MATH10282 Introduction to Statistics
Semester 2, 2017/18
Coursework assignment using R - mark scheme

Question 1 of 1

The data for this question are contained in the file `incomes.txt`, available on Blackboard. The data consist of the weekly incomes, in pounds (£), of 550 randomly selected individuals from a population.

(a) Read the data into a data frame called `incomes` in R. [2]

First download the file 'incomes.txt' from Blackboard, load up R and change the working directory to the folder where you saved the data. Then use the commmand:

```
incomes <- read.table(file="incomes.txt")
```

[2 marks - still award the marks if they didn't use the right variable name.]

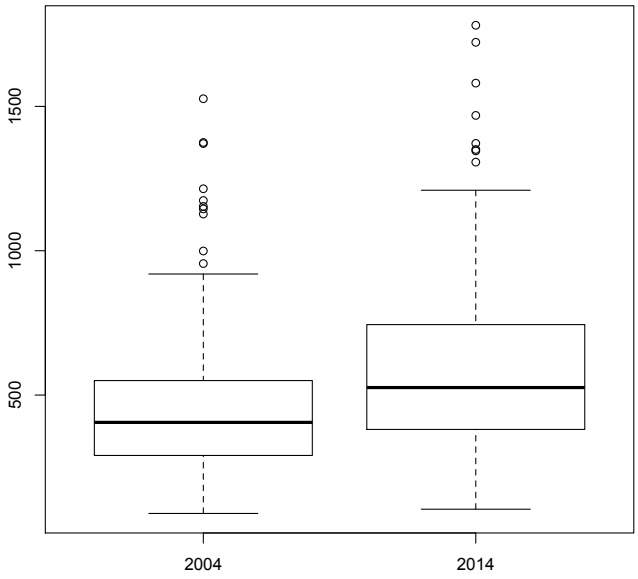
(b) We are now told by the data collection team that the first 300 rows are incomes from 2014, and the next 250 rows are incomes from 2004. Put this information into a second column in the data frame, called `incomes$year`. [*Hint: you may wish to use the command* `rep`.] [2]

```
incomes$year <- rep(c(2014,2004), c(300,250))
```

[2 marks - be lenient if they didn't save it in quite the right place, the main thing is the idea to use the rep command, and doing this reasonably correctly.]

(c) Draw a box plots for the income distributions in 2014 and another box plot for the income distribution in 2004. To facilitate comparison, put the two box plots on the same axes. Comment on the main features of the plots. Do you think that the distributions are symmetric? [4]

```
boxplot(income~year, data=incomes)
```



The incomes in 2014 are higher (presumably due to wage growth in the intervening 10 years). The distributions are positively skewed and there are many large outliers. All of the incomes are positive.

[2 marks for the plot + 2 marks for getting most of the main points.]

If they do two separate box plots with two different axes, e.g.

```
boxplot(incomes$income[1:300])
boxplot(incomes$income[301:550]}
```

then award one mark for the plot.

(d) Put the 2014 income values into a variable called `inc14`. [2]

```
inc2014 <- incomes$income[incomes$year==2014]   # subset the data
```

[2 marks. `incomes$income[1:300]` is fine, and it they save it to a different variable name that is also fine.]

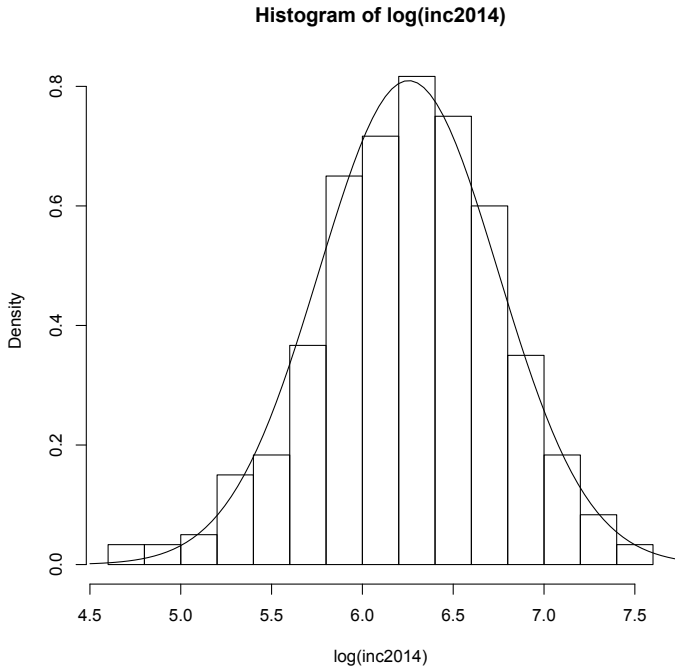
(e) Suggest a suitable transformation, $Y_i = g(X_i)$, that may enable a normal distribution, $Y_i \sim N(\mu_Y, \sigma_Y^2)$, to be fitted to the transformed 2014 data. Show that suitable estimates of the parameters are $\hat{\mu}_Y = 6.2541$, $\hat{\sigma}_Y^2 = 0.2428$. [3]

A log transformation may be suitable, i.e. $g = \log$. *We estimate* μ_Y *via* $\hat{\mu}_Y = \bar{y}$ *and* $\hat{\sigma}_Y = s_y$, *with values*

```
> ( mu.hat <- mean(log(inc2014)) )
6.254113
> ( sigma.hat <- sd(log(inc2014)) )
 0.4927045
> (sigma.hat^2)
0.2427577
```

(f) Plot a density histogram of the transformed income values, and superimposed the p.d.f. of the fitted normal density on the same plot. Comment on the goodness of fit. [3]

```
hist(log(inc2014),freq=F,breaks=20)
xs<-seq(from=4.5,to=8,length=100)
ys <- dnorm(xs, mean=mu.hat, sd=sigma.hat)
lines(xs,ys)
```



The fit seems good, as the heights of the histogram bars are close to the density curve.

[*2 marks for code + 1 mark for comment that fit is good (clear articulation of reason not needed)*]

(g) Use the fitted model to estimate the probability that a randomly selected individual in 2014 has a weekly income between £1000 and £1200. Compare this with the corresponding empirical probability. [4]

$$\hat{P}(1000 \leq Y \leq 1200) = \hat{P}(\log 1000 \leq \log Y \leq 1200)$$
$$= P(\log 1000 \leq N(6.2541, 0.2427^2) \leq \log 1200)$$
$$= P(N(6.2541, 0.2428^2) \leq \log 1200) - P(N(6.2541, 0.2428^2) \leq \log 1000)$$

```
> p1 <- pnorm(  log(1200), mean=mu.hat, sd=sigma.hat )
> p2<- pnorm(log(1000), mean=mu.hat, sd=sigma.hat)
> (p1-p2)
0.0474354
> mean( inc2014 < 1200 & inc2014 > 1000) # empirical probability
0.05
```

The model-based and empirical probabilities are close together, so there is no reason to doubt the fit of the model.

[2 marks for model-based probability (1 if they get the rough idea but wrong value); 1 mark for empirical probability; 1 mark for a sensible comment (e.g. close together)]

[Total 20 marks]