og for de log-transformerede data:

$$\begin{aligned}\tilde{M}_1 : Y_{gj} &\sim N(\tilde{\alpha}_g + \tilde{\gamma}_g x_{gj}, \tilde{\sigma}^2) \\ M_1 : \log(Y_{gj}) &\sim N(\alpha_g + \gamma_g \log(x_{gj}), \sigma^2)\end{aligned}$$

- Indlæs datasættet og fit modellen \tilde{M}_1 . Beregn residualerne, og undersøg ved en grafisk analyse, om de er normalfordelt og om variansen af Y_{gj} afhænger af x_{gj} . Gør det samme med modellen M_1 . Hvilken af de to modeller passer bedst til data?
- Opstil model M_2 for $\log(Y_{gj})$, hvor hældningen ikke afhænger af, om buddet var manipuleret (angiv som formel). Eftersis ved et test, at modellen M_1 kan reduceres til M_2 . Fit modellen M_2 og angiv alle estimerede parameter eksplicit, inklusive variansen, med de betegnelser du brugte i formlen.
- Eftersis, at modellen M_2 ikke kan reduceres yderligere.
- Beregn et 95%-konfidensinterval for den fælles hældning i M_2 , og test, om hældningen kan antages at være lig med 1.
- Brug modellen M_2 til at beregne et 90% prædiktionsinterval for prisen Y (i 1000 US\$), når ingeniørerne forudsiger omkostninger til at være 3 millioner US\$ og prisen ikke er manipuleret. (Vink: 3 mio = 3000 · 1000)

^aoprindelige data taget fra Mendenhall & Sincich (2012): *Regression Analysis*, 7th edition, Pearson / Prentice Hall, Boston.