# Midterm 2 Practice Questions (Part I)

S352

*Note: Practice the following questions to prepare for the written part of our in-class Midterm 2 Exam. In this part, you will be allowed to use a calculator and one double-sided (8.5 inches by 11.0 inches) sheet of handwritten notes.*

1. We have 100 IID observations from a Binomial distribution with $n = 3$ and $p$ unknown, given in the table below:

| Number of successes | Times observed |
|---|---|
| 0 | 11 |
| 1 | 34 |
| 2 | 40 |
| 3 | 15 |

   (a) Write down the likelihood as a function of $p$.

   (b) Let $\hat{p}$ be the MLE of $p$. $\hat{p}$ is the proportion of successes in all 300 trials. What is the value of $\hat{p}$?

   (c) What is the variance of $\hat{p}$? (Hint: Use the usual formula for the variance of sample proportion, $\hat{p}$.)

   (d) Suppose we wished to test the null hypothesis that $p = 0.5$. What is the Wald test statistic?

   (e) The P-value for the Wald test is about 0.30, which is not a small P-value. A statistics student concludes, "We can thus be sure that the the null hypothesis ($p = 0.5$) is true." Is this an appropriate conclusion? If not, give a better one.

2. I want to find out the average number of people per household in the U.S. I survey a simple random sample of U.S. households and obtain the results displayed in the table below.

| Household size | Number of households |
|---|---|
| 1 | 27 |
| 2 | 34 |
| 3 | 16 |
| 4 | 13 |
| 5 | 6 |
| 6 | 3 |
| 7 | 1 |

Now, household size doesn't have a Poisson distribution, because a Poisson has zeroes and you can't have zero people in a household. However, we can model the data as follows:

Let $Y = X + 1$ be the number of people in a randomly selected household, where $X$ has a Poisson distribution with parameter $\lambda$. Recall that this implies $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$.

(a) Find the mean household size for the 100 households in the sample. (It should be 2 point something.)

(b) Let $\hat{\lambda}$ be the MLE for $\lambda$. What is the numerical value of the MLE for $\lambda$?

(c) What is the variance of $\hat{\lambda}$?

(d) Find a 95% confidence interval for $\lambda$ using the Wald method.

(e) Find a 95% confidence interval for $\mathrm{E}(Y)$, the population mean household size.

3. In a population of numbers that obeys Benford's law, the probability that the first non-zero digit is $x$ is given by

$$f(x) = P(X = x) = \log_{10}(1 + 1/x)$$

for $x \in \{1, \ldots, 9\}$. Table 1 gives the numerical values of $f(x)$ to 4 decimal places.

(a) Suppose the first digits of the number of followers of T accounts obey the law (excluding accounts with no followers). If there are 38,663,000 T accounts, how many would we expect to have each first digit from 1 to 9?

(b) Table 1 shows the first digit of the number of followers for 38,663,000 accounts. Calculate the $G$-statistic for a likelihood ratio test of the hypothesis that the first digits of the number of T followers follow Benford's law.

| $x$ (First digit) | $f(x)$ | Number of accounts (thousands) |
|---|---|---|
| 1 | 0.3010 | 12614 |
| 2 | 0.1761 | 6443 |
| 3 | 0.1249 | 4563 |
| 4 | 0.0969 | 3581 |
| 5 | 0.0792 | 2951 |
| 6 | 0.0669 | 2533 |
| 7 | 0.0580 | 2227 |
| 8 | 0.0512 | 1988 |
| 9 | 0.0458 | 1763 |

Table 1: Distribution of the first digit of the number of followers for 38,663,000 T accounts.

(c) Comparing the statistic you calculated in (b) to a chi-squared distribution with _____ degrees of freedom gives a $P$-value of basically zero. What does this tell you?

4. A sleep researcher wishes to see which variables might help predict "sleep efficiency", which is the proportion of time in bed spent asleep. Their data set `sleep` includes measurements of the following variables on 438 individuals:

   - `Efficiency`: the proportion of time in bed spent asleep (between 0.5 and 0.99);
   - `Gender`: `Female` or `Male`;
   - `Alcohol`: ounces of alcohol consumed in the 24 hours prior to bedtime (between 0 and 5);
   - `Smoking`: whether or not the subject smokes (`No` or `Yes`);
   - `Age` of the subject in years (between 9 and 69.)

   (a) The researcher first wanted to see: is there a difference in average efficiency between females and males?

   Here's the result of a $t$-test:

```
> t.test(Efficiency ~ Gender, data = sleep)

Welch Two Sample t-test

data:  Efficiency by Gender
t = -0.24344, df = 431.92, p-value = 0.8078
alternative hypothesis: true difference in means between group Female and group Male i
95 percent confidence interval:
-0.02858878  0.02228741
sample estimates:
mean in group Female   mean in group Male
0.7872146            0.7903653
```

   Briefly answer the researcher's question.

   (b) Next the researcher wanted to predict efficiency using alcohol and smoking. They fit the following model:

```
> lm(Efficiency ~ Alcohol + Smoking, data = sleep)

Call:
lm(formula = Efficiency ~ Alcohol + Smoking, data = sleep)

Coefficients:
(Intercept)      Alcohol   SmokingYes
0.84985     -0.03092     -0.07285
```

   Explain what the numbers "$-0.03092$" and "$-0.07285$" mean.

   (c) Using the above model, write down a regression equation to predict efficiency from alcohol for *non-smokers*.

   (d) Using the above model, write down a regression equation to predict efficiency from alcohol for *smokers*.

(e) It turns out that adding an interaction improves the model:

```
> lm(Efficiency ~ Alcohol + Smoking + Alcohol:Smoking, data = sleep)

Call:
lm(formula = Efficiency ~ Alcohol + Smoking + Alcohol:Smoking,
data = sleep)

Coefficients:
(Intercept)             Alcohol          SmokingYes
0.83422            -0.01653            -0.02565
Alcohol:SmokingYes
-0.03791
```

Using the interaction model, write down a regression equation to predict efficiency from alcohol for *non-smokers*.

(f) Using the interaction model, write down a regression equation to predict efficiency from alcohol for *smokers*.

5. I have a data set called `gapminder07` that contains data (from 2007) on 142 countries of the world. Each row of the data set is a country. Variables include:

   - `lifeExp`: life expectancy of the country, in years
   - `gdpPercap`: the GDP per capita (a measure of income) of the country, in US dollars
   - `continent`: which continent the country is in: one of `Africa`, `Americas`, `Asia`, `Europe`, and `Oceania`.

   I fit a linear model to predict `lifeExp` using `continent` and the log of `gdpPercap` (with no interaction):

```
> gap.lm <- lm(lifeExp ~ continent + log(gdpPercap), data = gapminder07)
> gap.lm

Call:
lm(formula = lifeExp ~ continent + log(gdpPercap), data = gapminder07)

Coefficients:
(Intercept)  continentAmericas      continentAsia
20.138             11.694             10.114
continentEurope   continentOceania     log(gdpPercap)
11.268             12.929              4.631
```

(a) Gabon, a country in Africa, had a GDP per capita of $13,000 in the year 2007. What is the model's prediction for the life expectancy in Gabon?

(b) Find the regression line to predict life expectancy for countries in Asia as a function of `log(gdpPercap)`.

(c) Find the regression line to predict life expectancy for countries in the Americas as a function of `log(gdpPercap)`.

(d) On the axes provided on the following page, plot the regression lines for (b) and (c) for values of `log(gdpPercap)` from 7 to 10, labeling which line is which. Describe, using words and numbers, the difference between the two lines.

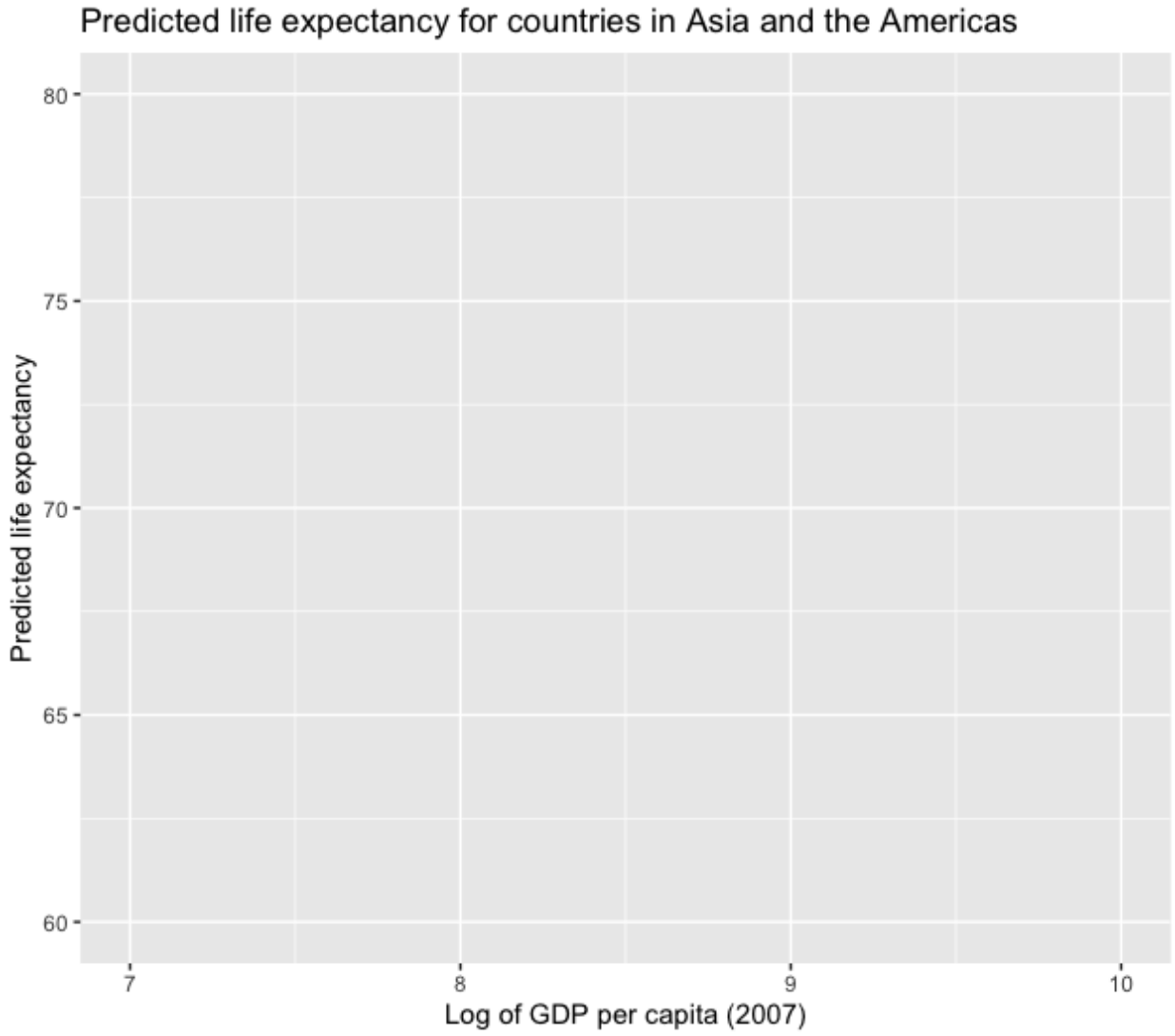


Figure 1: Graph axes for question 5(d).