FIT2086 ASSIGNMENT 2 SOLUTIONS

Question 1

1.1. The estimate for the sample mean is 65.75. By the standard formula, the 95% confidence interval is:

$$\begin{aligned}&\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right)\\ &= \left(65.75 - 1.96\frac{15.23}{\sqrt{8}}, 65.75 + 1.96\frac{15.23}{\sqrt{8}}\right)\\ &= (55.1961484481, 76.3038515519)\end{aligned}$$

The 95% confidence interval lies between 55.196 and 76.304. The conclusion can look like this:

"The estimated mean blood pressure of people from the Pima ethnic group with diabetes (sample size $n = 8$) is 65.75. We are 95% confident the population mean BMI for this group is between $55.196mmHg$ and $76.304mmHg$."

1.2. The mean blood pressure in Pima indians without diabetes is 68.1. The estimated mean difference is simply one mean minus the other, which is $65.75 - 68.1 = -2.35$ (that is, the population with diabetes has a higher mean blood pressure than the population without it).

The confidence interval can be calculated using the standard formula

$$\begin{aligned}min &= \hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\\ max &= \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\end{aligned}$$

Plugging in values, we get

$$\begin{aligned}min &= -2.35 - 1.96\sqrt{15.23^2/8 + 14.69^2/10} = -16.2886\\ max &= -2.35 + 1.96\sqrt{15.23^2/8 + 14.69^2/10} = 11.5886\end{aligned}$$

This yields the confidence interval (-16.29, 11.59). Because 0 lies within the 95% confidence interval, we cannot reject the possibility that these came from the same group. The conclusion can look like this:

"The estimated difference in mean blood pressure between people from the Pima ethnic group with diabetes (sample size $n = 8$) and without diabetes (sample size $n = 10$) is $-2.35mmHg$. We are 95% confident the population mean difference in blood pressure is between $16.29mmHg$ (blood pressure is lower in people with diabetes) up to $11.59mmHg$ (blood pressure is greater in people with diabetes). As the interval includes zero, we cannot rule out

the possibility of there being no difference at a population level between Pima people with and without diabetes."

If we swap the order we might get a difference of means of positive 2.35 instead of a negative - in this case the confidence interval is -11.5886 to 16.2886 and the summary should be appropriately reflect this.

1.3. We wish to determine whether the two groups are the same or not. Our null hypothesis, H_0, is that the means are the same at the population level, and our alternative hypothesis is that the means are not the same at the population level. That is,

$$H_0 : \mu_d = \mu_n$$
$$H_A : \mu_d \neq \mu_n$$

Because we know the variance for both populations, we can use the normal distribution to find our p-value. We use the standard formula:

$$\hat{\mu}_x - \hat{\mu}_y \sim N\left(0, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

Plugging in our values, we get:

$$\hat{\mu}_x - \hat{\mu}_y \sim N\left(0, \frac{15.23^2}{8} + \frac{14.69^2}{10}\right)$$
$$\therefore \hat{\mu}_x - \hat{\mu}_y \sim N(0, 50.573)$$

Our observed difference of means was -2.35. Our z-score for this difference of means is therefore

$$z = \frac{x - \mu}{\sigma}$$
$$\therefore z = \frac{-2.35}{\sqrt{50.573}}$$
$$\therefore z = -0.33045$$

Our p-value then is $2\,\mathbb{P}(Z < -|z|) = 2 \times 0.37053 = 0.74106$. Therefore, we have only very weak evidence against the null hypothesis that the population mean blood pressure of Pima indian people with and without diabetes is the same, and cannot reject the null.

Question 2

2.1. You can use some code like this:

```
y_s <- 1:10
mu_s <- c(1, 2, 4)
```

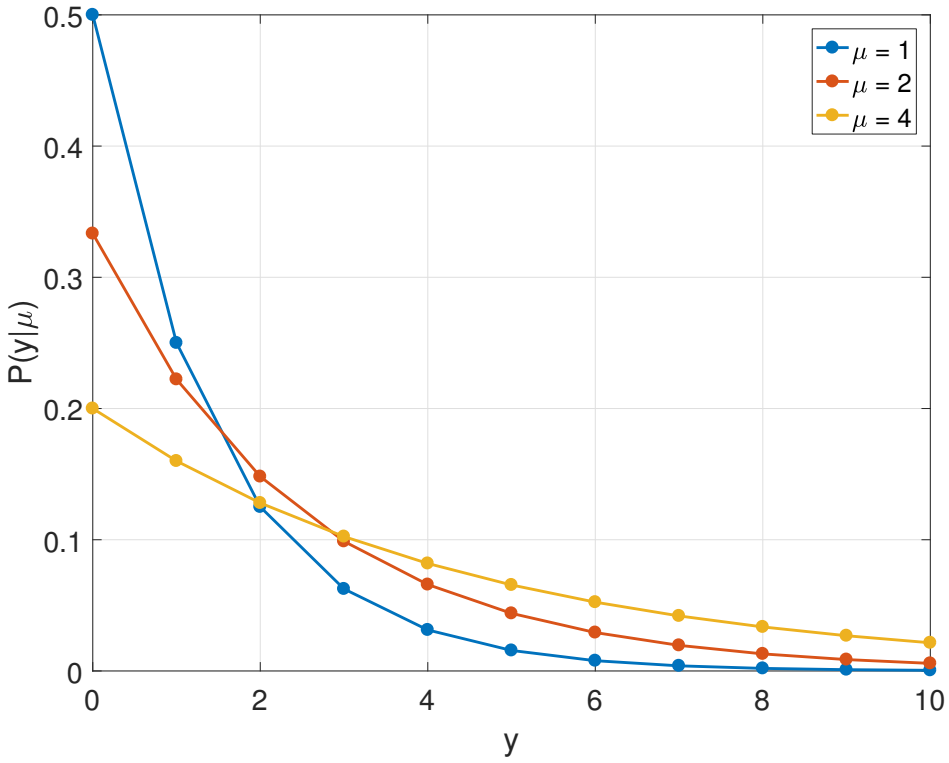

FIGURE 1. Geometric probability mass function plot

```
geo <- function (y, mu) {
  term_a <- mu / (1 + mu)
  term_b <- 1 / (1 + mu)

  return ( (term_a ^ y) * term_b)
}

scatters <- array(dim=c(10, length(mu_s)))

for (i in 1:length(mu_s)) {
  scatters[,i] = geo(y_s, mu_s[i])
}

print(scatters)

matplot(scatters, pch=1)
```

The plot should look something like Figure 1.

2.2. Since each y comes from an i.i.d. population, we can just multiply all the probabilities together to yield:

$$\begin{aligned}
&\prod_{i=1}^{n} P(y_i|\mu) \\
&= \prod_{i=1}^{n} \left(\frac{\mu}{1+\mu}\right)^{y_i} \left(\frac{1}{1+\mu}\right) \\
&= \left(\frac{\mu}{1+\mu}\right)^{m} \left(\frac{1}{1+\mu}\right)^{n} \text{ where } m = \sum_{i=1}^{n} y_i
\end{aligned}$$

2.3.

$$\begin{aligned}
& -\log\left(\left(\frac{\mu}{1+\mu}\right)^{m} \left(\frac{1}{1+\mu}\right)^{n}\right) \\
= \quad & -m(\log\mu - \log(1+\mu)) + n\log(1+\mu) \\
= \quad & (m+n)\log(1+\mu)) - m\log(\mu)
\end{aligned}$$

2.4. For an estimator with likelihood L, we find $\hat{\mu}$ such that $\frac{\partial}{\partial\mu}(-\log L) = 0$ for $\mu = \hat{\mu}$.

$$\begin{aligned}
&\frac{\partial}{\partial\hat{\mu}}(m+n)\log(1+\hat{\mu})) - m\log(\hat{\mu}) \\
&= \frac{m+n}{1+\hat{\mu}} - \frac{m}{\hat{\mu}} = 0 \\
&\therefore \hat{\hat{\mu}}(m+n) = m(1+\hat{\mu}) \\
&\therefore \hat{\mu}m + \hat{\mu}n = m + \hat{\mu}m \\
&\therefore \hat{\mu}n = m \\
&\therefore \hat{\mu} = \frac{m}{n}
\end{aligned}$$

which is the sample mean $\bar{Y}$.

2.5. The mean of the sample mean is $E[\bar{Y}] = \mu$, so the estimator is unbiased. The variance is $V[\bar{Y}] = \mu(1+\mu)/n$. Only award marks if both are correct.

3

3.1. Paul estimated $\frac{12}{14}$ matches correctly, giving him an approximate success rate of 0.8571, or 85.71%.

3.2. If Paul "got lucky" with random selection, then for a Bernoulli distribution we would expect to see $\theta = 1/2$. So, our hypothesis is

$$H_0 : \theta = 1/2$$
$$H_A : \theta \neq 1/2$$

From Paul's sample, we would estimate $\hat{\theta} = 12/14$. Using the standard formula, we calculate our z-score using the formula

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$$

where $\theta_0 = 1/2$.

In this case, we find $z = 2.67$. The probability of having a z-score less than -2.67 is approximately 0.00376, so the probability of the mean being at least as far away from 0.5 as 12/14 is $2 \times 0.00376 = 0.00752$. Thus our p-value for H_0 is 0.00752, which is much smaller than 0.05 (which is a reasonable boundary for statistical significance in this case). Therefore, this p-value suggests that our data does not support our hypothesis that Paul was guessing completely randomly, implying that Paul may have been choosing flags based on other factors.

As an alternative, I will also allow marks if students tested the one-sided hypothesis:

$$H_0 : \theta \leq 1/2$$
$$H_A : \theta > 1/2$$

i.e., that Paul is doing better than a random guess, in which case the p-value is $1 - P(Z < 2.67) = 0.003763$, though you should note to the students that in this case a two-sided test is more appropriate as we are interested to see if Paul's behaviour is *different* from random, not necessarily *better* than random.

3.3. Using the line

```
p.value <- binom.test(x=12, n=14, p=0.5)
```

we find our p-value to be 0.01294. The evidence against the null is weaker than the approximate test, but is still quite strong and suggests that we can reject the null hypothesis that he was guessing at random.

3.4. Given this analysis, it appears unlikely that Paul would be an oracle. It is notable that Paul usually chose the German flag, and Germany is a very good team which wins most of its games. Also, as one student noted in class, the first Spain v Germany match was predicted incorrectly but Paul veered from his "choose Germany" strategy the second time to secure a win, which might indicate some form of foul play. Perhaps they put the yummy food on the side more likely to win?

1. Question 4

4.1. All variables but `indus`, `chas` and `age` appear to be associated, though `chas` may be associated (p-value is borderline). The three strongest are either:

`lstat`, `dis` and `ptratio` – by p-value

or

`nox`, `rm` and `chas` (or `dis`, if they excluded `chas` due to its p-value) – by coefficient size.

4.2. The per-capita crime rate appears to have a negative correlation with the median house price, because its coefficient in the linear regression is negative (-0.191140). On the other hand, having frontage on the Charles River appears to have a positive correlation with the median house price, because its coefficient in the linear regression is positive (1.719044). So, it's better for house prices to have lower crime and be on the waterfront.

Ideally, the students would describe the effects in the following form, so please include this in your feedback:

- For every unit increase in per-capita crime rate, the median house price decreases by 0.191 thousand dollars.
- If a suburb has frontage on the Charles River, the median house price is higher by 1.719 thousand dollars, as compared to suburbs that do not have Charles River frontage.

4.3. Using the standard R step function, we find the following regression:

$$\begin{aligned}
\texttt{medv} \;=\; & 27.724105 + \\
& -0.190667 \times \texttt{crim} + \\
& 0.043092 \times \texttt{zn} + \\
& 1.814655 \times \texttt{chas} + \\
& -13.396245 \times \texttt{nox} + \\
& 4.852784 \times \texttt{rm} + \\
& -1.341063 \times \texttt{dis} + \\
& 0.464101 \times \texttt{rad} + \\
& -0.014154 \times \texttt{tax} + \\
& -0.789893 \times \texttt{ptratio} + \\
& -0.510858 \times \texttt{lstat}
\end{aligned}$$

4.4. If a council wants to improve median house prices, it could:

- Lower crime rates
- Zone more residential land
- Work to decrease the nitric-oxide concentration
- Provide incentives for residents to build more rooms in their homes

- Make highways more accessible
- Lower taxes
- Hire more teachers
- Attract "higher-status" residents, e.g. through regressive taxation

A suburb cannot be moved onto the Charles River, nor can it be moved closer to one of Boston's main employment centres.

4.5. We predict a median house price of 21.92865 for this house (which is in the thousands, so $21,928.65 is our estimate).

2. Question 5

This question is rather open-ended so marks can be given where students have made a reasonable and successful attempt to improve the accuracy of their model.

They might try some of the following things:

- Using interactions and non-linear transformations
- Trying other machine learning techniques
- Trying other tools (I know one student who wants to try it with sk-learn) - maybe someone will try TensorFlow for ultimate street cred?
- Who knows what else they'll try