Projekt 1: BMI undersøgelse

Formaliteter, opgavestruktur og forventninger til 1. obligatoriske opgave

Opgaven består af to dele. I første del skal der laves en deskriptiv analyse af data. Anden del handler primært om konfidensintervaller og hypotesetests.

Der er lagt op til, at man skal arbejde med opgaven i små "lette" trin. Opgaven skal i praksis løses ved hjælp af programmet R. Der er udarbejdet R-kode, som gør det nemt at komme i gang med projektet. Koden er dog ikke fuldstændig, og I opfordres til at udforske R samtidig med at I laver projektet. F.eks. kan I arbejde med at lave "pæne" titler til graferne eller benytte R's indbyggede funktioner til beregning af konfidensintervaller og test af hypoteser.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, matematisk notation, samt tekst der beskriver analysens resultater. Relevante grafer og tabeller skal indgå i sammenhæng med teksten - ikke som bilag. Præsenter resultaterne fra jeres analyser på samme måde, som I ville videreformidle dem til andre fagfæller.

Inddel besvarelsen i et underafsnit for hver af de stillede spørgsmål.

Besvarelsen skal afleveres som pdf-fil. R-kode bør ikke indgå i besvarelsen, men vedlægges som bilag (i form af en .R-fil). Besvarelsen samt bilag afleveres under Opgaver på Learn ved: `Projekt 1:BMI undersøgelse`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En normal side udgør 2400 anslag.

Grafer og tabeller kan IKKE stå alene - det er altså vigtigt, at I beskriver og fortolker

outputtet fra R med ord.

Grafer og tabeller indgår ikke i opgørelsen af besvarelsens længde. Det er dog IKKE i sig selv en fordel at medtage mange plots, hvis de ikke er relevante!

I må gerne arbejde sammen i grupper, men besvarelsen af opgaven skal skrives individuelt. Spørgsmål omkring projektet kan rettes til hjælpelæren, se retningslinjerne på siden `Projects` på kursets hjemmeside.

Problemstilling

I dette projekt ser vi på overvægt i Danmark. Overvægt, som måles ved det såkaldte BMI-indeks ("Body Mass Index"), er et stigende problem i den vestlige verden, herunder i Danmark. Overvægt associeres med følgesygdomme som f.eks. forhøjet blodtryk, hjertekarsygdomme, diabetes mv., og har således stor indvirkning på den danske økonomi, hvor udgifterne til sundhedsvæsenet er stigende. Man er derfor nødt til at sætte overvægtsproblemet på dagsordenen, og i den sammenhæng er der flere interessenter. Udover det offentlige er der nærmest en hel industri, der beskæftiger sig med overvægtsproblemer og sund livsstil: fra fitness til TV-programmer, bøger om sund livsstil, sunde fødevarer til producenter af mad og fastfood, osv.

Dels er der behov for en kortlægning af befolkningens BMI, men også en undersøgelse af hvilke faktorer, der har indflydelse på en persons BMI – f.eks. køn, alder og uddannelse. Andre spørgsmål melder sig også: Er der f.eks. sammenhæng mellem BMI og hvor hyppigt man spiser fastfood? Kan man gøre befolkningen mere opmærksom på fastfoods betydning for overvægtsproblematikken? Skal man f.eks. kræve en energi-/kaloriemærkning af fastfood-restauranternes måltider?

BMI er et mål for en persons overvægt, som defineres ved

$$BMI = \frac{vægt}{højde^2},$$

hvor *vægt* er personens vægt i kg og *højde* er personens højde i meter (m). BMI-indekset vurderes ud fra følgende tabel:

BMI	Vurdering
Under 18.5	Personen er undervægtig
Mellem 18.5 og 25	Personen er normalvægtig
Mellem 25 og 30	Personen er moderat overvægtig
Mellem 30 og 35	Personen er svært overvægtig (Fedme klasse I)
Mellem 35 og 40	Personen er svært overvægtig (Svær Fedme Klasse II)
Mindst 40	Personen er svært overvægtig (Ekstremt svær fedme Klasse III)

Ovenstående tabel viser WHO klassifikationen af BMI [1]

Indlæsning af data

Lav en mappe til projektet på din computer. Download materialet til projektet fra Learn og udpak ("unzip") det til den mappe, du lige har lavet.

Åbn derefter data-filen `bmi1_data.csv` (f.eks. i RStudio, File → Open File) for at se filens indhold. Bemærk at første linje indeholder variabelnavne (kaldes en *header*), og at de efterfølgende linjer indeholder de egentlige observationer. Observationerne af de enkelte variable er adskilt af et ';' (deraf `.csv`: "comma separated values", her dog semikolon).

I data-filen er der følgende søjler/variable:

Variabel	Forklaring
Køn (gender)	Respondentens køn
Højde (height)	Respondentens højde i cm
Vægt (weight)	Respondentens vægt i kg
Urbanitet (urbanity)	Størrelsen af den by, hvor respondenten bor
Fastfood (fastfood)	Antal dage pr. år, hvor respondenten spiser fastfood

Nogle af disse variable er kategoriserede variable. Yderligere information om hvordan disse er kodet fremgår af bilag 1 (s. 11).

Åbn filen `bmi1_dansk.R`, som indeholder R-kode der kan bruges til analysen. Først skal "working directory" sættes til den mappe på computeren, hvor filerne til projektet er gemt:

[1] http://www.si-folkesundhed.dk/upload/kap_21_overv%C3%A6gt_og_fedme.pdf Her kan man også læse om fedme som risikofaktor for hjerte-kar sygdomme og diabetes.

```
## I RStudio kan man nemt sætte working directory med menuen
## "Session -> Set Working Directory -> To Source File Location"
## Bemærk: i R bruges kun "/" til separering i stier
## (altså ingen backslash).
setwd("Erstat her med stien til den mappe, hvor projektfilerne er gemt.")
```

Nu kan datasættet indlæses i R med følgende kode:

```
## Indlæs data fra bmi1_data.csv
D <- read.table("bmi1_data.csv", header=TRUE, sep=";", as.is=TRUE)
```

Bemærk at der i R-koden bruges engelsk. Det er generelt ikke en god ide at bruge æøå osv. ved programmering. `D` bliver en "data.frame" (en slags tabel), som indeholder den indlæste data (se R-introen i kapitel 1.5 i bogen).

Beskrivende analyse (descriptive analysis)

Første del af projektet går ud på at lave en beskrivende analyse af data. I en rapport er det vigtigt at præsentere og beskrive data for læseren. Dette kan f.eks. gøres ved hjælp af opsummerende størrelser/nøgletal ("summary statistics") og passende figurer.

En simpel opsummering af det indlæste datasæt fås ved at køre følgende kode:

```
## Dimensionen af D (antallet af rækker og søjler)
dim(D)
## Søjle-/variabelnavne
names(D)
## De første rækker/observationer
head(D)
## De sidste rækker/observationer
tail(D)
## Udvalgte opsummerende størrelser
summary(D)
## En anden type opsummering af datasættet
str(D)
```

a) Lav en kort beskrivelse af datamaterialet: Hvilke variable indgår i datasættet? Er der tale om *kvantitative* og/eller *kategoriserede* variable? (Kategoriserede variable

dukker først op i kapitel 8, men det er bare variable, som inddeler observationerne i kategorier – f.eks. tre kategorier: lav, mellem og høj). Hvor mange observationer er der? Er der manglende værdier for nogen af variablene? Husk at kigge på den udvidede beskrivelse af variablene i bilag 1 (side 11).

Inden vi kan gå videre med analysen skal BMI-værdierne for stikprøven beregnes og tilføjes til datasættet som en ny variabel `bmi`. Dette kan gøres ved at køre følgende kode i R:

```
## Beregn BMI og tilføj som ny variabel i D
D$bmi <- D$weight/(D$height/100)^2
```

Et "density histogram" der beskriver den empiriske tæthed af BMI-værdierne i stikprøven (se kapitel 1.6.1) kan nu laves ved hjælp af følgende kode:

```
## Histogram der beskriver den empiriske tæthed for BMI
## (histogram for BMI normaliseret så arealet er lig 1)
hist(D$bmi, xlab="BMI", prob=TRUE)
```

b) Lav et density histogram for BMI. Beskriv fordelingen af BMI-værdierne i stikprøven ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan BMI være negativ? Er der stor spredning i observationerne?

Bemærk: I en *skæv* fordeling er sandsynlighedsmassen ikke symmetrisk fordelt omkring medianen. For en venstreskæv fordeling gælder der, at den længste hale ligger til venstre for midten (almindeligvis vil gennemsnittet også ligge til venstre for medianen). Tilsvarende gælder der, at for en højreskæv fordeling ligger den længste hale til højre for midten (almindeligvis med gennemsnit til højre for medianen).

En meget anvendelig operation i dataanalyse er at opdele data i delmængder. F.eks. kan funktionen `subset` bruges til at udtage en delmængde i R, se også bemærkningen på side 11.

Benyt den følgende R-kode til at lave to nye deldatasæt, ét med data for kvinder og ét med data for mænd:

```
## Opdel i to deldatasæt (hhv. kvinder og mænd)
Dfemale <- subset(D, gender == 0)
Dmale <- subset(D, gender == 1)
```

Ved hjælp af følgende kode kan der laves separate density histogrammer for BMI-værdierne for de to køn:

```
## Density histogrammer der beskriver den empiriske
## tæthed for BMI for hhv. kvinder og mænd
hist(Dfemale$bmi, xlab="BMI (kvinder)", prob=TRUE)
hist(Dmale$bmi, xlab="BMI (mænd)", prob=TRUE)
```

c) Lav density histogrammer af BMI for hhv. kvinder og mænd. Beskriv de empiriske fordelinger af BMI for mænd og kvinder ud fra disse histogrammer, som i det forrige spørgsmål. Ser der ud til at være forskelle i fordelingen af mænds og kvinders BMI (og hvis ja, hvori består disse forskelle)?

Følgende R-kode laver et boxplot af BMI opdelt efter køn:

```
## Boxplot af BMI opdelt efter køn
boxplot(Dfemale$bmi, Dmale$bmi, names=c("Kvinder", "Mænd"),
        xlab="Køn", ylab="BMI")
```

d) Lav et boxplot af BMI opdelt efter køn. Benyt derefter plottet til at beskrive den observerede fordeling af BMI for kvinder og mænd. Er fordelingerne symmetriske eller skæve? Ser det umiddelbart ud til, at der er forskelle mellem fordelingerne (hvis ja, hvilke)? Er der ekstreme observationer/outliers?

Man kan også beskrive den empiriske fordeling af BMI (herunder for mænd og kvinder hver for sig) ved hjælp af opsummerende størrelser/nøgletal som i følgende tabel:

Variabel: *BMI*	**Antal obs.**	**Stikprøve-gennem-snit**	**Stikprøve-varians**	**Stikprøve-standard-afvigelse**	**Nedre kvartil**	**Median**	**Øvre kvartil**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
Alle							
Kvinder							
Mænd							

For at udfylde de tomme celler i tabellen kan man f.eks. benytte R-kode som følgende (se også bemærkning på side 12 for tricks til udregningerne):

```
## Antal observationer i alt
## (medregner ej eventuelle manglende værdier)
sum(!is.na(D$bmi))
## Stikprøvegennemsnit (ej kønsopdelt)
mean(D$bmi, na.rm=TRUE)
## Stikprøvevarians (ej kønsopdelt)
var(D$bmi, na.rm=TRUE)
## osv.
##
## Argumentet 'na.rm=TRUE' sørger for at størrelsen
## udregnes selvom der eventuelt er manglende værdier
```

e) Udfyld tabellen ovenfor med de opsummerende størrelser for BMI for hele stikprøven og derefter separat for mænd og kvinder. Beskriv hvilken ekstra information, der kan udledes fra tabellen sammenlignet med boxplottet?

Statistisk analyse

Andel del af projektet går ud på at lave en simpel statistisk analyse vedrørende BMI for mænd og kvinder. Der skal opstilles statistiske modeller for BMI. Modellernes parametre skal estimeres, og der skal udføres hypotesetests og beregnes konfidensintervaller.

Konfidensintervaller og hypotesetests

Følgende R-kode kan benyttes til at lave et qq-plot med henblik på at vurdere, om den naturlige logaritme til BMI kan antages at være normalfordelt:

```
## Ny variabel 'logbmi' med log-transformeret BMI
D$logbmi <- log(D$bmi)
## qq-plot for log-transformeret BMI
qqnorm(D$logbmi)
qqline(D$logbmi)
```

I praksis vil der opstå situationer, hvor man f.eks. på baggrund af modelkontrollen *ikke* kan tillade sig at antage, at en models forudsætninger er opfyldte. Da vil man

ofte overveje, om det kunne hjælpe at foretage en transformation af data. Ofte kan en logaritmisk transformation være nyttig. (Se kapitel 3.1.9 i bogen). Bemærk at efter en log-transformation ændres fortolkningen af resultaterne på den oprindelige skala fra at omhandle middelværdien til at omhandle medianen, se afsnit 3.1.9. I dette projekt er der på forhånd blevet vurderet, at det er bedre at gennemføre den statistiske analyse på log-transformeret BMI i stedet for BMI.

f) Opskriv en statistisk model for logaritmen til BMI for hele befolkningen, hvor der ikke skelnes mellem kvinder og mænd (se bemærkning 3.2). Estimer modellens parametre (middelværdi og standardafvigelse). Foretag modelkontrol af de antagede forudsætninger (se kapitel 3 samt afsnit 3.1.8). Idet, konfidensintervaller og hypotesetests her involverer fordelingen af gennemsnit, kan det være nyttigt også at inddrage den centrale grænseværdisætning (sætning 3.14) i argumentationen.

g) Angiv formlen for et 95% konfidensinterval (KI) for middelværdien af logaritmen til BMI for hele befolkning (se afsnit 3.1.2). Indsæt tal og beregn intervallet. Angiv derefter et 95% KI for medianen af BMI for hele befolkningen (se afsnit 3.1.9).

h) Udfør et hypotesetest med henblik på at undersøge, om middelværdien af logaritmen til BMI er forskellig fra $\log(25)$. Dette kan gøres ved at teste følgende hypotese:

$$H_0 : \mu_{\text{logBMI}} = \log(25),$$
$$H_1 : \mu_{\text{logBMI}} \neq \log(25).$$

Dette svarer til at undersøge, om medianen af BMI er forskellig fra 25. Angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Kommenter herunder på, om det kan konkluderes, at mere end halvdelen af befolkningen er overvægtig.

Sammenlig resultaterne af hypotesetestet med outputtet fra følgende R-kode:

```
## T-test for en enkelt stikprøve foretaget på log-transformeret BMI
t.test(D$logbmi, mu=log(25))
```

Vi vil nu undersøge, om der er forskel på BMI for kvinder og mænd.

i) Angiv statistiske modeller for logaritmen til BMI for henholdsvis kvinder og mænd. Foretag modelkontrol af de antagede forudsætninger i de to modeller. Estimer modellernes parametre (middelværdi og standardafvigelse for hhv. kvinder og mænd).

j) Beregn 95% konfidensintervaller for middelværdien af logaritmen til BMI for hhv. kvinder og mænd (se afsnit 3.1.2). Benyt disse til at bestemme 95% konfidensintervaller for medianen af BMI for henholdsvis kvinder og mænd. Udfyld nedenstående tal med konfidensintervaller for de to medianer.

	Nedre grænse af KI	**Øvre grænse af KI**
Kvinder		
Mænd		

Sammenlign resultaterne for kvinder med outputtet fra følgende R-kode:

```
## Udtag data kun for kvinder
Dfemale <- subset(D, gender == 0)
## KI for middelværdien af log-BMI for kvinder
KI <- t.test(Dfemale$logbmi, conf.level=0.95)$conf.int
KI
## Transformer tilbage for at få KI for median BMI for kvinder
exp(KI)
```

k) Undersøg ved et hypotesetest, om der kan påvises en forskel på mænd og kvinders BMI. Opskriv hypotesen og angiv signifikansniveauet α, formlen for teststørrelsen, samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord.

Sammenlign resultaterne fra hypotesetestet med resultaterne af følgende R-kode:

```
## Sammenligning af logBMI for kvinder og mænd
t.test(D$logbmi[D$gender == 0], D$logbmi[D$gender == 1])
```

l) Kommenter om det var nødvendigt at udføre hypotesetestet i forrige spørgsmål, eller om samme konklusion kunne opnås ud fra konfidensintervallerne alene? (Se bemærkning 3.59 i bogen).

Korrelation

I forbindelse med den efterfølgende opbygning af modeller til beskrivelse af BMI, f.eks. i forbindelse med projekt 2, vil vi sætte fokus på yderligere sammenhænge mellem udvalgte variabler.

m) Angiv formlen til beregning af korrelationen mellem BMI og vægt. Indsæt tal og beregn korrelationen (indsæt kun i korrelationsformlen, dvs. sæt kun tre tal ind!). Beregn desuden de resterende parvise korrelationer, der involverer BMI, vægt og fastfood. Lav scatterplots, der illustrerer de parvise sammenhænge mellem disse variable. Vurder om sammenhængen mellem plots og korrelationer er som forventet.

Sammenlig de beregnede korrelationer med resultatet af følgende R-kode:

```
## Beregning af korrelation mellem udvalgte variable
cor(D[,c("weight","fastfood","bmi")], use="pairwise.complete.obs")
```

Bilag 1 Beskrivelse af variable i datasættet og deres kodning

Variabel	Betydning	Kode
Køn (gender)	Respondentens køn	0) Kvinde 1) Mand
Højde (height)	Respondentens højde i cm	
Vægt (weight)	Respondentens vægt i kg	
Urbanitet (urbanity)	"Ud fra din egen opfattelse bor du så i en storby, en by eller uden for bymæssig bebyggelse?"	1) Udenfor bymæssig bebyggelse 2) By med under 10.000 indbyggere 3) By med 10.000-49.999 indbyggere 4) By med 50.000-99.999 indbyggere 5) By med over 100.000 indbyggere
Fastfood (fastfood)	"Hvor ofte spiser du "fastfood" f.eks. McDonalds, Burger King, KFC, Sunset, Subway, Dominos, mad fra benzinstationer, pølsevogne, DSB-kiosker, 7-Eleven, pizzeriaer, o. lign?" **Bemærk** at denne variabel nu er "rekodet" til tilsvarende "antal dage pr. år" (værdien angivet i parentes for hver kategori).	1) Aldrig (0) 2) Sjældnere end 1 gang om året (1.0) 3) 1-11 gange om året (6.0) 4) 1-3 gange om måneden (24.0) 5) 1-2 gange om ugen (78.2) 6) 3-4 gange om ugen (182) 7) 5-6 gange om ugen (286.7) 8) Hver dag (365)

Remark .1 Ekstra tips til R

Dette er en valgfri ekstra bemærkning om R-kodning (ikke nødvendig for at løse opgaven). Der er mange måder hvorpå man kan udtage en delmængde i R:

```
## Ekstra bemærkning om måder at udtage delmænger i R
##
## En logisk (logical) vektor med sandt (TRUE) eller falsk (FALSE) for
## hver værdi i en kolonne i D - f.eks: Find alle kvinder i datasættet
D$gender == 0
## Vektoren kan bruges til at udtage data for kvinderne
D[D$gender == 0, ]
## Alternativt kan man bruge funktionen 'subset'
subset(D, gender == 0)
## Mere komplekse logiske udtryk kan laves, f.eks.:
## Find alle kvinder under 55 kg
subset(D, gender == 0 & weight < 55)
```

Remark .2 Ekstra tips til R

Endnu en bemærkning med ekstra R-tips for de interesserede. Man kan f.eks. lave tabellen mere effektivt med en for-løkke:

```
## Lav en for-løkke med beregning af et par opsummerende størrelser
## og gem resultatet i en ny data.frame
Tbl <- data.frame()
for(i in 0:1){
  Tbl[i+1, "mean"] <- mean(D$bmi[D$gender == i])
  Tbl[i+1, "var"] <- var(D$bmi[D$gender == i])
}
row.names(Tbl) <- c("Kvinder","Mænd")
## Se hvad der er i Tbl
Tbl

## I R er der endnu mere kortfattede måder sådanne udregninger kan
## udføres. For eksempel
aggregate(D$bmi, by=list(D$gender), function(x){
  c(mean=mean(x), var=var(x))
})
## Se flere smarte funktioner med: ?apply, ?aggregate og ?lapply
## og for ekstremt effektiv databehandling se f.eks. pakkerne: dplyr,
## tidyr, reshape2 og ggplot2.

## LaTeX tips:
##
## R-pakken "xtable" kan generere LaTeX tabeller og skrive dem direkte
## ind i en fil, som derefter kan inkluderes i et .tex dokument.
##
## R-pakken "knitr" kan anvendes meget elegant til at lave et .tex
## dokument der inkluderer R-koden direkte i dokumentet. Dette
## dokument og bogen er lavet med knitr.
```