University of New South Wales

MATH 2901

Higher Theory of Statistics

Assignment 1

Keegan Gyoery (z5197058), Edward McInnes (z5162873),
Alex Robinson (z5164884), Ruby Smith (z5113171)

March 15, 2018

1. (a) If event A is independent of itself, by the definition of independence, the following result holds.

$$\begin{aligned}
\mathbb{P}(A \cap A) &= \mathbb{P}(A)\mathbb{P}(A) \\
&= \big[\mathbb{P}(A)\big]^2 \\
\mathbb{P}(A) &= \big[\mathbb{P}(A)\big]^2 \\
\therefore \big[\mathbb{P}(A)\big]^2 - \mathbb{P}(A) &= 0 \\
\therefore \mathbb{P}(A)\big[\mathbb{P}(A) - 1\big] &= 0 \\
\therefore \mathbb{P}(A) &= 0 \quad \text{OR} \\
\mathbb{P}(A) &= 1
\end{aligned}$$

   (b) Suppose that event A has probability $\mathbb{P}(A) = 1$, and the event B has some probability $\mathbb{P}(B)$. Thus the following consequences arise.

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(B) \\
\mathbb{P}(A)\mathbb{P}(B) &= 1 \times \mathbb{P}(B) \\
&= \mathbb{P}(B) \\
\therefore \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B)
\end{aligned}$$

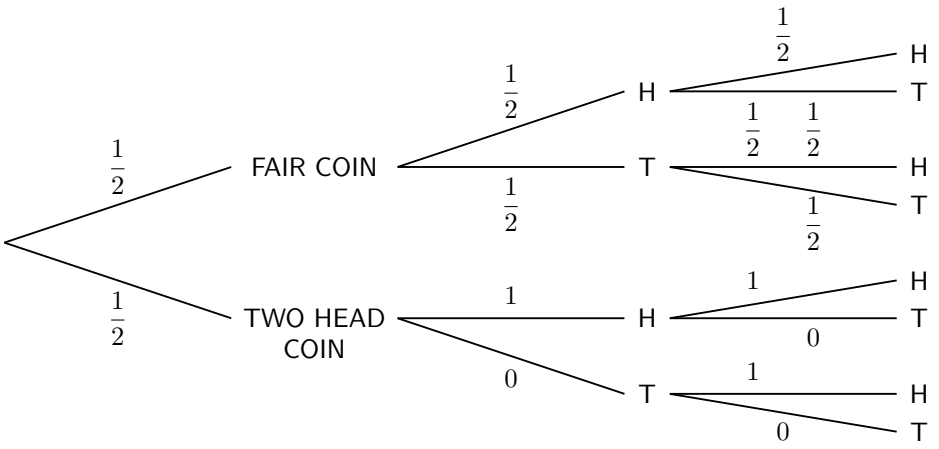
   Thus if event A has probability $\mathbb{P}(A) = 1$, events A and B are independent.

   Suppose now that event A has probability $\mathbb{P}(A) = 0$, and the event B has some probability $\mathbb{P}(B)$.

$$\begin{aligned}
\mathbb{P}(A \cap B) &= 0 \\
\mathbb{P}(A)\mathbb{P}(B) &= 0 \times \mathbb{P}(B) \\
&= 0 \\
\therefore \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B)
\end{aligned}$$

   Again, if event A has probability $\mathbb{P}(A) = 0$, events A and B are independent.

2. The following probability tree will be used to answer question 2. Furthermore, the notation FC will denote the Fair Coin.



(a)

$$\begin{aligned}\mathbb{P}(FC \mid H) &= \frac{\mathbb{P}(FC \cap H)}{\mathbb{P}(H)} \\ &= \frac{\frac{1}{2} \times \frac{1}{2}}{\left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times 1\right)} \\ &= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} \\ &= \frac{1}{3}\end{aligned}$$

Thus the probability of choosing the fair coin given that the coin shows heads after the first flip is $\frac{1}{3}$

(b)

$$\begin{aligned}\mathbb{P}(FC \mid H, H) &= \frac{\mathbb{P}(FC \cap H, H)}{\mathbb{P}(H, H)} \\ &= \frac{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}}{\left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times 1 \times 1\right)} \\ &= \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{2}} \\ &= \frac{1}{5}\end{aligned}$$

Thus the probability of choosing the fair coin given that the coin shows heads after the first flip, and heads after the second flip, is $\frac{1}{5}$

(c)

$$\begin{aligned}\mathbb{P}(FC \mid H, H, T) &= \frac{\mathbb{P}(FC \cap H, H, T)}{\mathbb{P}(H, H, T)} \\ &= \frac{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}}{\left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times 1 \times 1 \times 0\right)} \\ &= 1\end{aligned}$$

Thus the probability of choosing the fair coin given that the coin shows heads after the first flip, heads after the second flip, and tails after the third flip, is 1

3.

4.

5. (a)

$$\begin{aligned} F_X(x) &= \int_1^2 f_X(x)dx \\ &= 2\int_1^2 \frac{1}{x^2}dx \\ &= 2\left[\frac{-1}{x}\right]_1^2 \\ &= 2\left[\frac{-1}{2}+1\right] \\ &= 1 \end{aligned}$$

(b)

$$\begin{aligned} \mathbb{E}(X) &= \int_1^2 x f_X(x)dx \\ &= 2\int_1^2 \frac{1}{x}dx \\ &= 2\ln x\,\Big|_1^2 \\ &= 2[\ln 2 - \ln 1] \\ &= 2\ln 2 \end{aligned}$$

Let M be the location of the median.

$$\begin{aligned} F_X(x) &= \frac{1}{2} \\ \therefore \frac{1}{2} &= \int_1^M f_X(x)dx \\ &= 2\int_1^M \frac{1}{x^2}dx \\ &= 2\left[\frac{-1}{x}\right]_1^M \\ &= 2\left[\frac{-1}{M}+1\right] \\ &= \frac{-2}{M}+2 \\ \therefore \frac{2}{M} &= \frac{3}{2} \\ \therefore M &= \frac{4}{3} \end{aligned}$$

6.

7. (a) After importing the dataset into R, it tells us directly that there are 799 entries. Hence there are 799 observations in the Nervedata set.

(b)

Min.	0.50
1st Qu.	3.50
Median	7.50
Mean	10.95
3rd Qu.	15.00
Max.	69.00

Note that our dataset is labelled Nervedata. The code `summary(Nervedata)` output what is to the left. This returned the minimum and maximum values, the value of each quartile and the mean; all of which are important numbers to assess when analysing data.

(c) To produce a boxplot in R, the command `boxplot(Nervedata, main = "Time Between Successive Nerve Pulses", ylab = 1/50ths of a second)` was used, returning figure 1. As seen in figure one, the distribution is heavily skewed towards the lower times between nerve pulse. This is evidenced by the fact that the range of pulses goes from 0.5 to 69 seconds, but the median is 1.95 seconds. As the time between nerve pulses increases, the density of data points decreases. Also, there are many outliers that lay outside the box plot. These outliers are roughly defined as having a time value of greater than 32 seconds.

(d)

Min.	-0.6391
1st Qu.	1.2528
Median	2.0149
Mean	1.9113
3rd Qu.	2.7081
Max.	4.2341

i. To transform the data set by a logarithm, the code `LogNervedata = log(Nervedata)` was used. Then, the code `summary(LogNervedata)` was used to provide the numbers on the left. Most notably taking the logarithm of the data set has brought all the important values such as quartiles and the mean closer together.

ii. To produce a boxplot in R of the LogNervedata set, the code `boxplot(LogNervedata, main = "Log of the Time Between Successive Nerve Pulses", ylab = 1/50ths of a second)` was used, and returned figure 2. As seen from the boxplot, by taking the logarithm of the data, the distribution has become much less heavily skewed. Majority of the data points now occur between 1 and 3, and the max and min points are 4.23 and -0.69 respectively. Also, there are no outliers, as all points occur within the boxplot, which is in contrast to the original plot, where there were a large amount of outliers existing outside the plot. The boxplot shows that under the logarithm, this dataset becomes more equally distributed, resulting in a significantly more symmetric distribution compared to the original dataset.

(e)

Min.	0.7071
1st Qu.	1.8708
Median	2.7386
Mean	2.9694
3rd Qu.	3.8730
Max.	8.3066

i. To transform the data set appropriately, the code `SquarerootNervedata = ((Nervedata)^(1/2))` was used. Then to provide a numerical summary of this new distrubtion, the code `summary(SquarerootNervedata)` was used.

ii. To produce a boxplot in R of the SquarerootNervedata set the code `boxplot(SquarerootNervedata, main = "Square root of the Time Between Successive Nerve Pulses", ylab = 1/50ths of a second)` and returned figure 2. As seen from the boxplot, by taking the square root of the data, an effect similar to taking the logarithm has occurred. It has reduced the extent of the skewness towards lower values, however unlike the logarithm, this skewness is still clearly visible. Outliers still exist, as seen by the data points which lay outside the plot, however the number of outliers has been significantly reduced when compared to the original plot. Furthermore, in comparison to the original plot, the quartiles have been brought much closer together, as have the maximum and minimum values. Finally, this transformation has resulted in a more symmetric distribution in comparison to the original dataset.

(f)

Min.	0.25
1st Qu.	12.25
Median	56.25
Mean	229.19
3rd Qu.	225.00
Max.	4761.00

i. To transform the data set appropriately, the code `SquareNervedata =((Nervedata)^2)` was used. Then, to provide a numerical summary of this new distribution, the code `summary(SquareNervedata)` was used.

ii. To produce a boxplot in R of the SquareNervedata set, the code `boxplot(SquareNervedata, main = "Square of the Time Between Successive Nerve Pulses", ylab = 1/50ths of a second)` and returned figure 4. As seen from the boxplot, by taking the square of the data has drastically increased the spread of the data points, now ranging from a min of 0.25 to a max of 4761. The plot still shows a clear skewness towards lower values. In addition, the amount of outliers, like the original plot, remains very high. Also, the 1st, 2nd and 3rd quartiles, like the data points, have been spread apart. This transformation has definitely resulted in a less symmetric distribution than the original plot.

(g) Transformation 1: $\frac{1}{\text{original data}}$

Min.	0.01449
1st Qu.	0.06667
Median	0.13333
Mean	0.28376
3rd Qu.	0.28571
Max.	2.00000

i. To transform the data set appropriately, the code `ReciprocalNervedata = (1/Nervedata)` was used. Then to provide a numerical summary of this new distribution, the code `summary(ReciprocalNervedata)` was used.

ii. To produce a boxplot in R of the ReciprocalNervedata set the code `boxplot(ReciprocalNervedata, main = "Reciprocal of the Time Between Successive Nerve Pulses", ylab = HELP HERE` and returned figure 5. As seen from the boxplot, by taking the reciprocal of the data, an effect similar to taking the logarithm or square root has occurred. It has significantly reduced the amount of outliers compared to the original plot, although few still exist quite sporadically. The skewness towards lower values is clearly visible in this plot. Furthermore, compared to the original plot, the quartiles have been brought much closer together, as have the maximum and minimum values. Thus, this transformation has resulted in a more symmetric distribution in comparison to the original dataset.

Transformation 2: Cube of original data

Min.	0.1
1st Qu.	42.9
Median	421.9
Mean	6916.3
3rd Qu.	3375.0
Max.	328509.0

i. To transform the data set appropriately, the code `CubeNervedata = (1/Nervedata) HELP NEEDED????` was used. Then to provide a numerical summary of this new distribution, the code `summary(CubeNervedata)` was used.

ii. To produce a boxplot in R of the CubeNervedata set the code `boxplot(CubeNervedata, main = "Cube of the Time Between Successive Nerve Pulses", ylab = 1/50ths of a second)` and returned figure 6. As seen from the boxplot, by taking the cube of the data has drastically increased the spread of the data points, now ranging from a min of 0.1 to a max of 328509. The plot shows a clear skewness towards lower values. In addition, the amount of outliers remains extremely high, the highest out of all the different transformations. Also, the 1st, 2nd and 3rd quartiles, like the data points, have been spread apart. This transformation has definitely resulted in a less symmetric distribution than the original plot.

(h) The log transformation resulted in the most symmetric data for reasons explained in (d) part ii, as well as by assessing the log plot compared to all other transformations.