

# STA3030F – INFERENCE STATISTICS

## ASSIGNMENT 1 - 2020

March 2, 2020

---

THIS ASSIGNMENT WILL BE DONE INDIVIDUALLY

### Question 1

The Income and Expenditure Survey conducted by Statistics South Africa (Stats SA) for a given year produced data containing the following information:

- PROVINCE: Province where the household is located (WC=Western Cape; EC=Eastern Cape).
- GENDER: the gender of the head of the household.
- Race: the race group of the head of the household (1=black; 2=coloured; 3=Indian/Asian; 4=white).
- AGE: the age of the head of the household.
- AREATYPE: whether the household is in an urban or rural area (1=urban; 2=rural).
- HHSIZE: total number of people living in the household.
- HHEXPENDITURE: total monthly expenditure of the household.
- HHINCOME: total monthly income of the household.

Instructions:

- On page 4 each student has been randomly assigned an index. Take note of your index (the number next to your student number).
- Download the file “incexp” from the folder “Projects” under “Resources” on Vula and save it as **incexp.txt**.  
NB: The next bullet point assumes that the file incexp.txt is saved under the folder “assignment” on the hard drive, i.e “c:\assignment” .
- Use the following command to read the data into R:

```
mydata = read.table( "c:\\assignment\\incexp.txt",header=TRUE).
```

- Run the following commands in R:

```
age = mydata$AGE[index:(index+99)]
hhinc = mydata$HHINCOME[index:(index+29)]
```

NB: Replace “index” in the above R commands with your index. Check that your two samples “age” and “hhinc” are of size 100 and 30, respectively.

Your task is to solve the following inference problems using bootstrap re-sampling methods.

- Carry out exploratory data analysis (use graphical approaches where possible) in order to understand your data and provide some summary statistics.
- Construct a 95% confidence interval for the population mean age. Provide an interpretation of your 95% confidence interval.
- Use bootstrap re-sampling theory to arrive at a conclusion about the following hypothesis test. You should do this by calculating the p-value for the hypothesis test. State any assumption made.

$$H_0 : \mu_{age} \leq 43 \quad H_1 : \mu_{age} > 43$$

- Use the bootstrap samples to generate the histogram of the distribution of the sample mean age. Comment on the shape of the distribution.
- Construct a 90 % confidence interval of  $\varepsilon_{0.5}$ , the median household income.

## Question 2

On page 5 you have been randomly assigned two other data sets (**Data1** and **Data2**). The data sets are all taken (and adapted) from real-world studies covered in the book “Statistics for Management and Economics” (2000).

The file “Descriptions.pdf” contains a brief description of each of the datasets and can be found under the folder “Projects” on Vula.

All datasets can also be found under the “Projects” folder on Vula.

Use Data set 1 (**Data1**) to perform the following tasks

- Test whether there are significant differences in means between the two groups by using bootstrapping (Assume equality of variances).
- Provide a 95% confidence interval for the difference in means using the bootstrap approach.
- Use bootstrapping to test the variances for equality.
- Compare the results from bootstrapping and normal theory in a brief paragraph.

### Question 3

Use data set 2 (**Data2**) to test whether there are significant differences in means between the groups by using the bootstrap approach.

Your brief report should contain the following.

- (a) A clear statement of the hypotheses being tested.
- (b) A print-out of the first 3 bootstrap samples generated (In the appendix).
- (c) The p-value and conclusion based on the bootstrap results.
- (d) A one (short) paragraph interpretation of the results.

Note:

- For bootstrapping questions, use 4000 replications for question 1 and question 3; and use 5000 replications for question 2.
- Your project must include an introduction stating the objective of your study, a brief description of the raw data and your answers to all three questions. You must also include a plagiarism statement.
- **DUE DATE: 11 March 2020 before 14:00 (hand in a hard copy to reception and submit a soft copy on Vula)**

# QUESTION 1

USER ID	Index	USER ID	Index	USER ID	Index
ahnsoo002	131	jhnric014	166	pncnur001	246
allnae003	238	kglkhw001	306	pntste003	272
ancbuk001	144	kngann004	217	prkjoo001	266
blybon005	268	kngfin001	19	prkmah005	49
bgndee001	251	krjor001	60	ptlsah004	84
bggada002	167	lgnthi001	189	ptlles002	148
blxvhu001	124	lrxavi001	161	pllum011	216
bckang002	150	mbnthe023	320	plltri003	11
brdca001	133	mdnyan002	240	pswsis002	242
bthbon016	225	mhlslp057	294	rmtref003	18
errtri001	143	mhzkan001	207	rplher001	70
clxsip004	105	mlwsiy005	174	rsssam008	221
chkkur002	160	mlnrud002	255	slsgra001	274
chtsas002	216	mmgana001	267	slpleb004	99
chbmil003	186	mshpea001	77	shpman002	140
chksha007	167	msmnka004	244	shrrom003	189
chypai001	252	mttnin001	15	sbtelv002	121
cltane002	114	mbdnts001	89	sgwnka001	244
chnroy002	235	mddlut001	137	smpsed001	248
crwoli001	128	mlsluk002	202	sngdan006	127
cthrio001	333	mkhtha116	331	stwluc001	132
dkxqha001	155	mtlogo001	100	tljjam001	11
dcknab001	340	mlxcel001	188	tylles001	243
dbzsen001	292	mrtsar006	47	thsthe004	290
dynsak002	58	mndngo004	114	tlleda001	182
emrsam001	78	mphbok002	266	vzyhei003	243
estang001	242	mtshlu001	263	wynlyn004	147
evrll001	7	mdxkun001	303	wttmic012	68
frnsol002	155	mtstin007	269	ylzbuh001	12
frddan010	45	ndleda001	56	zlxnom027	80
frnkea002	117	nngvhu002	199		
fnduna001	294	ntsndi017	259		
frmgre001	166	ngtsim002	71		
gsxjos001	65	ngwdum008	206		
grtnin001	83	ngxawo001	127		
gqwmpu001	276	nkkzol001	98		
grysil001	84	nngbuh001	273		
gxtsol001	197	nrrmax001	287		
hmljoe001	217	nxmntu001	302		
hncdan001	311	nyvenn001	22		
hrsemm001	264	nzmfez001	118		

# QUESTIONS 2 & 3

USER ID	Data1	Data2	USER ID	Data1	Data2	USER ID	Data1	Data2
ahnsoo002	B2	aov3	jhnric014	A1	aov1	pncnur001	A2	aov5
allnae003	E2	aov6	kgkhkw001	B2	aov5	pntste003	A2	aov1
ancbuk001	A1	aov3	kngann004	C1	aov6	prkjoo001	E3	aov5
blybon005	B1	aov3	kngfin001	B2	aov2	prkmah005	A3	aov6
bgndee001	C1	aov7	krjor001	B2	aov7	ptlsah004	E2	aov6
bggada002	A2	aov4	lgnthi001	C1	aov2	ptlles002	C1	aov6
blxvhu001	A3	aov1	lrxavi001	C2	aov1	pllhum011	E2	aov5
bckang002	B1	aov2	mbnthe023	E2	aov4	pllttri003	E1	aov2
brdrca001	A1	aov1	mdnyan002	B1	aov3	pswsis002	C2	aov2
bthbon016	A3	aov7	mhlslp057	B2	aov6	rmtref003	E1	aov5
errtri001	A1	aov5	mhzkan001	B2	aov2	rplher001	B2	aov6
clxsip004	B2	aov2	mlwsiy005	E1	aov6	rsssam008	E2	aov2
chkkur002	E1	aov7	mlnrud002	B2	aov4	slsgra001	A1	aov7
chtsas002	E1	aov3	mmgana001	E2	aov3	slpleb004	B1	aov7
chbmil003	B1	aov6	mshpea001	E3	aov4	shpman002	C1	aov3
chksha007	E1	aov1	msmnka004	E2	aov6	shrrom003	A2	aov7
chypai001	C1	aov3	mttnin001	C2	aov7	sbtelv002	A3	aov5
cltane002	E2	aov2	mbdnts001	E3	aov6	sgwnka001	B1	aov7
chnroy002	A3	aov7	mddlut001	A2	aov3	smpsed001	A1	aov1
crwoli001	A1	aov3	mlsluk002	B1	aov2	sngdan006	A3	aov4
cthrio001	B2	aov2	mkhtha116	A2	aov6	stwluc001	A1	aov5
dkxqha001	A3	aov4	mtlogo001	B2	aov1	tljjam001	B2	aov6
dcknab001	B1	aov2	mlxcel001	A2	aov1	tylles001	E1	aov2
dbzsen001	E1	aov1	mrtzar006	A1	aov4	thsthe004	E1	aov5
dynsak002	E2	aov4	mndngo004	A1	aov5	tlleda001	B1	aov3
emrsam001	A3	aov2	mphbok002	A3	aov3	vzyhei003	E1	aov5
estang001	A1	aov1	mtshlu001	E1	aov3	wynlyn004	C1	aov7
evrll001	B1	aov4	mdxkun001	A3	aov1	wttmic012	E2	aov5
frnsol002	A2	aov1	mtstin007	A1	aov6	ylzbuh001	A3	aov7
frddan010	A2	aov4	ndleda001	C1	aov4	zlxnom027	A1	aov2
frnkea002	B2	aov1	nngvhu002	B1	aov7			
fnduna001	A1	aov5	ntsndi017	E3	aov4			
frmgre001	B1	aov4	ngtsim002	E3	aov1			
gsxjos001	B2	aov7	ngwdum008	C1	aov7			
grtnin001	E2	aov4	ngxawo001	E1	aov1			
gqwmpu001	E2	aov6	nkkzol001	B1	aov2			
grysil001	A3	aov6	nngbuh001	A3	aov3			
gxtsol001	B1	aov5	nrrmax001	A2	aov4			
hmljoe001	A3	aov3	nxmntu001	A2	aov3			
hncdan001	B2	aov5	nyvenn001	E1	aov7			
hrsemm001	A3	aov4	nzmfez001	A1	aov5			