I. Demokrati og økonomisk utvikling *(40%, hver deloppgave teller like mye)*

I denne oppgaven skal vi se på sammenhengen mellom demokrati og økonomisk utvikling. For å måle demokrati skal vi bruke en indeks for graden av liberalt demokrati fra V-Dem-prosjektet.[1] Dette er en indeks som går fra 0 (helt udemokratisk) til 1 (perfekt demokratisk). Økonomisk utvikling skal vi måle ved å bruke BNP-tall som er justert for forskjell i kjøpekraft i ulike land.[2] Vi har også data for antall innbyggere. Dataene dekker alle de landene vi har data for i 2015. Du kan laste ned dataene som enten en kommaseparert eller tabulatorseparert fil fra XX XX eller hente dem fra nettet ved hjelp av kommandoen

```
landdata <-
read.csv('https://www.uio.no/studier/emner/sv/oekonomi/ECON2130/v21/timeplan/data/landdata.csv').
```

a) Last dataene inn i en dataramme i R. Lag et histogram over variabelen `demokrati`. Kommenter fordelingen til variabelen.

Av naturlige årsaker er BNP høyere i større land. Derfor er det ofte nyttig å se på BNP per person.

b) Lag en ny variabel `bnp_pers` i datarammen din med BNP delt på antall innbyggere. Lag et spredningsdiagram (scatter plot) med demokrati-indeksen på x-aksen og BNP per innbygger på y-aksen. Diskuter hva vi kan lese ut av diagrammet.

Vi kan regne et land som demokratisk hvis det har en demokratiindeks over 0.5.

c) Lag en ny variabel `demokratisk` i datarammen din som er lik 1 for land med demokratiindeks over 0.5 og 0 for de under 0.5. Hva er andelen land som er demokratiske i datasettet?

Vi kan anta at datasettet vårt er en trekning av land fra universet av alle tenkelige land.

d) Hvordan vil du estimere andelen demokratiske land i dette universet? Er denne estimatoren forventningsrett?
e) Gjennomfør estimeringen fra oppgave d). Lag et 95 % konfidensintervall for estimatet ditt, og forklar hvordan vi skal tolke dette intervallet.
f) Vi ønsker å undersøke om demokratier er rikere (målt ved BNP per innbygger) enn ikke-demokratier. Sett opp de relevante hypotesene for å teste dette.
g) Gjennomføre en t-test av hypotesen at demokratier er rikere enn ikke-demokratier. Vær klar på hvilke forutsetninger du gjør og hvordan du vil tolke utfallet av testen.

Afghanistan har en demokrati-indeks på 0.23, så det er klassifisert som ikke-demokratisk.

h) Er det riktig å konkludere fra det du fant i oppgave g) at vi kan forvente at BNP per innbygger i Afghanistan vil gå opp hvis landet ble mer demokratisk?

Løsning

---

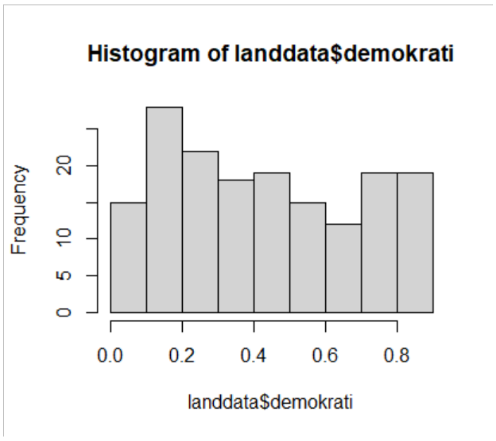
[1] For detaljer se https://www.v-dem.net

[2] Dataserien NY.GDP.MKTP.PP.CD fra World Development Indicators

a) *Fordelingen ligner ikke noen av de fordelingene vi har sett på, ligner mest på en uniform fordeling.*

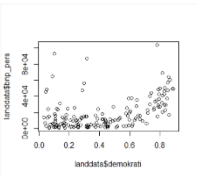
```
#les inn data
library(readr)
landdata <- read_csv("~/landdata.csv")

#plot histogram av demokrati
hist(landdata$demokrati)
```



b)

```
# Lag variabel bnp_pers og plot mot demikrati
landdata$bnp_pers <- landdata$bnp/landdata$befolkning
plot(landdata$demokrati, landdata$bnp_pers)
```



Vi ser at det er en tendens til samvariasjon med vi ser også at det er fattige land som scorer bra på demokrati og rike land som scorer dårlig.

c)

```
#Lag indeks demokartisk, plot for å sjekke og regn ut forventning og standardavvik
landdata$demokratisk <- ifelse(landdata$demokrati>0.5, 1,0)
plot(landdata$demokrati, landdata$demokratisk) #Vi kan plotte denne for å se om det har blitt rett.
mean(landdata$demokratisk)   [1] 0.3892216
```

d) Vi kan estimere andelen i hele universet ved andelen i utvalget. Denne er forventningsrett. Om andelen er μ og X_i er 0-1 indeksen for tilfeldig trukket land i, så er sannsynligheten for å trekke et demokratisk land lik μ, altså er $EX_i = \mu * 1 + (1-\mu) * 0 = \mu$. Da er

$$E\bar{X} = E\frac{\sum X_i}{N} = \frac{\sum EX_i}{N} = \frac{N\mu}{N} = \mu$$

For full uttelling bør kandidaten også begrunne hvorfor estimatoren er forventningsrett.

e)

mean(landdata$demokratisk)   [1] 0.3892216

For KI trenger vi først SD

sd(landdata$demokratisk) 1] 0.4890401

```
# regn ut standardfeil og lag øvre og nedre grense av konfidensintervall.

N<-length(landdata$demokrati)

SE<-sd(landdata$demokratisk)/sqrt(N)

Nedre_95<- mean(landdata$demokratisk)-1.96*SE   [1] 0.3150492

Ovre_95<- mean(landdata$demokratisk)+1.96*SE  [1] 0.463394
```

f) Vi ønsker nå å teste hypotesen H_0 at bnp_pers er den samme for land med demokratisk=1 som de med demokratisk=0.

g) Vi antar at utvalget av 167 land er tilfeldig trukket. Denne forutsetningen er kanskje ikke rimelig, da det er lettere å skaffe data fra rike, folkerike demokratier. De landene som lettere faller ut av et utvalg dersom de er små, diktaturer. Siden vi har relativt mange observasjoner er det rimelig å tenke at gjennomsnittlig bnp_pers i begge grupper er tilnærmet normalfordelt.

```
# t-test om bnp per capita er høyere for demokratisk =1 versus =0

t.test(bnp_pers~demokratisk,data=landdata)

        Welch Two Sample t-test

data:  bnp_pers by demokratisk

t = -5.6826, df = 118.9, p-value = 9.613e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -22159.82 -10707.21

sample estimates:

mean in group 0 mean in group 1

      12203.11        28636.62
```

h) Vi kan ikke tolke en slik årsakssammenheng ut av det. Det kan være at årsakssammenhengen går motsatt vei, at det er mer press mot demokrati i rike land. Eller det kan være en felles årsak, at land i vesten av ulike grunner både har demokrati og er rike. Sammenhengen kan også være spuriøs.

II Evaluering av behandling *(20 %, a og b teller 5% hver, c teller 10%)*

En studie ser på en mulig behandlingsform for en sykdom. Før studien gjennomføres regner vi med at følgende gjelder. Sykdommen er svært vanskelig å behandle så en regner med at bare 2% av foreslåtte behandlingsformer vil virke. Studien rekrutterer nok personer til at styrken er 80%, altså: om behandlingsformen virker har en 80% sannsynlighet for å få et signifikant resultat. Vi krever 5% signifikans for å si at studien har funnet et signifikant resultat.

a) Av 1000 foreslåtte behandlingsformer, hvor mange virker?

b) Hva er sannsynligheten for et signifikant resultat gitt at behandlingsformen ikke virker?

c) Hva er sannsynligheten for at behandlingsformen virker gitt at studien finner et signifikant resultat?

Svar

Med konkreter: a) Av 1000 behandlingsformer som studeres vil 2%, altså 20, ha en effekt. b) Siden kravet for signifikans er 5% er det 5% sannsynlighet for et signifikant resultat om behandlingsformen ikke virker. c) Av de 20 behandlingene som virker vil en finne en signifikant effekt i 80% av tilfellene, dvs 16 av tilfellene. Av de 980 behandlingsformene som ikke har noen effekt vil vi få et signifikant resultat i 5% av tilfellene, dvs i 49 tilfeller. Av de 49+16=65 tilfellene som gir et signifikant resultat er det 16 der behandlingsformen faktisk virker. Altså er sannsynligheten 16/65 = 24,6%.

c) Ved bruk av Bayes formel

$$P(virker|signifikant)$$

$$= \frac{P(signifikant|virker)P(virker)}{P(signifikant|ikke\ virker)P(ikke\ virker) + P(signifikant|virker)P(virker)}$$

$$= \frac{80\% * 2\%}{5\% * 98\% + 80\% * 2\%} = \frac{0{,}016}{0{,}049 + 0{,}016} = 24{,}6\%$$

III. Tillit *(40%, hver deloppgave teller like mye)*

I denne oppgaven bruker vi data fra World Values Survey.[3] Her er en rekke mennesker, totalt 68 895 personer, i mange land blitt stilt en rekke spørsmål. Her skal vi fokusere på to spørsmål, et om tillit og et om korrupsjon. Først har personene blitt spurt om de mener det er lite, middels, eller mye korrupsjon i landet sitt. Videre er de spurt om de mener at vi generelt kan stole på folk flest eller om vi snarere bør være forsiktige når vi er sammen med andre.

		Tillit		
Korrupsjon		Kan stole på folk flest	Bør være forsiktige	Total
	Lite	8391	18378	26769
	Middels	3347	12333	15680
	Mye	3114	23332	26446
	Total	14852	54043	68895

Anta først at de 68 895 personene utgjør hele populasjonen. Vi trekker en tilfeldig person fra denne populasjonen

a) Bruk tallene i tabellen til å finne sannsynligheten for at en person har høy grad av tillit, dvs. svarer at vi generelt kan stole på folk flest. Finn også sannsynligheten for at en person har høy grad av tillitt gitt at vedkommende tror det er lite korrupsjon?

14852/68895=21,5%

8391/18378= 45,6%

Betrakt svarene en person gir på de to spørsmålene som to stokastiske variabler.

b) Kan du da si om de er statistisk uavhengige?

Dersom de er uavhengige skal P(tillit)=P(tillit|lite korrupsjon), men det er ikke tilfellet.

Vi koder om variabelen om korrupsjon til å være 1 når de svarer lite, 2 når de svarer middels og 3 når de svarer mye. Den nye variabelen kalles korrupsjonsscore.

c) Hva blir forventet score på korrupsjon?

*g_korr = (26769*1 + 15680*2 + 26446*3)/68895 = 1,99*

d) Hva er varians og standardavvik for denne variabelen.

Varians = (26769(1-g_korr)^2 + 15680*(2-g_korr)^2 + 26446*(3-g_korr)^2)/68895 = 0,772*

SD = sqrt(varians) = 0,878

---

[3] Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven - Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi.org/10.14281/18241.

Personene i denne studien er selvsagt bare et lite utvalg av verdens befolkning, og i fortsettelsen skal vi ta hensyn til at dette bare er et lite utvalg.

Vi ønsker nå å gjennomføre en statistisk test av en hypotese om at det ikke er noen sammenheng mellom grad av tillit og holdning til korrupsjon. Siden begge variablene er diskrete må vi bruke en annen tilnærming enn vi pleier.

e) Anta du bare kjenner andelen i befolkningen med høy og lav tillit (jfr. spørsmål a) og andelen med hver av de tre nivåene på holdning til korrupsjon. Anta videre at de to variablene er stokastisk uavhengige. Hva er da sannsynligheten for at en person har høy tillit og mener at det er lite korrupsjon?

*Når de to er uavhengige blir P(A og B)=P(A)*P(B), altså er f.eks. P(lite og høy tillit)=P(lite)*P(høy tillit)=0,2155*0,3885 = 0,0838=8,38%. Sannsynligheten for høy tillit finner vi på samme måte som vi fant P(lite) i oppgave a)*

f) Bruk resonnementet fra oppgave e) til å finne sannsynligheten for å være i hver av de 6 cellene i tabellen. Finn forventet antall personer som er i hver celle.

Lite	*Høy*	*Lav*
Middels	*0,083761*	*0,304787*
Mye	*0,049063*	*0,17853*
Mye	*0,08275*	*0,301109*

	Høy	*Lav*
Lite	*5771*	*20998*
Middels	*3380*	*12300*
Mye	*5701*	*20745*

g) La E_i være forventningen du fant i oppgave f) og R_i være det reelle tallet mennesker vi observerer i cellen. Forklar hvorfor testobservatoren

$$A = \sum_{i=1}^{6} (E_i - R_i)^2$$

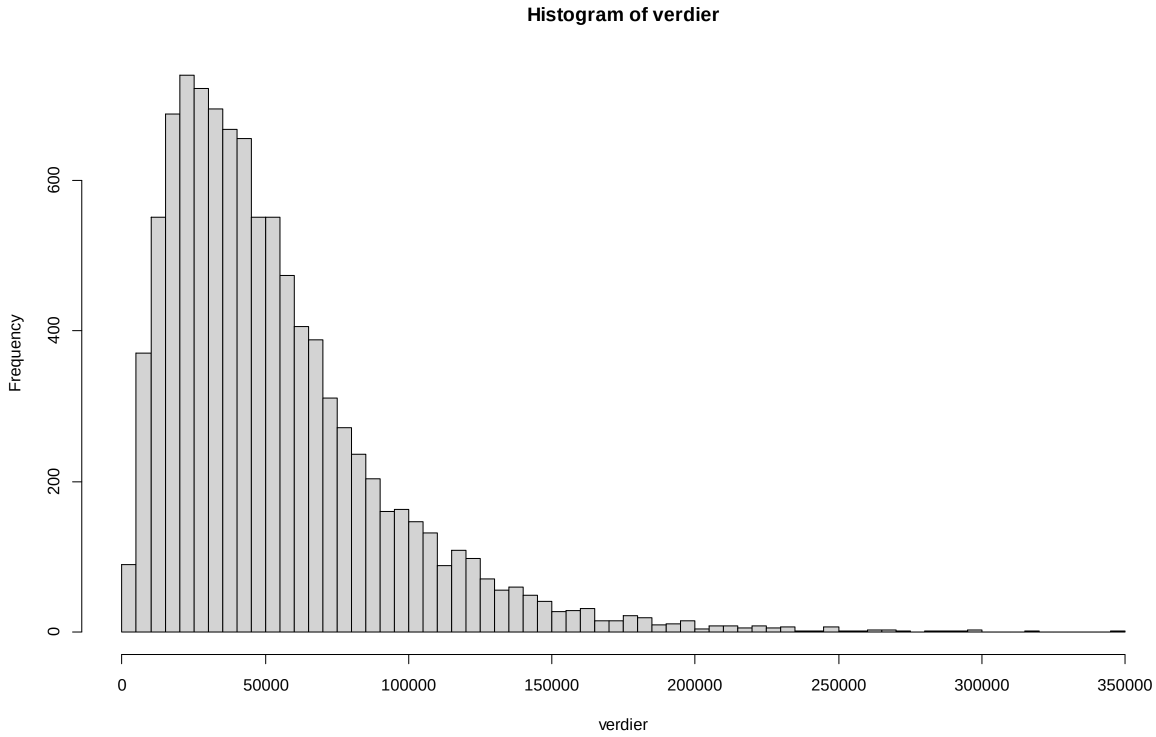
er en meningsfull testobservator for å teste hypotesen at de to stokastiske variablene tillit og holdning til korrupsjon er stokastisk uavhengige.

*Avvik fra prediksjonen er en indikasjon på at modellen ikke er riktig. Ved å kvadrere vil vi alltid få et positivt mål på avviket enten tallet blir for høyt eller lavt. Dette er derfor et rimelig mål på hvor mye vi bommer med prediksjonen. Med de aktuelle data er A=*27 120 005

h) La `h_exp` være vektoren med forventede verdier R_i og `p` vektoren med sannsynlighetene fra oppgave g). Vi kjører følgende R-skript:

```
verdier <- replicate(1e4,sum((h_exp-table(sample(1:6,
             size=68895, prob=p, replace=T)))^2))
hist(verdier, breaks = 100)
```

som gir histogrammet gjengitt nedenfor. Forklar hva dette skriptet gjør og hva det kan brukes til. Hva indikerer det om p-verdien til testen av hypotesen om stokastisk uavhengighet?



Skriptet genererer 10 000 trekninger av testobservatoren A under forutsetningen om stokastisk uavhengighet. Da er sann sannsynlighet for hvert av de 6 utfallene gitt med vektoren p. I hver replikasjon tekker vi N=68 895 personer gitt disse sannsynlighetene, teller opp hvor mange som er i hver gruppe, og summerer kvadratavviket fra vektoren av forventede verdier. Vi ser at den observerte A=27 120 005 er høyere enn noen av de 10 000 simulerte verdier, som indikerer en p-verdi p<0.0001. Men trenger flere simuleringer for å være helt presis på hvor lav den er.