I. CO2-utslipp

I denne oppgaven skal vi se på sammenhengen mellom lands CO2-utslipp og BNP. Vi skal bruke data for CO2-utslipp i tonn per person i 2019 fra *Our world in data* og sette sammen med tall på BNP per person og befolkning fra World Development Indicators.

Komma- og tabulatorseparerte filer er vedlagt denne oppgaven for nedlasting og du kan også hente dataene fra nettet med kommandoen
read.csv('https://www.uio.no/studier/emner/sv/oekonomi/ECON2130/v21/timeplan/data/co2.csv')
I datasettet er det følgende variable

`iso_code`	Landkode
`land`	Landnavn
`co2_per_capita`	CO2-utslipp per person (2019)
`bnp_per_capita`	BNP per person (2019)
`befolkning`	Antall innbyggere (2019)
`oecd`	Landet var medlem i OECD i 2019

a) Last dataene inn i en dataramme i R. Lag et histogram over variabelen `co2_per_capita`. Kommenter fordelingen til variabelen.

```
co2 <- read.csv('co2.csv')
hist(co2$co2_per_capita,breaks = 20)
```

Veldig skjev fordeling, mange med utslipp under 10, noen få opp mot 40

b) Lag et spredningsdiagram (scatter plot) med BNP per person på x-aksen og CO2-utslipp per person på y-aksen. Diskuter hva vi kan lese ut av diagrammet.

```
plot(co2$bnp_per_capita,co2$co2_per_capita)
```

Ganske klar positiv sammenheng. Alle veldig fattige land har lave CO2-utslipp. Ingen rike har helt lave. Men også en del mellominntektsland med høye utslipp, så litt tendens til pukkel-form.

Vi antar at landene i datasettet er et tilfeldig utvalg at alle verdens land.

c) Estimer forventet CO2-utslipp per capita i verdens land. Er estimatoren din forventningsrett og/eller konsistent, og hvilken fordeling har den?

En opplagt estimator på forventningen er gjennomsnittet. Vi finner

```
mean(co2$co2_per_capita)
## [1] 4.618363
```

Gjennomsnittet er en forventningsrett estimator for forventningen. Under generelle betingelser kan vi også bruke store talls lov for å vise konsistens. Det er ikke nødvendig å vise dette. Sidne vi har forholdsvis ange observasjoner er det også naturlig å trekke på sentralgrenseteoremet for å argumentere for at estimatoren er normalfordelt. Det er ikke realistisk å anta normalfordelte data gitt funnene i opg. a)

d) Lag et 90 % konfidensintervall på estimatet fra oppgave c). Vær klar på hvilke forutsetninger du gjør og hvordan du vil tolke intervallet.

Hvis vi antar at estimatoren er normalfordelt, jfr. opg. c), kan vi estimere et konfidensintervall med uttrykket $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$*. Vi finner*

```
qnorm(.95)
## [1] 1.644854
sd(co2$co2_per_capita)
## [1] 5.652454
```

Da blir konfidensintervallet

$$\left[4.62 - 1.645 \times \frac{5.65}{\sqrt{179}}, 4.62 + 1.645 \times \frac{5.65}{\sqrt{179}}\right]$$

som gir [3.93, 5.31].
Med sentralgrenseteoremet blir ikke testobservatoren t-fordelt, men kandidater som bruke t-fordelingen skal ikke straffes for det. Det er også mulig å bruke `t.test` *i R for å beregne konfidensintervallet, men da bør kandidaten gi litt begrunnelse for full skår.*
Tolkning er standard: I gjentatte trekninger av data vil intervaller beregnet på denne måten i 90% av tilfellene dekke den sanne parameteren. Kandidater bør trekkes for å gi utsagn om sannsynlighetsutsagn om den sanne parameteren.

OECD-landene er ofte sett på som en gruppe av rike land. Vi ønsker å teste om CO2-utslippene er høyere i OECD-landene enn i de andre landene.

e) Sett opp de relevante hypotesene for å teste dette.
La forventet utslipp i OECD-land være μ_O og la den være μ_A i de andre landene. Da kan vi bruke
H0: $\mu_O \leq \mu_A$
H1: $\mu_O > \mu_A$
Gitt utsagnet er det naturlig med en en-sidig test. Noe trekk for å sette den opp tosidig.

f) Gjennomfør en t-test av hypotesen at CO2-utslipp per capita er høyere i OECD-land enn i andre land med et signifikansnivå på 1 %. Vær klar på hvilke forutsetninger du gjør og hvordan du vil tolke utfallet av testen.
Igjen trekker vi på at estimatoren er normalfordelt fra opg. c). Da er testobservatoren

$$Z = \frac{\bar{X}_O - \bar{X}_A}{\sqrt{\frac{\sigma_O^2}{n_O} + \frac{\sigma_A^2}{n_A}}} \sim N(0,1)$$

Med et signifikansnivå på 1% forkaster vi H0 hvis vi observerer $Z > 2.33$. Vi finner

$$Z = \frac{7.81 - 3.82}{\sqrt{\frac{12.4}{36} + \frac{33.8}{143}}} = \frac{3.99}{\sqrt{0.581}} = 5.23$$

Siden det er over den kritiske verdien forkaster vi H0.
Det er helt greit, men helt unødvendig, å trekke inn Fischer-Behrens og Welch-tilnærmingen. Det er også mulig å bruke `t.test` *i R, men det krever en del forklaring av hva som foregår for full uttelling.*

Colombia ble medlem av OECD 28. april 2020.

g) Er det riktig å konkludere fra det du fant på oppgave f) at vi kan forvente at CO2-utslippene i Colombia da går opp?
Det vi har funnet over er en korrelasjon, men det er ingen grunn til å gi den en kausal tolkning. Derfor bør vi ikke forvente at utslippene i Colombia går opp som konsekvens av å ha blitt OECD-medlem. Kan legge til at hvis OECD-medlemskap fører til økonomisk vekst kan det føre til økte utslipp, men det har vi ikke egentlig studert kausalt i denne oppgaven.

h) Vi ønsker å gjennomføre testen fra oppgave f) med å bruke medianer i stedet for gjennomsnitt. Drøft om dette er en bedre eller dårligere tilnærming. Finn median CO2-utslipp i OECD-landene og de andre landene.
Med en skjev fordeling kan gjennomsnittet være lite representativt for et «normalt» land, det

kan derfor være hensiktsmessig å se på medianer.
La m_O være den sanne median CO2-utslipp blant OECD-land og m_A medianen blant de andre landene. Da er plausible hypoteser
H0: $m_O \leq m_A$
H1: $m_O > m_A$
Finner

```
median(co2$co2_per_capita[co2$oecd])
## [1] 7.531
median(co2$co2_per_capita[!co2$oecd])
## [1] 1.961
```

For å teste signifikansen til testen basert på median ønsker vi å bruke simulering. Spesifikt vil vi 1000 ganger definere 36 tilfeldige land som «OECD-land» og finne medianen til denne gruppa.

i) Lag et R-skript som gjennomfører denne simuleringen og finn forkastningsverdien for medianen i en test med signifikansnivå 5 %. Gjennomfør testen.

```
set.seed(12345)
median_diff <- function() {
  sim_oecd <- sample(1:179,36) # Trekker 36 tilfeldige land
  return(median(co2$co2_per_capita[sim_oecd])-
median(co2$co2_per_capita[-sim_oecd]))
}
differanser <- replicate(1000,median_diff())
quantile(differanser,probs = .95)
##     95%
## 1.822725
```

Så en en-sidig test med 5% signifkansnivå har forkastning hvis vi observerer en differanse mellom medianene på 1.82. Vi observerte 7.53-1.96=5.57, og kan forkaste H0 på 5% signifikansnivå.

II. Evaluering av emne

I emnet ECON2130 hadde vi en midveisevaluering hvor 22 studenter svarte på noen spørsmål om lærerne, undervisningen, og sin egen opplevelse av kurset. Blant spørsmålene de svare på var om de deltok på spørsmål og svar (QnA) rundene alltid, noen ganger eller aldri. De svarte også på om de hadde problemer med å få utbytte av forelesningene eller seminarer på grunn av mangelfulle forkunnskaper. I tabellen nedenfor kan du se hvordan svarene fordelte seg

Deltok på QnA	**Mangelfullt utbytte på grunn av mangelfulle forkunnskaper**				
		Ja	Til en viss grad	Nei	**Total**
	Aldri	0	3	2	**5**
	Noen ganger	1	5	7	**13**
	Alltid	1	1	2	**4**
	Total	**2**	**9**	**11**	**22**

Vi lager en ny variabel, MangelfulltUtbytte som er 1 om de svarer «Ja», 0 om de svarer «Til en viss grad» og -1 om de svarer «Nei».

a) Hva er forventet verdi av variabelen MangelfulltUtbytte for en tilfeldig trukket respondent.

Svar: (11-2)/22 = 9/22 = 40,9%

b) Hva er forventet verdi av variabelen MangelfulltUtbytte for en tilfeldig respondent, gitt at vedkommende aldri deltok på QnA?

Svar: 2/5 = 9/22 = 40,9%

Foreleserne lurer på om tilbøyeligheten til å delta på QnA forelesninger er ulik for de med mangelfulle forkunnskaper og de som ikke har mangelfulle forkunnskaper.

c) Formuler en hypotese og en test du ville bruke for å svare på dette. (Du skal ikke gjennomføre testen!)

Det er flere måter å løse det på. Vi ønsker å sammenligne to grupper, men det er tre svaralternativ for mangelfulle forkunnskaper. En mulighet er å sammenligne de som svarer ja og de som sier nei, og utelate de som sier «til en viss grad». Eller en kan slå sammen de som sier ja og de som sier «til en viss grad» i en gruppe, da de alle gir uttrykk for at forkunnskapene ikke var ideelle.

Det er tilsvarende tre svaralternativer på om en deltar på QnA. En mulighet er å se på andelen som svarer «alltid». En nullhypotese er at andelen som svarer alltid er den samme for de som svarer Ja og de som svarer Nei. En kan også lage en numerisk variabel av QnA spørsmålet, aldri=0, alltid=2 og mellomalternativet er 1. Nyllhypotesen er at forventet verdi gitt ja er samme som forventet verdi gitt nei.

Det var totalt 151 studenter oppmeldt til eksamen i kurset, og 22 som valgte å svare på spørreundersøkelsen.

d) Diskuter om dette er et tilfeldig utvalg og i hvilken grad dette har betydning for konklusjonene vi kan trekke fra dette datamaterialet.

Dette er ikke et tilfeldig utvalg. Det er de 22 studentene som av en eller annen grunn valgte å svare på spørreskjemaet. Det kan tenkes at de som valgte å svare gjorde det av en bestemt grunn, kanskje fordi de var spesielt misfornøyde med noe og ville gi tilbakemelding om dette. Eller de var ekstra begeistret for kurset og derfor ekstra villige til å svare på henvendelser. Det kan også være de mest pliktoppfyllende studentene. Uansett hva grunnen er, er det sannsynlig at disse studentene ikke er representative. Det betyr at vi ikke kan stole på resultatene uansett utfallet på den statistiske testen.

Oppgave 3

La X og Y være to uavhengige stokastiske variable. X har forventning μ og varians σ^2, mens Y har forventning 2μ og varians $2\sigma^2$. La $Z = aX + bY$.

a) Finn forventning og varians til Z.

$$EZ = a\mu + 2b\mu = (a + 2b)\mu$$

$$VarZ = a^2\sigma^2 + 2b^2\sigma^2 = (a^2 + 2b^2)\sigma^2$$

b) Vi ønsker å bruke Z til å estimere μ. Hvordan vil du velge a slik at Z blir en forventningsrett estimator av μ for hver av tilfellene $b = \frac{1}{2}; b = \frac{1}{3}\ og\ b = \frac{1}{4}$?

$$a + 2b = 1 \quad a = 1 - 2b, \qquad a = 0; a = \frac{1}{3}; a = \frac{1}{2}$$

c) Hva blir variansen til de tre forventningsrette estimatorene du fant i b)? Hvilken estimator vil du foretrekke? Forklar intuisjonen bak hvorfor dette er den beste estimatoren.

$$VarZ = (a^2 + 2b^2)\sigma^2 =$$
$$Når\ b = \frac{1}{2}: \left(0 + \frac{2}{4}\right)\sigma^2 = \frac{\sigma^2}{2} > \frac{\sigma^2}{3};$$
$$Når\ b = \frac{1}{3}: \left(\frac{1^2}{3} + 2\frac{1^2}{3}\right)\sigma^2 = \frac{\sigma^2}{3};$$
$$Når\ b = \frac{1}{4}: \left(\frac{1^2}{2} + 2\frac{1^2}{4}\right)\sigma^2 = \left(\frac{1}{4} + \frac{2}{16}\right)\sigma^2 = \frac{3}{8}\sigma^2 > \frac{3}{9}\sigma^2 = \frac{\sigma^2}{3}$$

$b = \frac{1}{3}$ gir lavest varians.

Intuisjonen er at X og Y har ulik varians, og dette er vektingen som best reduserer total varians

Oppgave 4

Vi kaster to rettferdige terninger der alle utfall fra 1 til 6 er like sannsynlige. La T_1 være antallet øyne opp på den første terningen, mens T_2 er antallet øyne som vender opp på den andre terningen. Endelig er $S = T_1 + T_2$ summen av antall øyne.

a) Hva er sannsynligheten $P(S \geq 5 | T_1 = 3)$?

Dette er sannsynligheten for 2 eller flere øyne, altså alt unntatt 1. Sannsynlighet = 5/6

b) Hva er sannsynligheten $P(S \geq 5 | T_1 \leq 2)$?

Her er enten $T_1 = 1$, og vi trenger 4 eller flere øyne, som har sannsynlighet 3/6. Eller så er enten $T_1 = 2$, og vi trenger 3 eller flere øyne, som har sannsynlighet 4/6. De to er like sannsynlige, altså er (3+4)/12. Det er trolig lettere å se i en tabell: Det er tolv ruter med $T_1 \leq 2$, de grønne. Syv av dem (de mørkegrønne) har en sum lik 5 eller høyere.

	$T_1 = 1$	2	3	4	5	6
$T_2 = 1$	$S = 2$	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12