Helle Sørensen, Anders Tolver (MATH, KU)
Efterår, 2021

Opgaver til Statistisk Dataanalyse 1

Opgave HS.28 (Colasmagning)

Tyve personer, der alle mener at de bedre kan lide Coca Cola end Pepsi, har deltaget i et eksperiment. De smagte på begge slags cola og skulle udpege hvilken af smagsprøverne der var en Coca Cola. Tretten personer svarede korrekt. Vi antager at alle 20 personer har samme sandsynlighed for at svare korrekt, og denne sandsynlighed kaldes p.

1. Forklar hvorfor binomialfordelingen kan bruges til beskrivelse af eksperimentet.
2. Bestem et estimat for p.
3. Bestem to 95% konfidensintervaller for p: både det simple konfidensinterval givet ved formel (11.6) og „det forbedrede“ konfidensinterval vha. `prop.test`.
4. Hvilken hypotese svarer til at personerne ikke kan smage forskel på de to colaer? Test hypotesen vha. `binom.test`. Hvad er konklusionen?
5. Hvor mange personer, ud af de 20, skulle have svaret korrekt for at vi ville have forkastet hypotesen?

   Vink: Prøv dig frem med `binom.test`.

Opgave HS.29 (Konkurser)

I en undersøgelse holdt man øje med 600 virksomheder der var repræsentativt udvalgt blandt danske virksomheder. I løbet af et år gik 8 virksomheder konkurs.

1. Bestem et estimat og et 95% konfidensinterval for konkursrisikoen blandt danske virksomheder i det pågældene år.

Man holdt tilsvarende øje med et repræsentativt udvalg af virksomheder fra brancerne *Landbrug, skovbrug og fiskeri* (LSF) og *Bygge og Anlæg* (BA). I LSF gik 4 ud af 493 virksomheder konkurs, i BA gik 10 ud af 497 virksomheder konkurs.

2. Bestem et estimat og et 95% konfidensinterval for *forskellen* i konkursrisikoen mellem de to brancher.
3. Udfør et hypotesetest hvor det undersøges om konkursrisikoen var den samme i to bracher det pågældende år. Husk at angive hypotesen.

Opgave HS.30 (Stress i virksomheder)

I finansiering taler man sommetider om at en virksomhed kan være stresset, fx pga. dårlige aktiekurser eller dårlig omsætning. For en specifik definition af stress har man vurderet 518 virksomheder i to på hinanden følgende måneder. Resultatet følger af tabellen nedenfor:

		Måned 2	
		Stresset	Ikke stresset
Måned 1	Stresset	14	52
	Ikke stresset	51	401

1. Forklar hvorfor den relevantehypotese er en hypotese om uafhængighed snarere end en hpotese om ens sandsynligheder i to binomialfordelinger.
2. Udfør et test for uafhængighed, og forklar hvad resultatet siger om forekomsten af stress i virksomheder: Forekommer stress tilfældigt over tid eller snarere i "i klumper".

Opgave HS.31 (Prædiktion af gæt)

Igen-igen en opgave der bruger gættene på antal punkter, men nu med multipel regression og prædiktion. Data ligger i filerne `punktplot2017.xlsx` og `punktplot2017.txt`.

Husk at det korrekte antal punkter var 86, 142 og 47.

1. Opskriv og fit den multiple regressionsmodel med `figur2` som respons og `figur1` og `figur3` som forklarende variable.
2. Betragt to personer (A og B), og antag at person A gætter 1 højere på både figur 1 og figur 3 end person B. Hvor meget højere vil du forvente at person A gætter på figur 2?
3. Betragt en person der gættede korrekt på både figur 1 og figur 3. Vil det være usædvanligt at han/hun også gætter rigtigt på figur 2?

   Vink: Hvilken slags interval har du brug for? Lav et nyt datasæt med en kommando a la `newData <- data.frame(figur1=**, figur3=**)`

Tidligere opgaver har vist at det er hensigtsmæssigt at arbejde med de log-transformerede variable.

4. Opskriv og fit multiple regressionsmodel med `log(figur2)` som respons og `log(figur1)` og `log(figur3)` som forklarende variable.
5. Betragt to personer (A og B), og antag at person A gætter 20% højere på både figur 1 og figur 3 end person B. Hvor meget procent højere vil du forvente at person A gætter højere end person B?

   Vink: At `figur1` er 10% højere for person A end for person A, betyder at

   $$\frac{\texttt{figur1}_\text{A}}{\texttt{figur1}_\text{B}} = 1.2.$$

   Hvad kan du så sige om differensen $\log(\texttt{figur1}_\text{A}) - \log(\texttt{figur1}_\text{B})$? Tilsvarende for `figur3`. Brug så modellen til beregne den forventede værdi for differensen $\log(\texttt{figur2}_\text{A}) - \log(\texttt{figur2}_\text{B})$. Hvor mange procent svarer det til at $\texttt{figur2}_\text{A}$ er større end $\texttt{figur2}_\text{B}$?
6. Brug den nye model til at svare på spørgsmål 3 igen, og sammenlign resultaterne.
7. Lav modelkontrol for begge modeller (uden og med log-transformation). Hvilket af de to prædiktionsintervaller vil du stole mest på?

Opgave HS.32 (Øjenfarve og kritisk frekvens)

Dette er en omskrivning af November 2015, opgave 2.

Hvis hastigheden hvormed en lyskilde blinker (målt i antal blink per sekund) bliver tilstrækkelig høj, så vil et menneske ikke længere kunne se at lyset blinker, men vil i stedet for opfatte lyset som en konstant lyskilde. Den *kritiske frekvens* for en given person defineres som det højeste antal blink per sekund hvorved personen stadigvæk kan opfatte, at en blinkende lyskilde blinker.

En eksperimentel psykolog undersøgte om den *kritiske frekvens* afhænger af personens øjenfarve. Nedenstående tabel indeholder den *kritiske frekvens* for 19 personer inddelt efter forsøgspersonernes øjenfarve:

Øjenfarve		
Blå	Brun	Grøn
25.7 27.2 29.9 28.5 29.4 28.3	26.8 27.9 23.7 25.0 26.3 24.8 25.7 24.5	26.4 24.2 28.0 26.9 29.1

Data er tilgængelige i filerne `lyskilde.xlsx` og `lyskilde.txt`.

Nedenfor findes output fra en R analyse af dette datasæt. Man kan uden begrundelse antage, at den benyttede model er statistisk valid.

(a) Opskriv en statistisk model der kan bruges til at undersøge om øjenfarven har betydning for en persons *kritiske frekvens*. Fit modellen i R og udfør modelkontrol. Lav herunder en skitse af de relevante figurer og kommentér graferne.

(b) Kan man ud fra datasættet konkludere, at øjenfarven har betydning for en persons *kritiske frekvens*? Husk at angive den relevante hypotese.

(c) Angiv et 95% konfidensinterval for forskellen mellem gennemsnittene for den *kritiske frekvens* i populationerne af personer med henholdsvis brune og grønne øjne.

(d) I hvilket interval ligger den *kritiske frekvens* for 95% af mennesker med blå øjne?

(e) Mellem hvilke øjenfarver er der signifikant forskel i populationsgennemsnittene af den *kritiske frekvens*, når der tages højde for multipel testing via en Bonferroni korrektion? Svaret skal begrundes.