Statistical Methods For Researchers
ST511: Homework #1

Due on September, 2019 at 11:59pm

Instructor: Dr. Shuting Wang Section 2

Fausto Rodríguez Zapata

Problem 1

1.3

In 2014, Congress cut $8.7 billion from the Supplemental Nutrition Assistance Program (SNAP), more commonly referred to as food stamps. The rationale for the decrease is that providing assistance to people will result in the next generation of citizens being more dependent on the government for support. Hoynes (2012) describes a study to evaluate this claim. The study examines 60,782 families over the time period of 1968 to 2009 which is subsequent to the introduction of the Food Stamp Program in 1961. This study examines the impact of a positive and policy driven change in economic resources available in utero and during childhood on the economic health of individuals in adulthood. The study assembled data linking family background in early childhood to adult health and economic outcomes. The study concluded that the Food Stamp Program has effects decades after initial exposure. Specifically, access to food stamps in childhood leads to a significant reduction in the incidence of metabolic syndrome (obesity, high blood pressure, and diabetes) and, for women, an increase in economic self-sufficiency. Overall, the results suggest substantial internal and external benefits of SNAP.

a. Describe the sample.
   R/ "60,782 families over the time period of 1968 to 2009" it is not clear how many families received SNAP benefits.

b. What characteristics of the population are of interest to the researchers?
   R/ Increase of resources due to SNAP during gestation and childhood, health and economic status of individual in adulthood.

c. If the sample measurements are used to make inferences about the population characteristics, why is a measure of reliability of the inferences important?
   R/ Because in order to make an informed decision the evidence in favor of or against the policy must be reliable, or a least uncertainty must be quantified and weighted into the decision.

d. Identify the population that is of interest to the researchers.
   R/ The set of individuals receiving SNAP benefits.

Problem 2

1.5

During the 2004 senatorial campaign in a large southwestern state, illegal immigration was a major issue. One of the candidates argued that illegal immigrants made use of educational and social services without having to pay property taxes. The other candidate pointed out that the cost of new homes in their state was 20–30% less than the national average due to the low wages received by the large number of illegal immigrants working on new home construction. A random sample of 5,500 registered voters was asked the question, "Are illegal immigrants generally a benefit or a liability to the state's economy?" The results were as follows: 3,500 people responded "liability," 1,500 people responded "benefit," and 500 people responded "uncertain."

a. What is the population of interest?
   R/ Registered voters. From the opening sentence one could assume that the population of interest is specifically the registered voters in the unnamed southwestern state. However there is no explicit specification of the geographical scale corresponding to the 5500 voters sample, it could have been drawn at the national level also.

Problem 2 continued on next page. . .

b. What is the population from which the sample was selected?
R/ From the text it is not clear whether the 5500 subjects were sampled just in the unnamed southwestern state or throughout the US.

c. Does the sample adequately represent the population?
R/ Assuming a normal approximation, the confidence interval for a proportion is:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

In our case, $\hat{p}$ is the estimator for a proportion of voters in the population, z is the percentile for the standard normal distribution corresponding to an α confidence level, and n is the sample size.

With a 95% confidence, $\alpha = 0.05$ and $z = 1.96$. For an expected $\hat{p} = 0.5$ and $n = 5500$, the margin of error would be:

$$MOE = 1.96\frac{0.5}{\sqrt{5500}}$$

$$MOE \approx 0.0132$$

A litle above a percentage point.

This means that $n = 5500$ is enough to estimate proportions of a population with a 95% confidence within ±1% margin.

If the sample was randomly selected, yes it will be representative of its corresponding population, whether it is drawn from the unnamed southwestern state or the whole US ($N \approx 329$ millions).

d. If a second random sample of 5,000 registered voters was selected, would the results be nearly the same as the results obtained from the initial sample of 5,000 voters? Explain your answer.
R/

$$MOE = 1.96\frac{0.5}{\sqrt{5000}}$$

$$MOE \approx 0.0139$$

If both samples are drawn randomly from the same population we do no expect them to differ much, because given their size both are representative and big enough to calculate $\hat{p}$ within a percentage point.

Problem 3

2.3

A news report states that minority children who take advanced mathematics courses in high school have a first-year GPA in college that is equivalent to that of white students. The newspaper columnist suggested that the lack of advanced mathematics courses in high school curricula in inner-city schools was a major cause of the low college success rate of students from inner-city schools. What confounding variables may be present that invalidate the columnist's conclusion?
R/

Problem 4

2.7

The circuit judges in a rural county are considering a change in how jury pools are selected for felony trials. They ask the administrator of the courts to assess the county residents' reaction to changing the requirement

Problem 4 continued on next page. . .

for membership in the jury pool from the current requirement of all registered voters to a new requirement of all registered voters plus all residents with a current driver's license. The administrator sends questionnaires to a random sample of 1,000 people from the list of registered voters in the county and receives responses from 253 people.

a. What is the population of interest?
   R/

b. What is the sampling frame?
   R/

c. What possible biases could be present in using the information from the survey?
   R/

Problem 5

2.10

The New York City school district is planning a survey of 1,000 of its 250,000 parents or guardians who have students currently enrolled. They want to assess the parents' opinion about mandatory drug testing of all students participating in any extracurricular activities, not just sports. An alphabetical listing of all parents or guardians is available for selecting the sample. In each of the following descriptions of the method of selecting the 1,000 participants in the survey, identify the type of sampling method used (simple random sampling, stratified sampling, or clus- ter sampling).

a. Each name is randomly assigned a number. The names with numbers 1 through 1,000 are selected for the survey.
   R/

b. The schools are divided into five groups according to grade level taught at the school: K–2, 3–5, 6–7, 8–9, 10–12. Five separate sampling frames are constructed, one for each group. A simple random sample of 200 parents or guardians is se- lected from each group.
   R/

c. The school district is also concerned that the parent's or guardian's opinion may differ depending on the age and sex of the student. Each name is randomly as- signed a number. The names with numbers 1 through 1,000 are selected for the survey. The parent is asked to fill out a separate survey for each of their currently enrolled children.
   R/

Problem 6

2.17

A medical study is designed to evaluate a new drug, D1, for treating a particular illness. There is a widely used treatment, D2, for this disease to which the new drug will be compared. A placebo will also be included in the study. The researcher has selected 10 hospitals for the study. She does a thorough evaluation of the hospitals and concludes that there may be aspects of the hospitals that may result in the elevation of responses at some of the hospitals. Each hospital has six wards of patients. She will randomly select six

Problem 6 continued on next page. . .

patients in each ward to participate in the study. Within each hospital, two wards are randomly assigned to administer D1, two wards to administer D2, and two wards administer the placebo. All six patients in each of the wards will be given the same treatment. Age, BMI, blood pressure, and a measure of degree of illness are recorded for each patient upon entry into the hospital. The response is an assessment of the degree of illness after 6 days of treatment.
In place of the design described above make the following change. Within each hospital, the three treatments will be randomly assigned to the patients, with two patients in each ward receiving D1, two patients receiving D2, and two patients receiving the placebo.
Identify:

a. Type of design.
R/

b. Explanatory and response variables.
R/

c. Levels of treatment.
R/

d. Which key principles of experimental design were used.
R/

Problem 7

Give an appropriate positive constant c such that $f(n) \leq c \cdot g(n)$ for all $n > 1$.

1. $f(n) = n^2 + n + 1$, $g(n) = 2n^3$
2. $f(n) = n\sqrt{n} + n^2$, $g(n) = n^2$
3. $f(n) = n^2 - n + 1$, $g(n) = n^2/2$

Solution
We solve each solution algebraically to determine a possible constant c.

Part One

$$\begin{aligned} n^2 + n + 1 &= \\ &\leq n^2 + n^2 + n^2 \\ &= 3n^2 \\ &\leq c \cdot 2n^3 \end{aligned}$$

Thus a valid c could be when $c = 2$.

Part Two

$$\begin{aligned} n^2 + n\sqrt{n} &= \\ &= n^2 + n^{3/2} \\ &\leq n^2 + n^{4/2} \\ &= n^2 + n^2 \\ &= 2n^2 \\ &\leq c \cdot n^2 \end{aligned}$$

Problem 7 continued on next page. . .

Thus a valid c is $c = 2$.

Part Three

$$\begin{aligned} n^2 - n + 1 &= \\ &\leq n^2 \\ &\leq c \cdot n^2/2 \end{aligned}$$

Thus a valid c is $c = 2$.

Problem 8

Let $\Sigma = \{0, 1\}$. Construct a DFA A that recognizes the language that consists of all binary numbers that can be divided by 5.

Let the state q_k indicate the remainder of k divided by 5. For example, the remainder of 2 would correlate to state q_2 because $7 \mod 5 = 2$.

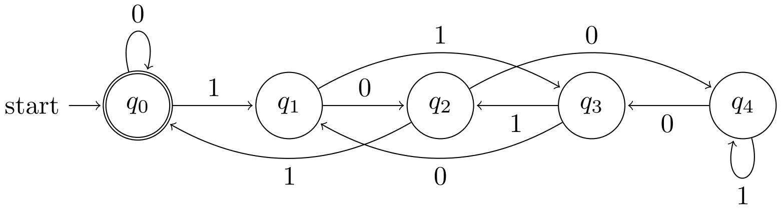

Figure 1: DFA, A, this is really beautiful, ya know?

Justification

Take a given binary number, x. Since there are only two inputs to our state machine, x can either become $x0$ or $x1$. When a 0 comes into the state machine, it is the same as taking the binary number and multiplying it by two. When a 1 comes into the machine, it is the same as multipying by two and adding one.

Using this knowledge, we can construct a transition table that tell us where to go:

	$x \mod 5 = 0$	$x \mod 5 = 1$	$x \mod 5 = 2$	$x \mod 5 = 3$	$x \mod 5 = 4$
$x0$	0	2	4	1	3
$x1$	1	3	0	2	4

Therefore on state q_0 or ($x \mod 5 = 0$), a transition line should go to state q_0 for the input 0 and a line should go to state q_1 for input 1. Continuing this gives us the Figure 1.

Problem 9

Write part of **Quick-Sort(**$list, start, end$**)**

```
function QUICK-SORT($list, start, end$)
    if $start \geq end$ then
        return
    end if
    $mid \leftarrow$ PARTITION($list, start, end$)
    QUICK-SORT($list, start, mid - 1$)
    QUICK-SORT($list, mid + 1, end$)
end function
```

Algorithm 1: Start of QuickSort

Problem 10

Suppose we would like to fit a straight line through the origin, i.e., $Y_i = \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathrm{E}[e_i] = 0$, and $\mathrm{Var}[e_i] = \sigma_e^2$ and $\mathrm{Cov}[e_i, e_j] = 0, \forall i \neq j$.

Part A

Find the least squares esimator for $\hat{\beta}_1$ for the slope β_1.

Solution

To find the least squares estimator, we should minimize our Residual Sum of Squares, RSS:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 x_i\right)^2 \end{aligned}$$

By taking the partial derivative in respect to $\hat{\beta}_1$, we get:

$$\frac{\partial}{\partial \hat{\beta}_1}(RSS) = -2\sum_{i=1}^{n} x_i(Y_i - \hat{\beta}_1 x_i) = 0$$

This gives us:

$$\begin{aligned} \sum_{i=1}^{n} x_i(Y_i - \hat{\beta}_1 x_i) &= \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 \\ &= \sum_{i=1}^{n} x_i Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Solving for $\hat{\beta}_1$ gives the final estimator for β_1:

$$\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}$$

Problem 10 continued on next page...

Part B

Calculate the bias and the variance for the estimated slope $\hat{\beta_1}$.

Solution

For the bias, we need to calculate the expected value $\mathrm{E}[\hat{\beta_1}]$:

$$\begin{aligned}
\mathrm{E}[\hat{\beta_1}] &= \mathrm{E}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i \mathrm{E}[Y_i]}{\sum x_i^2} \\
&= \frac{\sum x_i (\beta_1 x_i)}{\sum x_i^2} \\
&= \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1 \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1
\end{aligned}$$

Thus since our estimator's expected value is β_1, we can conclude that the bias of our estimator is 0.

For the variance:

$$\begin{aligned}
\mathrm{Var}[\hat{\beta_1}] &= \mathrm{Var}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \sigma^2 \\
&= \frac{\sigma^2}{\sum x_i^2}
\end{aligned}$$

Problem 11

Prove a polynomial of degree k, $a_kn^k + a_{k-1}n^{k-1} + \ldots + a_1n^1 + a_0n^0$ is a member of $\Theta(n^k)$ where $a_k \ldots a_0$ are nonnegative constants.

Proof. To prove that $a_kn^k + a_{k-1}n^{k-1} + \ldots + a_1n^1 + a_0n^0$, we must show the following:

$$\exists c_1 \exists c_2 \forall n \geq n_0,\ c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n)$$

For the first inequality, it is easy to see that it holds because no matter what the constants are, $n^k \leq a_kn^k + a_{k-1}n^{k-1} + \ldots + a_1n^1 + a_0n^0$ even if $c_1 = 1$ and $n_0 = 1$. This is because $n^k \leq c_1 \cdot a_kn^k$ for any nonnegative constant, c_1 and a_k.

Taking the second inequality, we prove it in the following way. By summation, $\sum\limits_{i=0}^{k} a_i$ will give us a new constant, A. By taking this value of A, we can then do the following:

$$\begin{aligned} a_kn^k + a_{k-1}n^{k-1} + \ldots + a_1n^1 + a_0n^0 = & \\ & \leq (a_k + a_{k-1} \ldots a_1 + a_0) \cdot n^k \\ & = A \cdot n^k \\ & \leq c_2 \cdot n^k \end{aligned}$$

where $n_0 = 1$ and $c_2 = A$. c_2 is just a constant. Thus the proof is complete. □

Problem 18

Evaluate $\sum_{k=1}^{5} k^2$ and $\sum_{k=1}^{5}(k-1)^2$.

Problem 19

Find the derivative of $f(x) = x^4 + 3x^2 - 2$

Problem 6

Evaluate the integrals $\int_0^1 (1-x^2)\mathrm{d}x$ and $\int_1^\infty \frac{1}{x^2}\mathrm{d}x$.