STATS 2107
Statistical Modelling and Inference II
Assignment 1

Jono Tuke

Semester 2 2017

CHECKLIST

- ☐: Have you shown all of your working, including probability notation where necessary?
- ☐: Have you given all numbers to 3 decimal places?
- ☐: Have you included all R output and plots to support your answers where necessary?
- ☐: Have you included all of your R code?
- ☐: Have you made sure that all plots and tables each have a caption?
- ☐: If before the deadline, have you submitted your assignment via the online submission on MyUni?
- ☐: Is your submission a single pdf file - correctly orientated, easy to read? If not, penalties apply.
- ☐: Penalties for more than one document - 10% of final mark for each extra document. Note that you may resubmit and your final version is marked, but the final document should be a single file.
- ☐: Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero.
- ☐: Assignments emailed instead of submitted by the online submission on MyUni will not be marked and will recieve zero.
- ☐: Have you checked that the assignment submitted is the correct one, as we cannot accept other submissions after the due date?

Due date: Friday 11th August 2017 (Week 3), 5pm.

Q1. Independence of S^2 and $\bar{X}$ in normal distributions

This question may be handwritten and then scanned to pdf.

Suppose $Y_1,\ Y_2,\ \ldots,\ Y_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables and let $\bar{Y}$ denote the sample mean.

(a) Find $E(\bar{Y}^2)$.

[2 marks]

(b) For each i with $1 \leq i \leq n$, prove that $\bar{Y}$ and $Y_i - \bar{Y}$ are uncorrelated, *i.e.*,

$$cov(Y_i - \bar{Y}, \bar{Y}) = 0.$$

[4 marks]
[Question total: 6]

Q2. Chi-square distributions

This question may be handwritten and then scanned to pdf.

Suppose $Z_1,\ Z_2,\ \ldots,\ Z_p$ are i.i.d. $N(0, 1)$ random variables and let

$$X = \sum_{i=1}^{p} Z_i^2.$$

(a) Find the moment generating function of X and hence, name the distribution of X.

You may assume, from the tutorial, that $M_{Z^2}(t) = (1-2t)^{-1/2}$, where $Z \sim N(0,1)$.

[5 marks]

(b) Suppose $Y_1, Y_2, \ldots, Y_p$ are independent normal random variables with different means and variances, that is,

$$Y_i \sim N(\mu_i, \sigma_i^2), i = 1, \ldots, p.$$

Show that

$$W = \sum_{i=1}^{p} \frac{(Y_i - \mu_i)^2}{\sigma_i^2} \sim \chi_p^2,$$

where χ_p^2 denotes the chi-squared distribution with p degrees of freedom.

[2 marks]
[Question total: 7]

Q3. Binomial estimators

This question may be handwritten and then scanned to pdf.

Let

$$X \sim Bin(n, p).$$

In this question, we are going to find an unbiased estimator for the parameter p^2. Let

$$\hat{p}^2 = \left(\frac{X}{n}\right)^2.$$

(a) Find $E[\hat{p}^2]$, and hence state the bias of $\hat{p}^2$.

[2 marks]

(b) Show that

$$E\left[\frac{\hat{p}(1-\hat{p})}{n-1}\right] = \frac{p(1-p)}{n}.$$

[3 marks]

(c) Using the first two parts, find an unbiased estimator for p^2.

[1 mark]
[Question total: 6]

Q4. Gumtree univariate analysis

This question should be typed up in latex, word, or Rmarkdown, then converted into a pdf for upload.

All tables and figures need to be captioned for full marks.

In each assigment, we will analyse the gumtree dataset that you see in the practicals. In the final assignment - Assignment 6 - you will put together your answers from each assignment to form a group project report.

The research goal is to find a linear model that can be used to predict price of a dog advertised on gumtree.

For each of the following variables:

- Price
- Cross

- Pet Offered By
- Microchip
- Vaccination
- Desexing status
- Relinquished or not

perform the following:

1. Clean each of the variables using the methods described in Practical 1. For full marks you must include **commented** code and the output to explain why and how you cleaned each variable. Also state whether the variable is quantitative or categorical.
2. For each of the variables, produce an appropriate plot to look at the data. For categorical variables produce a bar-chart, and for quantitative variables produce a histogram. Include all the plots in your assignment, ensuring that they are **labelled** and **captioned**.
3. For each quantitative variable, identify whether it is unimodel or bimodel, also whether it is symmetric, left-skewed or right-skewed. For the categorical variables identify the most common level.

As an example, here is my version of the first variable for you.

Price

The variable price is a quantitative continuous random variable.

Cleaning

First we check that if there are missing values:

```
table(gumtree$price)
```

```
##
##    0    1  100 1000  110 1100  120 1200 1250  130 1300 1350  140 1400 1499
##  576    9  190   72    1    9    8   41    8    7   12    3    1    7    1
##  150 1500 1600  170 1700  174  175  180 1800  190 1900 1950   20  200 2000
##  170   78   10    1    4    1    1    4   22    2    2    2    1  236   38
## 2100  220 2200 2300 2400   25  250 2500 2600 2700  275  280    3  300 3000
##    1    2   12    3    2    1  156   34    2    1    2    5    1  201   14
## 3250  330   35  350 3500  370  375  380  390 3900  395 3950   40  400 4000
##    1    1    1  127   12    1    3    4    2    1    1    1    2  183    4
##  420  425  440  450 4500  465  470  475  480  485  490  495  499   50  500
##    1    1    1   99    5    1    1    1    5    2    4    4    6   75  392
## 5000  550 5500   60  600 6000  650 6500  670   70  700 7000  730   75  750
##   11   29    3    5   62    7   27    1    1    5   36    2    1    3   12
##  799   80  800 8000  808  850   90  900  950  999   NA
##    1   18   57    2    1   13    2   28   12    3  212
```

As there are values `NA`, I convert them to `NA`:

```
gumtree$price[gumtree$price == "NA"] <- NA
```

No, I check that this worked:

```
table(gumtree$price)
```

```
##
##    0    1  100 1000  110 1100  120 1200 1250  130 1300 1350  140 1400 1499
##  576    9  190   72    1    9    8   41    8    7   12    3    1    7    1
```

```
##  150 1500 1600  170 1700  174  175  180 1800  190 1900 1950   20  200 2000
##  170   78   10    1    4    1    1    4   22    2    2    2    1  236   38
## 2100  220 2200 2300 2400   25  250 2500 2600 2700  275  280    3  300 3000
##    1    2   12    3    2    1  156   34    2    1    2    5    1  201   14
## 3250  330   35  350 3500  370  375  380  390 3900  395 3950   40  400 4000
##    1    1    1  127   12    1    3    4    2    1    1    1    2  183    4
##  420  425  440  450 4500  465  470  475  480  485  490  495  499   50  500
##    1    1    1   99    5    1    1    1    5    2    4    4    6   75  392
## 5000  550 5500   60  600 6000  650 6500  670   70  700 7000  730   75  750
##   11   29    3    5   62    7   27    1    1    5   36    2    1    3   12
##  799   80  800 8000  808  850   90  900  950  999
##    1   18   57    2    1   13    2   28   12    3
```

All good, so now I check type of variable

```
class(gumtree$price)
```

```
## [1] "character"
```

As I know that price is a quantitative, we change it to numeric:

```
gumtree$price  <- as.numeric(gumtree$price)
```

Univariate plot

As price is a quantitative continuous random variable, I produce a histogram (Figure 1).

```
ggplot(gumtree,aes(x = price)) +
  geom_histogram(col = "black", fill = "orange") +
  labs(x = "Price of dog in dollars.")
```

Discussion

The random variable price appears to be unimodal and right skewed.

[24 marks]

[Question total: 24]

[[Assignment total: 43]]

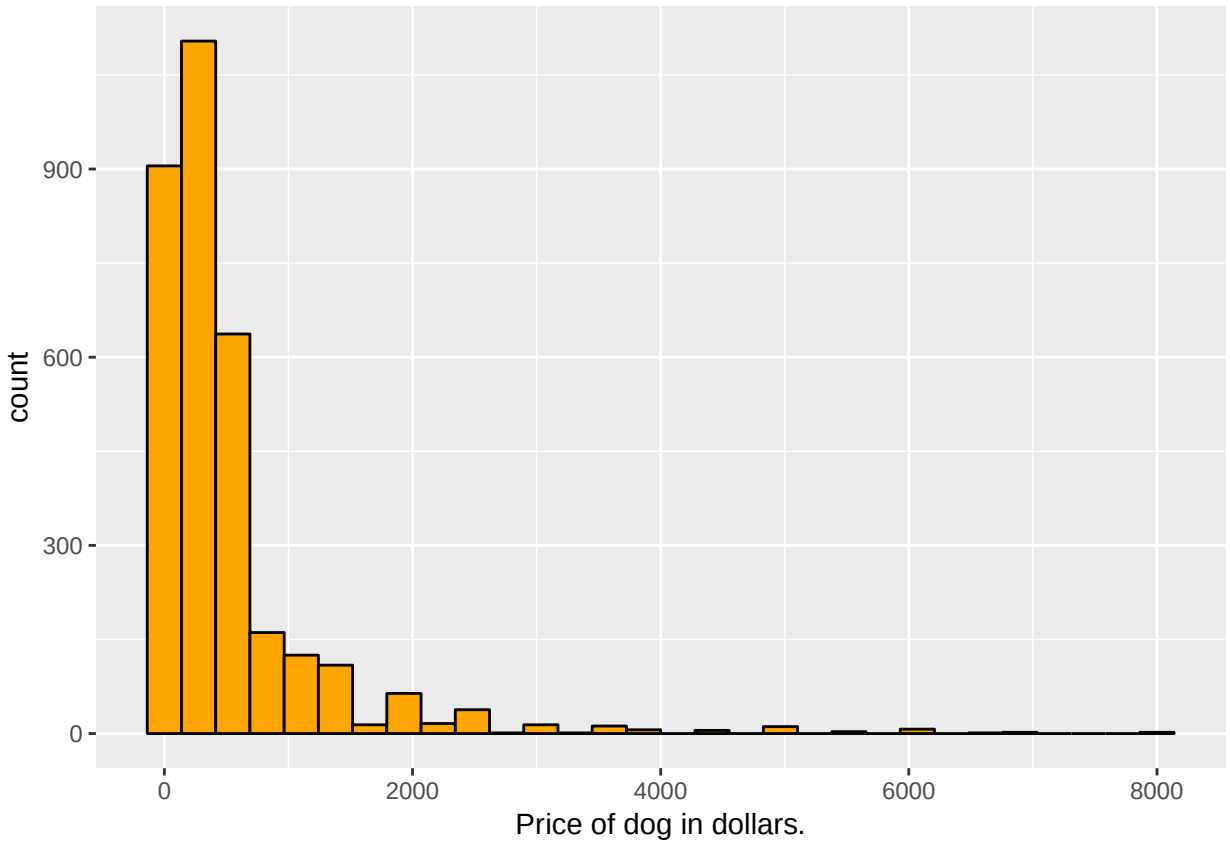

Figure 1: Histogram of price of advertised dogs on gumtree.