Kasper Kann
Daniel Behr

05/01/2021

studienr: 20134818
studienr: 20195227

Eksamensopgave 1

Denne eksamensopgave er en teoretisk opgave, hvor det ikke forventes at I bruger R (men I må meget gerne hvis I vil). Mht. selve opgavebesvarelsen er det vigtigt at I ikke blot kan udregne facit, men også kan skrive formlerne op og forklare teorien/principperne bag.

En sælger ved en brugtvognsforhandler er på provision, og når han sælger en bil, får han 4200 kr for en personbil og 4800 kr for en varebil. Han forventer at kunne sælge et antal person- og varebiler pr. dag med flg. sandsynligheder:

Antal varebiler (Y)	Antal personbiler (X)			
	0	1	2	3
0	0.102	0.142	0.061	0.102
1	0.061	0.081	0.102	0.061
2	0.081	0.102	0.061	0.044

1. Hvad er sa
2. Givet at sæ sandsynligehden for at sælger har
3. Beregn det de antal varebiler, sælgeren kan s
4. Beregn sta standardafvigelsen på antallet af
5. Beregn den biler og varebiler, som sælgeren m
6. Er salget a
7. Beregn kov olgte varebiler.
8. Beregn sta n på en dag.

HUSK PLAGIAT

Discret random variable

Eksamen 1

Opgave 1

	Antal Personbiler (X)			
Antal vararbiler (Y)	0	1	2	3
0	0.102	0.142	0.061	0.102
1	0.061	0.081	0.102	0.061
2	0.081	0.102	0.061	0.044

Kasper Kann
Daniel Behr

05/01/2021

studienr: 20134818
studienr: 20195227

Statistik eksamen

Indholdsfortegnelse

Eksamen 1 ..... **2**
Opgave 1 2
Opgave 2 3
Opgave 3 3
Opgave 4 4
Opgave 5 5
Opgave 6 5
Opgave 7 5
Opgave 8 6
EXAM2 ..... **7**
Opgave 1 7
Opgave 2 8
Opgave 3 8
Opgave 4 9
Opgave 5 10

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

Sandsynligheden for at sælge 0 personbiler er givet ved summen af kolonen, hvor der solgt 0 personbiler:

$$0.102 + 0.061 + 0.081 = 0.244 = 24.4\%$$

Konklusion

Opgave 2

Vi skal finde ud af hvad chancen er for at sælgeren sælger 0 varebiler(A), givet at sælgeren allerede har solgt 0 personbiler(B).
Her skal vi bruge formlen:

Conditional probability distribution

$$P(A|B) = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{P(B)}}$$

Vi tager sandsynligheden for at der bliver solgt 0 varebiler (0.102) og dividere med den samlede sandsynlighed(0.244) for at der bliver solgt 0 personbiler.

$$\frac{0.102}{0.244} = 0.4180 = 41.80\%$$

Konklusion

Opgave 3

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

Her skal vi finde det forventede gennemsnit, som beskrives på formlen:

$$E[X] = \mu = \sum_{x} xP(x)$$

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Vi tilføjer en række og en kolonne der summerer sandsynligheden for hele rækken eller kolonnen.

Antal personbiler (X)					
Antal varerbiler (Y)	0	1	2	3	Sum
0	0.102	0.142	0.061	0.102	0.407
1	0.061	0.081	0.102	0.061	0.305
2	0.081	0.102	0.061	0.044	0.288
Sum	0.244	0.325	0.224	0.207	1

marginal probability distribution table

Vi vil nu finde det forventede antal personbiler som sælgeren kan forvente og sælge om dagen:

$$\mu_x = (0 \cdot 0.244) + (1 \cdot 0.325) + (2 \cdot 0.224) + (3 \cdot 0.207) = 1.394$$

x P(x)

personbiler

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

Vi vil nu finde det forventede antal varebiler som sælgeren kan forvente og sælge om dagen:

$$\mu_y = (0 \cdot 0.407) + (1 \cdot 0.305) + (2 \cdot 0.288) = 0.881$$ varechiler

Sælgeren kan altså forvente og sælge 1.394 personbiler og 0.881 varebiler om dagen.

Opgave 4

Vi skal finde standardafvigelsen for personbiler som sælgeren dagligt kan sælge. Her skal vi først finde variansen og så tage kvadratroden af den for at finde standardafvigelsen. Vi vil benytte formlen for variansen σ^2:

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 P(x),$$

som også kan skrives:

$$\sigma^2 = E[X^2] - \mu^2 = \sum_x x^2 P(x) - \mu^2$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Standardafvigelse

$\sigma = \sqrt{\sigma^2}$ pop

$S = \sqrt{S^2}$ Sample

Vi bruger den første version til at finde variansen for personbiler:

$$\sigma_x^2 = (0-1.394)^2(0.244) + (1-1.394)^2(0.325) + (2-1.394)^2(0.224) + (3-1.394)^2(0.207) = 1.14$$

Vi tager nu kvadratroden af variansen for at få standardafvigelsen:

$$\sqrt{1.14} = 1.06$$
$$\sigma_X = 1.06$$

Så vi kan nu sige, at standardafvigelsen for personbiler er 1.06. Dette vil altså sige, at salget pr. dag for personbiler kan afvige med 1.06.

Vi vil nu finde standardafvigelsen for varebiler som sælgeren dagligt kan sælge. Vi vil bruge samme metode som ved personbiler, hvor vi først finder variansen σ^2.:

$$\sigma_Y^2 = (0-0.881)^2(0.407) + (1-0.881)^2(0.305) + (2-0.881)^2(0.288) = 0.680839$$

Vi tager nu kvadratroden af variansen for at finde standardafvigelsen:

$$\sqrt{0.680839} = 0.8251297$$
$$\sigma_Y = 0.8251297$$

Så vi kan nu konkludere, at standardafvigelsen for salg af varebiler er 0.825. Dette betyder, at salget af varebiler pr. dag kan svinge med 0.825.

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

Opgave 5

Her vil vi finde den forventede samlede provision, som sælgeren har på en dag. Her bruger vi vores to sandsynligheder fra opgave 3 og ganger med provisionen for henholdsvis personbiler og varebiler.:

$$\mu_W = \underset{\mu_X \quad a \quad \mu_Y \quad b}{(1.394 \cdot 4200) + (0.881 \cdot 4800)} = 5854.8 + 4228.8 = 10.083.6\, kr.per\, dag$$

Så den forventede samlede provision er på 10083.6kr pr. dag. $\mu_W = E[W] = a\mu_X + b\mu_Y$

Opgave 6

To variable (personbiler og varebiler) er uafhængige, hvis dette udsagn er sandt:

$$P(x,y) = P(x) \cdot P(y)$$

$P(x|y) = P(x)$

$P(y|x) = P(y)$

Tester for udsagn:

$$P_X(0) \cdot P_Y(0)$$

$$0.407 \cdot 0.244 = 0.099$$

$Cov(x,y) = 0$

Vi ved at sandsynligheden er givet ved 0.102 og da:

$$0.102 \neq 0.099$$

Så er X og Y afhængige. Dette betyder, at de kan påvirke hinandens værdier.

Opgave 7

Vi vil nu finde kovariansen, og for at finde kovariansen, der skal middelværdien(mean) for E(XY) findes:

$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$

Formel:

$$cov(X,Y) = E[XY] - \mu_X\mu_Y = \sum_X \sum_Y xyP(x,y) - \mu_X\mu_Y$$

Vi tager udgangspunkt i tabellen fra opgave 3.
Her starter vi med at tage række 1 gange den på de respektive kolloner: 1, 2 og 3, og derefter gør vi det samme med række 2. Række 0 undlader vi, fordi det ikke giver mening at gange med 0.

Kasper Kann
Daniel Behr

05/01/2021

studienr: 20134818
studienr: 20195227

$$E(XY) = 1\big((0.081 \cdot 1) + (0.102 \cdot 2) + (0.061 \cdot 3)\big) + 2(0.102 \cdot 1) + (0.061 \cdot 2) + (0.044 \cdot 3)) = 1.18$$

Her tager vi vores formel i brug.

$$Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y = 1.18 - 1.394 \cdot 0.881 = 1.18 - 1.2281 = -0.0481$$

Kovariansen kan svinge fra $[-\infty; \infty]$, og er ikke ligesom korrelation, som går fra -1 til 1. Derfor kan vi i dette tilfælde sige, at Kovariansen mellem X og Y er en smule negativ, det vil sige at hvis X (personbiler) går op, så går Y (varerbiler) ned.

$$Corr(X,Y) = \delta = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Opgave 8

Vi vil nu finde standardafvigelsen for sælgerens samlede provision for en dag

Vi benytter formlen:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$\sigma = \sqrt{Var(aX + bY)}$$

Vi finder først variansen ved hjælp af formlen:

$Var(aX+bY)$

$$Var(X,Y) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abCov(X,Y)$$

Vi indsætter nu mine tal i formlen:

$Var(aX+bY)$ $\quad 0{,}825$

$$Var(X,Y) = 4200^2 \cdot 1.06^2 + 4800^2 \cdot 0.881^2 + 2 \cdot \big(4200 \cdot 4800(-0.0481)\big) = 35763661$$

Vi indsætter nu variansen, og tager kvadratrodet for at finde standardafvigelsen:

$Var(aX+bY)$ $\quad \sigma = \sqrt{Var(X,Y)} = 5980.27$

Derfor er standardafvigelsen 5980.27 kr for sælgerens samlede provision for en dag.
Det vil altså sige, at sælgeres samlede provision for en dag kan afvige med 5980.27kr.

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

EXAM2

Opgave

Denne eksamensopgave er hovedsageligt en teoretisk opgave, hvor det kun forventes at I bruger R til få ting som tabelopslag og udregning af estimater. Mht. selve opgavebesvarelsen er det vigtigt at I ikke blot kan udregne facit, men også kan skrive formlerne op og forklare teorien/principperne bag.

En batteriproducent oplyser, at levetiden målt i timer for en ny type af batterier er normalfordelt med middelværdi $\mu = 26$ og varians $\sigma^2 = (1.5)^2$. I spørgsmål 1-3 antag at producentens oplysninger er korrekte. Opgaverne **skal** løses vha. standardisering.

- Hvad er sandsynligheden for, at et batteri har en længere levetid end 28 timer?
- Hvilket antal timer er der 95% sandsynlighed for et batteris levetid overstiger.
- Find et interval symmetrisk omkring middelværdien, der med 95% sandsynlighed indeholder levetiden for et batteri.

For at efterprøve producentens påstand udtages en stikprøve på 10 batterier, der udsættes for en standard levetidstest. Resultaterne er angivet i vektoren nedenfor.

```
levetid <- c(25.0, 26.3, 23.7, 24.5, 26.8, 25.6, 22.7, 27.9, 27.3, 25.2)
```

- Estimer middelværdi μ og varians σ^2 for levetiden på batterierne på baggrund af stikprøven.
- Hvis producentens oplysninger er korrekte, hvad er da sandsynligheden for, at en ny stikprøve vil have et gennemsnit, der er endnu mindre fordelagtig end i stikprøven ovenfor.

Teori hørende til eksamenspørgsmålet

- Kontinuerte stokastiske variable, f.eks. middelværdi, varians, tæthedsfunktion og fordelingsfunktion.
- Normalfordelingen: Tæthed (skitse), middelværdi, varians, standardisering.
- Beregning af sandsynligheder i en vilkårlig normalfordeling, dvs. standardisering og opslag på computer eller i tabel.
- Symmetriske sandsynlighedsintervaller omkring middelværdien i en normalfordeling.
- Estimation af middelværdi og varians.
- Central grænseværdisætning og stikprøvefordeling for middelværdi.

random variable Kontinous

Opgave 1

Vi har en normal fordelt variabel X, med en normal fordeling. Vi kan finde Z-værdien ved at bruge formlen:

$$Z = \frac{X - \mu}{\sigma}$$

Z-score angiver hvor mange standardafvigelser man afviger fra gennemsnittet

Vi kender alle variabelværdierne:

$$X = 28$$
$$\mu = 26$$
$$\sigma = 1.5$$

Ved indsættelse fås:

$$Z = \frac{28 - 26}{1.5} = \frac{2}{1.5} = 1.33$$

Det giver os en Z værdi på 1.33, som giver os værdien 0.9082 jf. tabel side 742. Denne værdi kan vi nu bruge til at finde sandsynligheden.

P-værdi: er sandsynligheden for at observere en værdi der er ligesåeller mere ekstrem, end hvad der allerede er observeret.

Kasper Kann
Daniel Behr

05/01/2021

studienr: 20134818
studienr: 20195227

Vi skal nu finde den kumulative sandsynlighed. Dette gøres ved at tage det fulde areal(1) og trække 0.9082 fra:

$$1 - Z = 1 - 0.9082 = 9.18\%$$

Det vil sige at der er 9.18% chance for at et batteri har en levetid over 28 timer.

Normal fordeling
$Z \sim N(\mu, \sigma^2)$
Standard normal
$Z \sim N(0,1)$
(left-skewed og right-skewed)

Chi-i-anden

Opgave 2

Denne gang bruger vi R til at bestemme Z:

"over stiger"

qnorm(0,025) hvis begge males

$$Z = qnorm(0.05) = -1.64$$

Vi indsætter nu alle værdier vi kender i formlen fra opgave 1, og vil finde X:

$$-1.64 = \frac{X - 26}{1.5}$$

$$X = 26 - 1.64 \cdot 1.5 = 23.54\ timer$$

Det vil sige at 95% af batterierne vil have en længere levetid end 23.54 timer.

Opgave 3

Vi skal nu finde en symmetrisk fordeling omkring vores middelværdi på 26. Det gør vi ved at bruge vores konfidensinterval for en 95 % normalfordelingen omkring middelværdien. Vi finder først vores nedre grænse og derefter vores øvre grænse. Med disse to grænser kan vi klarlægge batteriets levetid med 95% præcision.

Vi ved at z-scoren for 95% er 1.96 på baggrund af denne tabel, som er taget direkte fra undervisningen:

$CL = 1 - \alpha$	α	$\alpha/2$	$z_{\alpha/2}$
90	0.1	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.576

Nedre grænse:

$$X = \mu - z \cdot \sigma$$

$$X = 26 - 1.96 \cdot 1.5 = 23.06$$

Øvre grænse:

$$X = \mu + z \cdot \sigma$$

$$X = 26 + 1.96 \cdot 1.5 = 28.94$$

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

Det betyder at der er 95% sandsynlighed for at batteriets levetid falder indenfor intervallet [23.06;28.94 timer].

Opgave 4

Stikprøve

```
levetid <- c(25.0, 26.3, 23.7, 24.5, 26.8, 25.6, 22.7, 27.9, 27.3, 25.2)
```

Vi vil nu finde middelværdien (sample mean) for stikprøven noteret som "levetid":

$$\mu_X = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots x_N}{N} \qquad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Og da vi ved, at det er stikprøvemiddelværdi, skal det noteres som $\bar{x}$

$$\bar{x} = \frac{25.0 + 26.3 + 23.7 + 24.5 + 26.8 + 25.6 + 22.7 + 27.9 + 27.3 + 25.2}{10} = 25.5$$

Middelværdien for stikprøven er altså 25.5

Vi vil nu finde sample variansen for stikprøven:

Pop

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

$$s^2 = \frac{\begin{array}{c}(25.0 - 25.5)^2 + (26.3 - 25.5)^2 + \\ (23.7 - 25.5)^2 + (24.5 - 25.5)^2 + (26.8 - 25.5)^2 + (25.6 - 25.5)^2 + (22.7 - 25.5)^2 \\ +(27.9 - 25.5)^2 + (27.3 - 25.5)^2 + (25.2 - 25.5)^2\end{array}}{10 - 1}$$

$$s^2 = \frac{0.89 + 4.24 + 1.7 + 13.6 + 3.33 + 2.64}{9} = 2.64$$

Dermed er variansen for stikprøven 2.64.

Mangler

$E[\bar{x}] = \mu = 25,5$

$E[s^2] = \sigma^2 = 2,64$

Kasper Kann 05/01/2021 studienr: 20134818
Daniel Behr studienr: 20195227

Opgave 5

Her skal vi bruge de to kendte variabler fra opgaven: Populationsmiddelværdi $\mu = 26$, og populationsvariansen = $\sigma^2 = (1.5)^2$. Derudover så kender vi også allerede stikprøvemiddelværdien fra tidligere $\bar{x} = 25.5$. Nu kan vi bruge disse parametre til at bestemme Z-værdien ved hjælp af følgende formel: $Z = \frac{\bar{x}-\mu}{\sigma}$

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ har ikke mange observationer

Vi bruger de tre variable til at bestemme Z værdien:

$$\bar{x} = 25.5$$

$$\mu = 26$$

$$\sigma = 1.5$$

Ved indsættelse fås:

$$Z = \frac{\bar{x}-\mu}{\sigma} = \frac{25.5-26}{1.5} = -0.33$$

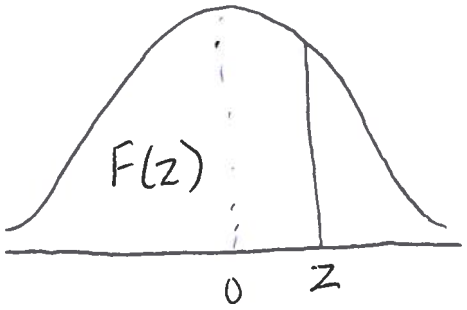
Slår vi denne værdi op i Appendix table 1, s. 742, får vi værdien: 0.6293.

Nu trækker vi denne værdi fra det samlede areal (1), og får vores sandsynlighed.

$$1 - 0.6293 = 0.3707$$

Dette vil sige, at der med 37,07% sandsynlighed er chance for, at den nye stikprøve er mindre fordelagtig end den første stikprøve.

Hvis man kan bruge begrebet "mindre fordelagtigt", som jo i sig selv er to modstridende ord, så kan man derfor sige, at levetiden på batterierne, derfor har en risiko på 37.07% for at have en kortere levetid.



F(z) = kumulative fordelings funktion

1 − z for at finde en mere extrem værdi

Exam 3

Statistics 1

Kasper Kann og Daniel Behr
Student ID: 20134818 og 20195227
januar 2021

Aalborg University Business School

Contents

1 Opgave 1 **1**

1.1 Lav et histogram over løn og log-løn sammen med en normalfordelingstæthed – hvilken ser mest normalfordelt ud? Hint til kommandoer (husk at slette eval = FALSE i Rmd-fil hvis denne bruges som skabelon): 1

2 Opgave 2 **2**

2.1 Estimer middelværdi og varians for log-løn. 2

3 Opgave 3 **3**

3.1 Bestem et 90% og 95% konfidensinterval for middel log-løn. 3

4 Opgave 4 **4**

4.1 Test på 5% signifikansniveau om middel log-løn er lig med 1.6. Bestem p-værdien for testet (brug evt.t.test – dog skal man kunne forklare alle mellemregninger). 4

5 Opgave 5 **6**

5.1 Bestem 95% konfidensinterval for variansen for log-løn. 6

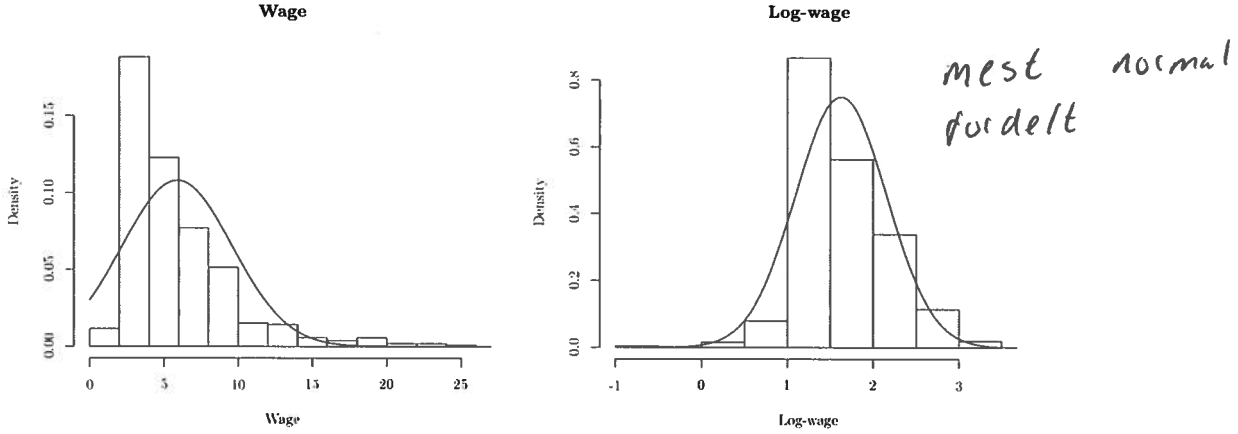
6 Opgave 6 **7**

6.1 Test på 5% signifikansniveau om variansen af log-løn er større end 0.25. 7

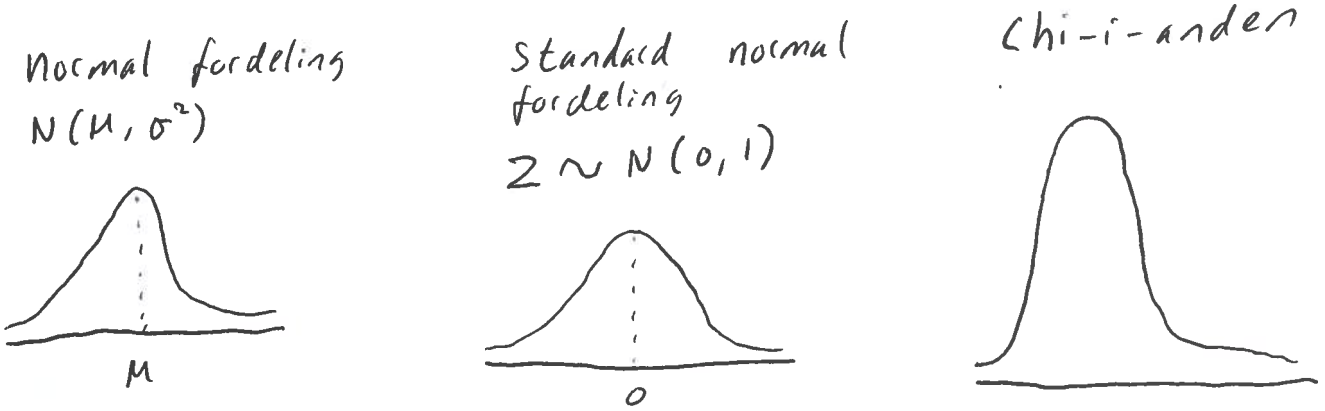
1 Opgave 1

1.1 Lav et histogram over løn og log-løn sammen med en normalfordelingstæthed – hvilken ser mest normalfordelt ud? Hint til kommandoer (husk at slette eval = FALSE i Rmd-fil hvis denne bruges som skabelon):

```
wage1 <- read.table("C:/Users/Kann/Downloads/wage1.dat", header = TRUE)
par(mfrow = c(1, 2))
hist(wage1$wage, prob = TRUE, main = "Wage", xlab = "Wage")
curve(dnorm(x, mean(wage1$wage), sd(wage1$wage)), from = 0, to = 30, add = TRUE)
hist(wage1$lwage, prob = TRUE, main = "Log-wage", xlab = "Log-wage")
curve(dnorm(x, mean(wage1$lwage), sd(wage1$lwage)), from = -1, to = 5, add = TRUE)
```

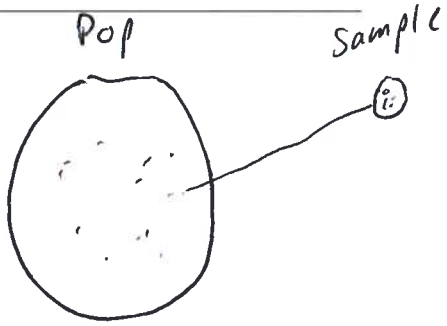


Vi kan se at wage er right-skewed og log-wage er left-skewed. Når noget er right-skewed, så ligger middelværdien til højre for toppunktet. For left-skewed er det omvendt, her ligger middelværdien til venstre for toppen.



2 Opgave 2

2.1 Estimer middelværdi og varians for log-løn.



Her bruger vi formlen for samplemiddelværdi og samplevarians.

Sample

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

Pop

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Finder først middelværdien på den nemme måde, og så gør vi det også på den sværere måde, hvor vi følger ligningerne opstillet ovenfor:

```
#Finder middelværdien
mean_wage1 <- mean(wage1$lwage)
#Finder variansen
var_wage1 <- var(wage1$lwage)

#Finder den på en sværere måde, hvor vi bruger formlen ovenfor:
alt_mean_wage1 <- sum(wage1$lwage) / length(wage1$lwage)
alt_var_wage1 <- sum(((wage1$lwage) - mean_wage1)^2)/(length(wage1$lwage)-1)

#Opstiller en tabel med middelværdi og varians
tabel <- c(mean_wage1,var_wage1)
names(tabel) <- c("Middelværdi", "Varians")
tabel
## Middelværdi      Varians
##   1.6229658   0.2825763
```

3 Opgave 3

3.1 Bestem et 90% og 95% konfidensinterval for middel log-løn.

Z-værdi for vi har mere end 30 obs

$Z = \frac{X-\mu}{\sigma}$

Her skal vi finde T-scoren For en sample middelværdi, som er på følgende formel:

$$\bar{X} \pm t_{v,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

- Først finder vi vores T-værdi ved 90% konfidensinterval og den øvre og nedre grænse.

```
#Finder T-scoren
t_value <- qt(0.95, df = length(wage1$lwage)-1)
#øvre grænse
t_upper <- mean_wagel + t_value * (sqrt(var_wagel)/sqrt(length(wage1$lwage)))
#Nedre grænse
t_lower <- mean_wagel - t_value * (sqrt(var_wagel)/sqrt(length(wage1$lwage)))
c(t_lower, t_upper)
## [1] 1.584774 1.661158
```

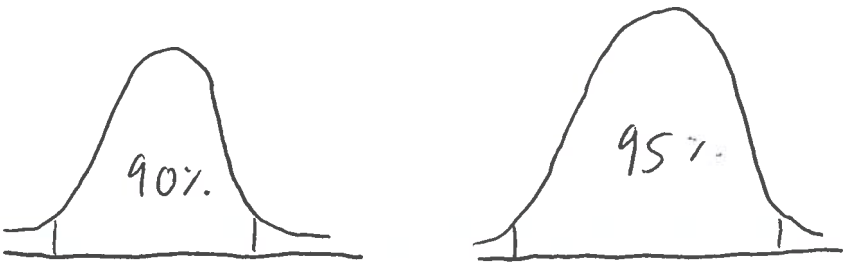
v = frihedsgrader

På et 90% konfidensinterval, kan vi sige, at middelværdien på log-løn ligger inden for [1.58:1.66]

- Vi vil nu finde konfidensintervallet for et 95% konfidensinterval:

```
#Finder T-scoren
t_value2 <- qt(0.975, df = length(wage1$lwage)-1)
#øvre grænse
t_upper2 <- mean_wagel + t_value2 * (sqrt(var_wagel)/sqrt(length(wage1$lwage)))
#Nedre grænse
t_lower2 <- mean_wagel - t_value2 * (sqrt(var_wagel)/sqrt(length(wage1$lwage)))
c(t_lower2,t_upper2)
## [1] 1.577433 1.668499
```

På et 95% konfidensinterval, kan vi sige, at middelværdien på log-løn ligger inden for [1.57:1.66]
*Når man øger konfidensintervallet, så øger man også arealet under bell-kurven, som viser en normalfordeling. Intervallet bliver derfor større.



4 Opgave 4

4.1 Test på 5% signifikansniveau om middel log-løn er lig med 1.6. Bestem p-værdien for testet (brug evt.t.test – dog skal man kunne forklare alle mellemregninger).

Opstiller Hypoteser:

$$H0 : \mu[lwage] = 1.60$$

alternativ → $$H1 : \mu[lwage] \neq 1.60$$

Først finder vi standardafvigelsen med hjælp fra tidligere varians.

$$\sigma = \sqrt{s^2} \qquad s = \sqrt{s^2} \qquad \sigma = \sqrt{\sigma^2}$$

```
sd_wage1 <- sd(wage1$lwage)
sd_wage1
## [1] 0.5315791
```

Finder antallet af observationer

```
n_1 <- length(wage1$lwage)
n_1
## [1] 526
```

Finder standard fejlen ved hjælp af:

$$se = \frac{\sigma}{\sqrt{n}}$$

```
se_1 <- sd_wage1 /sqrt(n_1)
se_1
## [1] 0.02317795
```

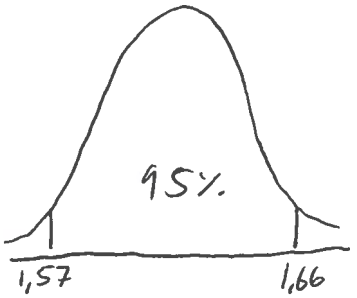
Finder vores T-statistik og p-værdien.

```
#T-stats
t_stats <- mean_wage1/se_1
t_stats
## [1] 70.02197

#Finder nu p-værdien
2*(1 - pt(t_stats, df = n_1-1))
## [1] 0
```

matematisk ikke 0

Her er det vigtigt at påpege, at en P-værdi matematisk set ikke kan vævre nul, men bare meget tæt på nul. Vi går derfor ud fra, at R har afrundet den. Det kan vi også teste ved hjælp af t.test-funktionen i R.

```
t.test(wage1$lwage)
##
##  One Sample t-test
##
## data:  wage1$lwage
## t = 70.022, df = 525, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.577433 1.668499
## sample estimates:
## mean of x
##  1.622966
```



Her kan vi konkludere, at vi kan afvise vores nulhypotesn inden for et 5% siknifikansniveau, fordi at p-værdien $<$ 0.05. Vi kan også se, at T-værdi, konfidensinterval og mean passer overens med vores udregninger fra tidligere opgaver.

5 Opgave 5

5.1 Bestem 95% konfidensinterval for variansen for log-løn.

Her skal vi finde et konfidensinterval på 95% for variansen, så vi skal altså have fat i nogle formler for Lower-confidens-limit(LCL) og Upper-confidens-limit(UCL) for chi-square distributionen:

Chi-i-anden: når pop varians ikke er kendt.

$$LCL = \frac{(n-1)}{\chi^2_{n-1,\alpha/2}} S^2$$

$$UCL = \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} S^2$$

Først kan vi slå vores chi-square værdier op ved hjælp af qhicsq() funktionen i R på et 5 procents siknifikansniveau.

$\chi^2_{n-1,\alpha/2}$
$\chi^2_{n-1,1-\alpha/2}$

```
wl_0.025 <- qchisq(p = 0.05/2, df = (length(wage1$lwage)-1), lower.tail = FALSE)
wl_0.975 <-qchisq(p = 1-0.05/2, df = (length(wage1$lwage)-1), lower.tail = FALSE)
c(wl_0.025, wl_0.975)
## [1] 590.3827 463.4049
```

Vi kan nu finde vores upper og lower limit for vores konfidensinterval.

```
#Nedre grænse
wl_LCL <- ((length(wage1$lwage)-1)*var_wage1)/wl_0.025
#Øvre grænse
wl_UCL <- ((length(wage1$lwage)-1)*var_wage1)/wl_0.975
c(wl_LCL, wl_UCL)
## [1] 0.2512821 0.3201360
```

Vi kan derfor sige at med 95% sandsynlighed, der falder variansen for logløn inden for intervallet [0.251:0.320]

6 Opgave 6

6.1 Test på 5% signifikansniveau om variansen af log-løn er større end 0.25.

Her skal vi opstille en en-sidet hypotesetest, hvor vi opstiller følgende:

$$H0 : \sigma^2[lwage] = 0.25$$
$$H1 : \sigma^2[lwage] > 0.25$$

Her skal vi bruge test-statistikken som måler hvor stor variansen er relativt til den hypotetiske værdi. Den opskrives således:

$$\chi^2_{n-1} = \frac{s^2(n-1)}{\sigma_0^2}$$

Så sætter vi vores kendte værdier ind og vores 0.25 varians-test.

```
t_stat_0.25 <- (((length(wage1$lwage)-1)*var_wage1)/0.25)
t_stat_0.25
## [1] 593.4103
```

Med denne værdi, kan vi finde vores P-værdi, som kan bruges til at bekræfte eller afkræfte hypotesen:

```
p_værdi_wl <- 1-pchisq(t_stat_0.25, df = (length(wage1$lwage)-1))
p_værdi_wl
## [1] 0.02036618
```

Vi har en p-værdi < 0.05, og vi kan derfor afvise vores hypotese på et 5% signifikansniveau. Dette vil altså sige, at variansen [redacted] er statistisk signifikant højere end 0.25

Exam 4

Statistics 1

Kasper Kann og Daniel Behr
Student ID: 20134818 og 20195227
januar 2021

Aalborg University Business School

Contents

1 Opgave 1 **1**

1.1 Lav et boxplot der sammenligner log-loen for maend og kvinder. 1

2 Opgave 2 **2**

2.1 Estimer middelvaerdi og varians for log-loen for hhv. maend og kvinder. 2

3 Opgave 3 **3**

3.1 Test paa signifikansniveau 5% om der er en forskel i middel log-loen for maend og kvinder (en p-vaerdi skal rapporteres). 3

4 Opgave 4 **5**

4.1 Bestem et 95% konfidensinterval for forskellen i middel log-loen for maend og kvinder. 5

5 Opgave 5 **6**

5.1 Estimer andelen af gifte personer hhv. i og udenfor storbyer 6

6 Opgave 6 **7**

6.1 Test på 5% signifikansniveau om der er en forskel i andelen af gifte i og udenfor storbyer. . . 7

7 Opgave 7 **8**

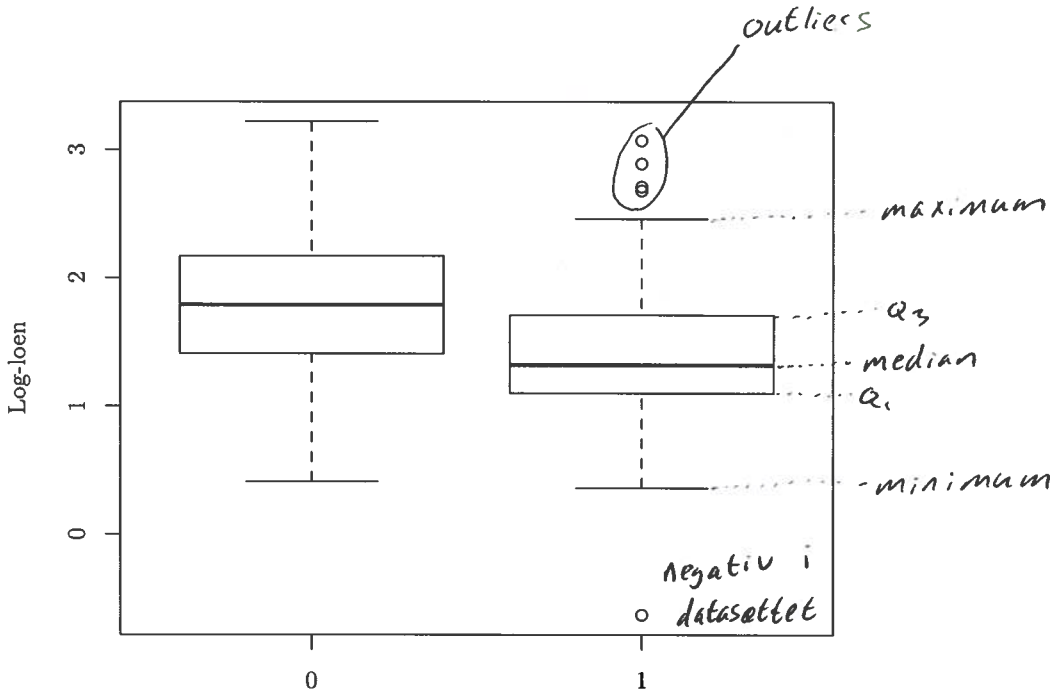
7.1 Bestem et 95% konfidensinterval for forskellen i andelen af gifte i og udenfor storbyer. 8

```
wage1 <- read.table("C:/Users/Kann/Downloads/wage1.dat", header = TRUE)
```

1 Opgave 1

1.1 Lav et boxplot der sammenligner log-loen for maend og kvinder.

```
boxplot(lwage~ female, data = wage1, xlab = "Maend, kvinder", ylab = "Log-loen")
```



Her kan vi se, at der i vores datasaet er en tendens til at loennen hos maend er hoejere end hos kvinder. Vi kan ogsaa se, at der er en stor spredning i Log-loennen for baade maend og kvinder.Det kan ogsaa ses at mindstevaerdien for begge koen er stor set identiske undtaget en enkel outlier.

2 Opgave 2

2.1 Estimer middelvaerdi og varians for log-loen for hhv. maend og kvinder.

Her kan vi bruge foelgende formler:

Sample

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Pop

$$\mu = \sum_{i=1}^{N} \frac{X_i}{N}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{X_i}{N}$$

Det kan findes baade paa den lette maade ved at benytte R's automatiske funktioner eller vi kan udregne det manuelt:

```
#Sorterer vores data.
library(dplyr)
male <- wage1 %>%
  filter(female == 0, !is.na(female))

female <- wage1 %>%
  filter(female == 1, !is.na(female))

#udregner middelværdi
mean_wagel_female <- mean(female$lwage)
mean_wagel_male <- mean(male$lwage)
#udregner varians
var_wagel_female <- var(female$lwage)
var_wagel_male <- var(male$lwage)

#En måde, hvor man ikke bare lader R udregne det automatisk..
alt_mean_wagel_male <- sum(male$lwage) / length(male$lwage)
alt_var_wagel_male <- sum(((male$lwage) - alt_mean_wagel_male)^2)/(length(male$lwage)-1)

#Opstiller vores værdier i en tabel
mean_var_res <- matrix(c(mean_wagel_female, mean_wagel_male,
                         var_wagel_female, var_wagel_male), nrow = 2, byrow = TRUE)
rownames(mean_var_res) <- c("Mean", "Varians")
colnames(mean_var_res) <- c("Female","Male")
mean_var_res
##            Female      Male
## Mean     1.4159127 1.8133942
## Varians 0.1973183 0.2860298
```

3 Opgave 3

3.1 Test paa signifikansniveau 5% om der er en forskel i middel log-loen for maend og kvinder (en p-vaerdi skal rapporteres).

Her skal vi lave en hypoteste for en 2-sidet hypotese test, og vi starter derfor med at opstille en hypotese.

$$H0 : \mu[female_lwage] = \mu[men_lwage]$$
$$H1 : \mu[female_lwage] \neq \mu[men_lwage]$$

alternativ

Her bruger vi formlen hvor populationsvarisen er ukendt, men forskellige. Derudover bruger vi formelen for T-statistik.

$$se_d = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$
$$t = \frac{\bar{x} - \bar{y}}{se_d}$$

Udregner foerst standardfejlen og derefter T-statistikken.

```
#Standardfejl
se_wl <- sqrt((var_wagel_male)/(length(male$lwage))+((var_wagel_female)/length(female$lwage)))
#T-stat
t_wl <- ((mean_wagel_male - mean_wagel_female) / se_wl)

#Opstiller tabel
se_t <- c(se_wl, t_wl)
names(se_t) <- c("Standardfejl","T-stat")
se_t
## Standardfejl       T-stat
##   0.04274241   9.29946317
```

Nu vil vi gerne finde P-vaerdien, men foerst skal vi finde frihedsgraderne. Det goeres ved formlen:

$$v = \frac{[(\frac{s_x^2}{n_x} + \frac{(s_y^2)}{n_y}]^2}{(\frac{s_x^2}{n_x})^2/(n_x - 1) + (\frac{s_y^2}{n_y})^2/(n_y - 1)}$$

Det er en lidt større udregning i R, og vi har derfor delt det op i flere dele.

```
dof <- ((var_wagel_male/length(male$lwage))+(var_wagel_female/length(female$lwage)))^2/
(((var_wagel_male/length(male$lwage)^2)/
(length(male$lwage-1)+((var_wagel_female/length(female$lwage)^2)/
(length(female$lwage-1))))))

q <- ((((var_wagel_male/length(male$lwage)))+(var_wagel_female/length(female$lwage)))^2)
```

```
s <- ((var_wagel_male/length(male$lwage))^2)/(length(male$lwage)-1)

o <- ((var_wagel_female/length(female$lwage))^2)/(length(female$lwage)-1)
dof_1 <- q/(s+o)
dof_1
## [1] 518.7179
```

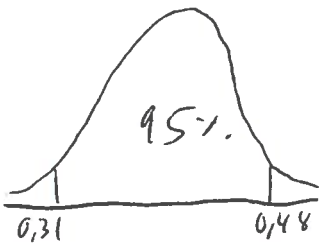
Nu kan vi finde vores P-vaerdi ved hjaelp af T-statistikken og degrees of freedom(df). Reelt kunne vi have undladet selv at finde degrees of freedom(df) og lade R goere det paa forskellige maader. Herunder vises 3 metoder:

```
p_wl <- 2*(1-pf(t_wl, df1=(length(male$lwage)-1), df2= (length(female$lwage)-1)))
p_wl2 <- 2*(1-pt(t_wl, df = dof_1))
t_wl3 <- t.test(male$lwage, female$lwage, var.equal  = FALSE)
```

Her kan vi se, at begge metoder til p-vaerdien giver det samme, og at T-testen ogsaa giver en p-vaerdi taet paa nul. P-værdien kan matematisk ikke være 0, men meget tæt på, derfor må vi gå ud fra, at R har afrundet den.

```
c(p_wl, p_wl2)
## [1] 0 0
t_wl3
##
##  Welch Two Sample t-test
##
## data:  male$lwage and female$lwage
## t = 9.2995, df = 518.72, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.313512 0.481451
## sample estimates:
## mean of x mean of y
##  1.813394  1.415913
```

matematisk ikke lig 0



Naar p-vaerdien < 0.05, saa kan vi paa et 5% signifikansniveau forkaste nulhypotesen om, at maend og kvinder skulle have den samme Log-loen.

4 Opgave 4

4.1 Bestem et 95% konfidensinterval for forskellen i middel log-loen for maend og kvinder.

Her skal vi finde vi finde konfidensintervallet ved hjælp af forskellen i middelværdier, t-værdien, standard-afvigelsen og degrees of freedom. Det kan gøres ved hjælp af følgende formler:

Standard fejlen

$$se_d = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$t = \frac{\bar{x} - \bar{y}}{se_d}$$

$$\bar{x} \pm t_{n-1,\alpha/2} se_d$$

Først finder vi forskellen på middelværdien mellem mænd og kvinder, og derefter finder vi standardfejlen:

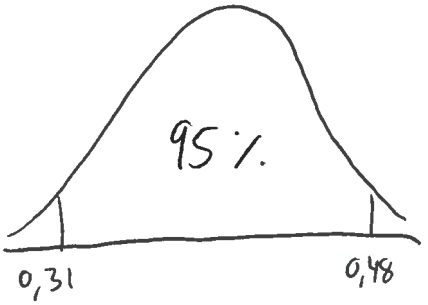
```
#Finder forskellen i middelværdien mellem mænd og kvinder
mean_wagel_diff <- mean(male$lwage) - mean(female$lwage)
#Finder standardafvigelsen
se_wagel <- sqrt(((var_wagel_male)/length(male$lwage)) + (var_wagel_female)/length(female$lwage))
```

Nu kan jeg finde vores T-værdi og bruge den til at finde vores grænser UCL og LCL.

```
#T-værdien
t_value_diff <- qt(0.975, df = dof_1)
#Øvre grænse
t_upper_diff <- mean_wagel_diff + t_value_diff * (se_wagel)
#Nedre grænse
t_lower_diff <- mean_wagel_diff - t_value_diff * (se_wagel)
```

Det vil altså sige at vi med en 95% sandsyndlighed kan sige, at forskellen for middelværdi i log-løn vil falde indenfor intervallet:

```
## [1] 0.313512 0.481451
```



Tester ved hjælp af R om vi har fået det rigtige resultat:

```
t.test(male$lwage, female$lwage)
##
##  Welch Two Sample t-test
##
## data:  male$lwage and female$lwage
## t = 9.2995, df = 518.72, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.313512 0.481451
## sample estimates:
```

```
## mean of x mean of y
##  1.813394  1.415913
```

5 Opgave 5

5.1 Estimer andelen af gifte personer hhv. i og udenfor storbyer

Her laver vi forskellige sorteringer i vores datasæt, og vi kan på den måde udregne de to andele.

```
gifte <- factor(wage1$married == 1)
storbyen <- factor(wage1$smsa == 1)
tab <- (xtabs(~storbyen+gifte, data=wage1))
tab <- tab[ , c("TRUE", "FALSE")]
tab
##          gifte
## storbyen TRUE FALSE
##    FALSE  100    46
##    TRUE   220   160
```

Udregner først px, som er andelen af gifte i storbyen, og derefter py, som er andelen af gifte på landet.

```
p_x <- 220/(220+100) # Gifte i storbyen
p_y <- 100/(100+220) # Gifte på landet

px_py <- c(p_x, p_y)
names(px_py) <- c("Gifte i storbyen","Gifte på landet")
px_py
## Gifte i storbyen  Gifte på landet
##           0.6875           0.3125
```

6 Opgave 6

6.1 Test på 5% signifikansniveau om der er en forskel i andelen af gifte i og udenfor storbyer.

Vi opstiller en nulhypotese, der kigger på forskellen ved hjælp af en Z-score. Hvis scoren er indenfor eller lig intervallet for et 5% siknifikansniveau, så kan vi ikke afvise hypotesen.

$$H0 : z \geq -z_{\alpha/2} \qquad H_0 : \hat{p}_x - \hat{p}_y = 0$$

$$H1 : z < -z_{\alpha/2} \qquad H_1 : \hat{p}_x - \hat{p}_y \neq 0$$

Her kan vi finde vores Z-score for at se, om vi kan afvise hypotesen om, at der skulle være en forskel på andelen af gifte i og udenfor storbyen: Først opstiller vi en tabel med de Z-værierne:

$CL = 1 - \alpha$	α	$\alpha/2$	$Z_{\alpha/2}$
90	0.1	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.006	2.576

Figure 1: Z
-værdier for forskellige konfidensintervaller

```
#Først angiver vi antallet af gifte.
antal_x_værdier <- 220+100
antal_y_værdier <- 100+220
# Vi skal nu finde forskellen på de to propertioner
p_0_6 <- (antal_y_værdier*p_y + antal_x_værdier*p_x)/(antal_x_værdier+antal_y_værdier)
#Udregner vores Z-score ved hjælp af standarderror
se_0_del_1 <- (p_0_6*(1-p_0_6))/(antal_x_værdier)
se_0_del_2 <- (p_0_6*(1-p_0_6))/(antal_y_værdier)
se_0_6 <- sqrt(se_0_del_1 + se_0_del_2)
z_score_6 <- (p_x-p_y)/se_0_6
z_score_6
## [1] 9.486833
z_score_5_procent <- qnorm((0.05/2))
z_score_5_procent
## [1] -1.959964
```

Med et 5% siknifikansniveau kan vi ~~ikke~~ afvise H0, fordi vores Z-score falder ~~indenfor~~ det accepterede interval, som grænser sig til 1.96. Det vil sige, at der ~~ikke~~ er en statistisk siknifikant forskel på gifte i storbyerne og på landet.

7 Opgave 7

7.1 Bestem et 95% konfidensinterval for forskellen i andelen af gifte i og udenfor storbyer.

Det kan findes ved følgende formel:

$$\hat{p_y} - \hat{p_x} \pm 1.96 * se_d$$

Finder først standardfejlen og derefter vores øvre og nedregrænse.

$$Se_d = \sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n_x} + \frac{\hat{P}_y(1-\hat{P}_y)}{n_y}}$$

```
#Standardfejl
se_del_1 <- (p_x*(1-p_x))/(antal_x_værdier)
se_del_2 <- (p_y*(1-p_y))/(antal_y_værdier)
se_d_6 <- sqrt(se_del_1) + sqrt(se_del_2)

py_px = p_y-p_x # Gifte på landet minus gifte i storbyen

#øvre grænse
py_px_UCL <- py_px + 1.96 * se_d_6

#nedre grænse
py_px_LCL <- py_px - 1.96 * se_d_6
```

Vi har fundet et konfidensinterval, som med 95% sandsynlighed kan vise forskellen i andelen af gifte i og udenfor storbyen. Det viser også, at der er en negativ sammenhæng mellem gifte på landet og i storbyen. Vi kan altså sige, at der med 95% sandsynlighed er en negativ sammenhæng indenfor dette interval.

```
c(py_px_LCL, py_px_UCL)
## [1] -0.4765716 -0.2734284
```

Exam 5

Statistics 1

Kasper Kann og Daniel Behr
Student ID: 20134818 og 20195227
januar 2021

Aalborg University Business School

Contents

1 Opgave 1 1

1.1 Udfør følgende opgaver ved at bruge den fulde stikprøve (2000-2015): 1

1.1.1 Beregn de sammenfattende statistik for de tre variable. Organiser dine resultater i en enkelt tabel. 1

1.1.1.1 Vis data fordelingen for de tre variable. På baggrund af din statistiske viden, hvad kan du sige om fordelingen af disse tre variable? 2

1.1.1.2 Beregn den simple korrelation mellem variablene. Ved at se på disse sammenhænge, hvad kan du fortælle om sammenhængen blandt variablene? 3

2 Opgave 2 4

2.1 Dan to scatter plots, i) unemployment rate og investment, ii) unemployment rate og current account. Kommenterer på forholdet mellem variable. 4

3 Opgave 3 5

3.1 Estimer en regressionsmodel efter følgende funktion: 5

3.1.1 Estimer ovenstående regressionsmodel under anvendelse af følgende prøver: 5

3.1.1.1 Hvad kan du sige om virkningerne af investeringer og løbende poster på arbejdsløshed? 6

3.1.1.2 Hvad kan du sige om virkningerne af investeringer og løbende poster på arbejdsløshed? 8

3.1.1.3 Diskuter hvorvidt estimaterne for de tre modeller er ens eller forskellige? . . 8

3.1.1.4 Udfør hypotesetest på dine regressionsmodeller. Er disse resultater statistisk signifikante? 8

1 Opgave 1

1.1 Udfør følgende opgaver ved at bruge den fulde stikprøve (2000-2015):

1.1.1 Beregn de sammenfattende statistik for de tre variable. Organiser dine resultater i en enkelt tabel.

Vi har først navngivet vores variabler til nogle kortere navne, og de er som følger:
Vores variabler er delt op med navne og årstal:

1. Navne:
   - u = arbejdsløshed
   - c = betalingsbalancen
   - inv = investeringsraten ift. BNP.
2. Årstal:
   - 00-15 = 2000-2015
   - 00-08 = 2000-2008
   - 09-15 = 2009-2015

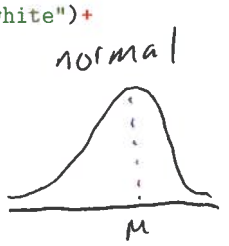
Vi bruger kun årstallet 2000-2015 i denne opgave, men vil senere komme til at bruge nogle af de andre årstal.
Ved hjælp af stargazer fåes tabellen:

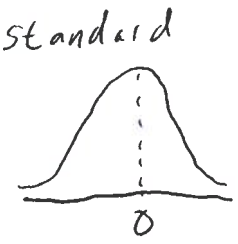
Table 1: Statistik for hele perioden

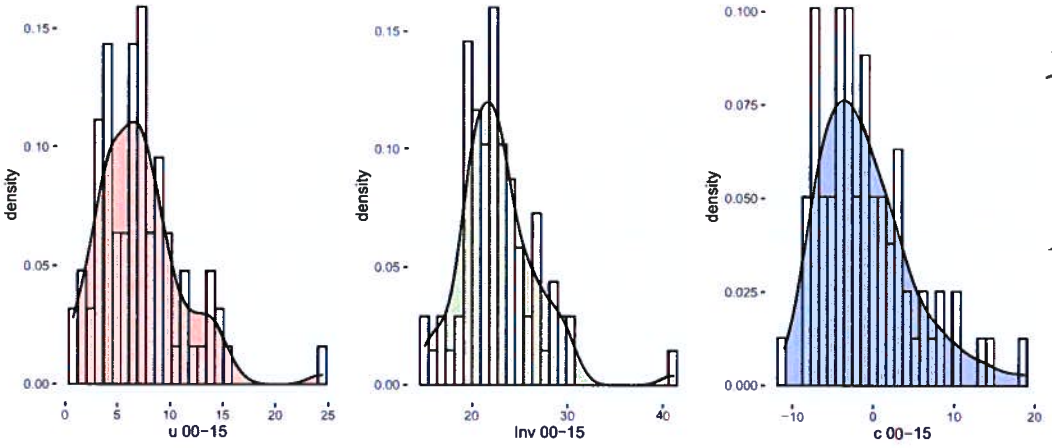
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
inv_00_15	77	22.827	4.038	15.220	20.150	24.690	41.070
u_00_15	77	7.144	4.014	0.810	4.340	8.930	24.440
c_00_15	77	−0.964	5.761	−11.030	−4.970	2.140	18.840

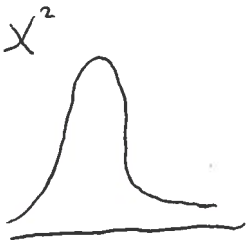
1.1.1.1 Vis data fordelingen for de tre variable. På baggrund af din statistiske viden, hvad kan du sige om fordelingen af disse tre variable?

```
plot_u <- ggplot(data, aes(x=u_00_15)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")+
  scale_x_continuous("u 00-15")

plot_inv <- ggplot(data, aes(x=inv_00_15)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="green") +
  scale_x_continuous("Inv 00-15")


plot_c <- ggplot(data, aes(x=c_00_15)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="blue")+
  scale_x_continuous("c 00-15")


grid.arrange(plot_u, plot_inv, plot_c, ncol=3)
```









Her kan vi se på vores histogrammer, at de alle sammen er right-skewed. Vi kan dog se på betalingsbelancen, at den er lidt tættere på at være en normalfordeling end de to andre. Grunden til at de er right-skewed er fordi, at der flest værdier tæt på minimumsværdien. Det vil altså sige, at vores gennemsnit er tættere på minimumsværdien end maksimumsværdien, og de derved bevæger sig mere mod venstre og efterlader en "højre-hale".

$$Cov(X,Y) = \sum_x \sum_y (x_i - \mu_x)(y_i - \mu_y) P(x,y)$$

1.1.1.2 Beregn den simple korrelation mellem variablene. Ved at se på disse sammenhænge, hvad kan du fortælle om sammenhængen blandt variablene? Her skal vi bruge formlen for Korrelation, hvis man skal lave det i hånden:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E[XY] - \mu_X \mu_Y$$

$$\delta = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Korrelationen angiver en sammenhæng mellem to variable. En korrelation tager en værdi mellem -1 og 1, hvor -1 er en perfekt negativ sammenhæng, og 1 er en perfekt positiv sammenhæng. 0 Angiver, at der ikke er en sammenhæng mellem de to variable. R kan udregne Korrelationen nemt, og det gør vi i denne opgave:

```
#Investeringer og arbejdsløshed
inv_u <- cor(data$inv_00_15, data$u_00_15)
#Investeringer og betalingsbalancen
inv_c <- cor(data$inv_00_15, data$c_00_15)
#Arbejdsløshed og betalingsbalancen
u_c <- cor(data$u_00_15, data$c_00_15)

c(inv_u,inv_c,u_c)
## [1] -0.18023269 -0.01260807 -0.33492828
```

Vi kan altså se, at sammenhængen for følgende er:

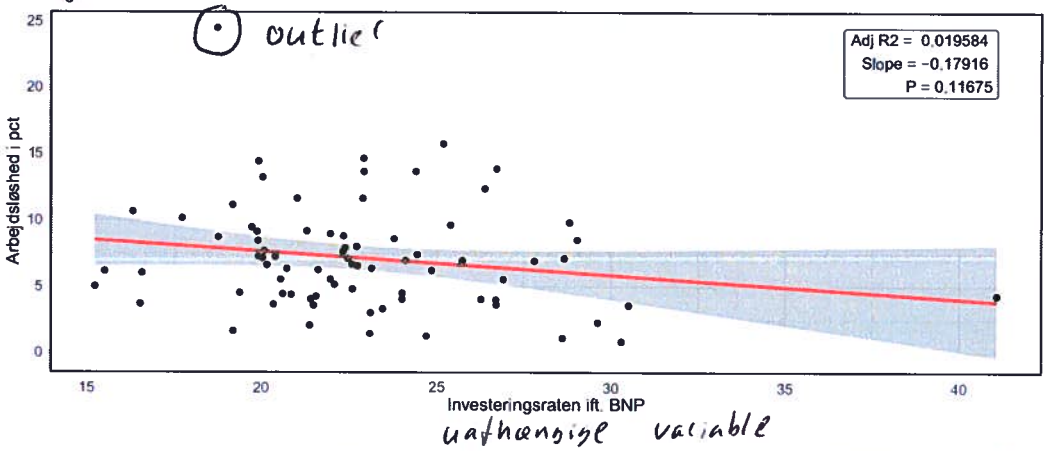
1. Investeringer og ledighed:
    - Her kan vi se, at sammenhængen mellem investeringer og Ledighed er negativ. Dog er den ikke stærkt negativ, så det tyder også på, at der er andre faktorer, som kan påvirke sammenhængen mellem investeringsrater og ledighed.
2. Investeringer og betalingsbalancen:
    - Her er sammenhængen stadigvæk negativ, men den er meget tæt på nul. Det betyder, at der i vores stikprøve stadigvæk er en negativ sammenhæng mellem investeringer og betalingsbalancen, men den er meget svag.
3. Arbejdsløshed og betalingsbalancen:
    - Her er sammenhængen også negativ, dog i højere grad end de to andre. Dette giver også god mening, fordi at en højere ledighed vil svække betalingsbalancen.

$$Corr(X,Y) = \delta = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

2 Opgave 2

2.1 Dan to scatter plots, i) unemployment rate og investment, ii) unemployment rate og current account. Kommenterer på forholdet mellem variable.

Sammenhæng mellem arbejdsløsheden og investeringsraten(2000-2015)
Figur 1

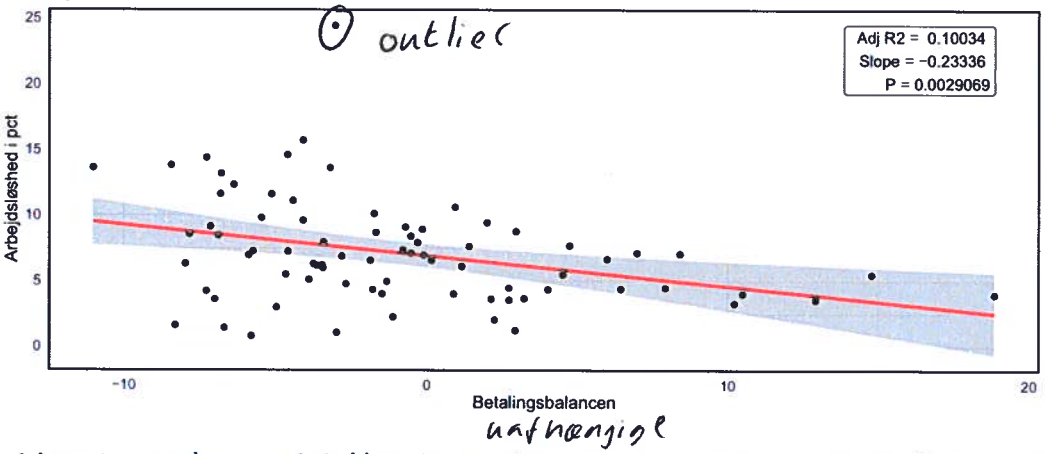


afhængige-variable

grå skygge: 95 % konfidens-interval

I figur kan vi se, at der er en negativ sammenhæng mellem Investeringer ift. BNP og arbejdsløsheden. vi ser dog også, at der er nogle outliers, og der er faktisk også en masse punkter, som ligger uden for konfidensintervallet på 95%, som er markeret af den grå skygge. P-værdien er heller ikke under 0.05, og vi kan derfor ikke sige, om denne negative sammenhæng er statistisk siknifikant på et 5% siknifikansniveau. Dette tyder på, at der er andre faktorer som kan påvirke arbejdsløsheden, som vi også har konkluderet tidligere.

Sammenhæng mellem arbejdsløsheden og betalingsbalancen(2000-2015)
Figur 2



afhængige

I figur 2 kan vi se, at der er en forholdsvis høj negativ sammenhæng mellem betalignsbalancen og arbejdsløsheden. Her har vi også en meget lav P-værdi, som betyder, at der er en statistisk siknifikant sammenhæng. Da P-værdien er under 0.01, så kan vi på et 99% siknifikansniveau bekræfte, at der er en negativ sammenhæng mellem betalingsbalancen og arbejdsløsheden. Der er dog en masse punkter(lande), som ikke falder indenfor konfidensintervallet. Dette kunne tyde på at selve regressionen måske ikke forklarer den bedste sammenhæng, og man er derfor nødt til at bruge flere statistiske værktøjer.

Regressionsmodel:
En regressionsmodel er en model, der måler Sammenhængen mellem en afhængig variabel (y), og X antal uafhængige variable (x)

3 Opgave 3

3.1 Estimer en regressionsmodel efter følgende funktion:

$$Unemployment\ rate = \beta_0 + \beta_1(investment) + \beta_1(current\ account) + u$$

Ligningen ovenfor er forsimplet ud fra følgende ligning:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Variablernes definitioner:

y	er den afhængige variabel
x_1 *og* x_2	er de uafhængige variabler
u	er fejlledet
β_0	er skæringen med y-aksen (en kostant)
$\beta_1, \beta_2, \beta_k$	er parameterer for hver uafhængig variabel$(x_1, x_2 ..... x_k)$, som man estimerer

3.1.1 Estimer ovenstående regressionsmodel under anvendelse af følgende prøver:

- Fulde stikprøve (gennemsnit af 2000-2015)
- Pre krise prøve (gennemsnit af 2000-2009)
- Post krise prøve (gennemsnit af 2009-2015)

For at estimere ovenstående parametrer, antager vi at modellen er en linær regressionsmodel med 2 uafhængige variabler. Ligningen for dette har vi fået givet i starten af opgagen, hvor investments og current account er vores to uafhængige variabler.
Skæringen med Y-aksen findes ved:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

og vi estimerer

$$\beta_1 \ og \ \beta_2$$

ved følgende formler:

$$\beta_1 = \frac{S_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{S_{x_1}(1 - r^2_{x_1 x_2})}$$

$$\beta_2 = \frac{S_y(r_{x_2 y} - r_{x_1 x_2} r_{x_1 y})}{S_{x_2}(1 - r^2_{x_1 x_2})}$$

Variablernes definitioner:

$r_{x_1 y}$	er stikprøve korrelationen mellem X_1 *og* Y
$r_{x_2 y}$	er stikprøve korrelationen mellem X_2 *og* Y
$r_{x_1 x_2}$	er stikprøve korrelationen mellem X_1 *og* X_2
S_{x_1}	er stikprøve standardafvigelsen for X_1
S_{x_2}	er stikprøve standardafvigelsen for X_2
S_y	er stikprøve standardafvigelsen for Y

3.1.1.1 Hvad kan du sige om virkningerne af investeringer og løbende poster på arbejdsløshed?

Her laver vi en regressionstabel ved hjælp af stargazer pakken

```
model1 <- lm( u_00_15 ~ inv_00_15 + c_00_15 , data = data)
model2 <- lm(u_00_08 ~ inv_00_08 + c_00_08, data = data)
model3 <- lm(u_09_15 ~ inv_09_15 + c_09_15, data = data)
stargazer(model1,model2, model3,
          style = "qje", header = F, type = "latex", title = "Regressions resultater",
          dep.var.labels=c("Arbejdsløshed 2000-2015",
                           "Arbejdsløshed 2000-2008",
                           "Arbejdsløshed 2009-2015"),
          covariate.labels=c("Investeringer 2000-2015",
                   "Betalingsbalance 2000-2015",
                   "Investeringer 2000-2008",
                   "Betalingsbalance 2000-2008",
                   "Investeringer 2009-2015",
                   "Betalingsbalance 2009-2015"))
```

I regressionsmodellerne får vi først opgivet værdien, der viser sammenhængen mellem de to variabler, og nedenunder i parantes vises standarderror. Stjernene viser i hvor høj grad, der er en statistisk sammenhæng og på hvilket siknifikansniveau, man kan forkaste H0. Hvis der ikke er en stjerne, så er der meget svag statistisk siknifikans mellem variablerne, og man ville formentlig ikke kunne bruge det til noget.

I vores tabel ser vi også, at alle værdier er negative. Det betyder, at når vores værdier i kolonnen med investeringer og betalingsbalancen stiger, så vil de tilhørende værdier i arbejdsløsheden falde.

Table 2: Regressions resultater

	Arbejdsløshed 2000-2015	Arbejdsløshed 2000-2008	Arbejdsløshed 2009-2015
	(1)	(2)	(3)
Investeringer 2000-2015	−0.183* β_1		
	(0.107) — standard error		
Betalingsbalance 2000-2015	−0.235*** β_2		
	(0.075)		
Investeringer 2000-2008		−0.153 β_1	
		(0.100)	
Betalingsbalance 2000-2008		−0.181** β_2	
		(0.069)	
Investeringer 2009-2015			−0.354*** β_1
			(0.099)
Betalingsbalance 2009-2015			−0.230*** β_2
			(0.082)
Constant	11.103*** β_0	10.270*** β_0	15.239*** β_0
	(2.475)	(2.313)	(2.318)
N	77	77	77
R^2	0.146	0.098	0.200
Adjusted R^2	0.123	0.074	0.178
Residual Std. Error (df = 74)	3.759	3.732	4.360
F Statistic (df = 2; 74)	6.336***	4.023**	9.249***

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

3.1.1.2 Hvad kan du sige om virkningerne af investeringer og løbende poster på arbejdsløshed?

I følge vores tabel, så kan vi sige, at der hele tiden er et negativt forhold. Før kriseperioden kan vi se, at det negative forhold ikke er så stærkt, men efter krisen øges koefficienten i en negativ retning, og det indikerer en stærkere statisitsk sammenhæng mellem investeringernes og betalingsbalancens effekt på arbejdsløsheden.

3.1.1.3 Diskuter hvorvidt estimaterne for de tre modeller er ens eller forskellige?

Som skrevet tidligere, så er der forskel på, hvorvidt siknifikansniveauet er stærkt eller svagt, som det er angivet ved stjernerne. Udover det, så er der ikke forskel på andet end, at den negative sammenghæng varierer i forskellige perioder, men den er stadigvæk negativ. Alt i alt er den eneste forskel på graden af den negative sammenhæng og hvorvidt resultatet er statistisk siknifikant eller ej.

3.1.1.4 Udfør hypotesetest på dine regressionsmodeller. Er disse resultater statistisk signifikante?

Vi skal opstille 2 nulhypoteser, fordi vi har 2 uafhængige variabler: Investeringer og betalingsbalancen: Opstiller nulhypotese 1: Investeringer har ingen effekt på ledigheden.

$H_0 : \beta_0 = 0$

$$H0 : \beta_1 = 0$$

$H_0 : \beta_0 \neq 0$

Alternativ hypotese:

$$H1 : \beta_1 \neq 0$$

Opstiller nulhypotese 2: Betalingsbalancen har ingen effekt på ledigheden:

$$H0 : \beta_2 = 0$$

Alternativ hypotese:

$$H1 : \beta_2 \neq 0$$

Her kan vi ved hjælp af R både få T-værdi og P-værdien. På denne måde, så kan vi hurtigt acceptere eller forkaste hypoteserne. H0 kan forkastes hvis T-værdien er:

$$T > 1.96, T < -1.96$$

Derudover kan H0 forkastes, hvis P-værdien er < 0.05. Disse værdier gælder, når man tester på et 95% siknifikansniveau.

Nu laver vi på vores hypotesetest på vores 3 modeller: Model 1, som er perioden 2000-2015:

```
summary(model1)
##
## Call:
## lm(formula = u_00_15 ~ inv_00_15 + c_00_15, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9234 -2.0666 -0.0265  1.8927 16.0577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.10327    2.47502   4.486 2.61e-05 ***
## inv_00_15   -0.18338    0.10678  -1.717  0.09009 .
## c_00_15     -0.23498    0.07485  -3.140  0.00243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.759 on 74 degrees of freedom
## Multiple R-squared:  0.1462, Adjusted R-squared:  0.1231
## F-statistic: 6.336 on 2 and 74 DF,  p-value: 0.002884
```

Her kan vi se, at investeringerne ikke har en tilstrækkelig T eller P-værdi til at kunne forkastes nulhypotesen på et 95% siknifikansniveau, og der er derfor ikke en entydig sammenhæng mellem investeringer og arbejdsløshed.
Samtidig kan vi se, at betalingsbalancen har en T-værdi under -1.96 og P-værdien er under 0.05. Dette betyder, at vi kan forkaste vores nulhypotese på et 95% siknifikansniveau, og der derfor er en klar sammenhæng betalingsbalalcen og dens effekt på arbejdsløsheden.

I model 2, kan vi se, at det er fuldstændig samme resultat, som i model 1, med undtagelse af siknifikansniveauet på betalingsbalancen. Dette kan ses ved, at vi er gået en stjerne ned, og vi kan derfor ikke forkaste på samme siknifikansniveau.

```
summary(model2)
##
## Call:
## lm(formula = u_00_08 ~ inv_00_08 + c_00_08, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5357 -2.4331 -0.3786  2.1093 16.3078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 10.27043    2.31254   4.441 3.08e-05 ***
## inv_00_08   -0.15260    0.10012  -1.524   0.1317
## c_00_08     -0.18071    0.06896  -2.621   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.732 on 74 degrees of freedom
## Multiple R-squared:  0.09806,    Adjusted R-squared:  0.07368
## F-statistic: 4.023 on 2 and 74 DF,  p-value: 0.02196
```

I model 3, som er perioden efter krisen, så kan vi se en meget kraftig sammenhæng mellem både investeringer og betalingsbalancens effekt på arbejdsløsheden. Vi kan derfor sige, at vi afvise nulhypotesen på et siknifikansniveau på 95%. Resultaterne er derfor statistisk siknifikante.

```
summary(model3)
##
## Call:
## lm(formula = u_09_15 ~ inv_09_15 + c_09_15, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2326 -2.7039 -0.5137  1.8486 15.5822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.23867    2.31840   6.573 6.08e-09 ***
## inv_09_15   -0.35361    0.09944  -3.556 0.000661 ***
## c_09_15     -0.22988    0.08227  -2.794 0.006624 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.36 on 74 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:  0.1784
## F-statistic: 9.249 on 2 and 74 DF,  p-value: 0.0002597
```