1. About 22% of the residents of California were born outside the United States. You choose an SRS of 1,000 California residents for a sample survey on immigration issues. You want to find the probability that at least 250 of the people in your sample were born outside the United States.

[4] (a) You would like to use the normal approximation for the sampling distribution of a sample proportion to answer this question. Is it appropriate to do so in this case? Why or why not?

In this case n=1000, and \( p=0.22, \ q=0.78 \)

We have \( np=1000*0.22 = 220, \) and \( nq = 780. \) As both \( np \) and \( nq \) are greater than 10, we can use the normal approximation of the binomial.

[6] (b) Using the correct approach, determine the probability that at least 250 of the people in your sample were born outside the United States.

\[
\text{Standard deviation} = \sqrt{n* p*(1-p)} = \sqrt{1000*0.22*(1-0.22)} = 13.1
\]

Now \( E(x) = 1000*0.22=220.\)

Using continuity correction, the z-score for 250, \( z_{250} = (249.5-220)/13.1=2.25. \)

Using the normal table, we find the p-value for \( z_{250} = 2.25 \) is 0.4878. Therefore, the probability that at least 250 of the people in the sample will be born outside the US is 0.5 – 0.4878 = 0.0122. This means that there is a 1.22% chance that the sample will have 250 or more people born outside the US.

2. The following is adapted from “Improving Classroom Practice: Implementing Action Research,” by Rock D. Moore of Fort Hays State University. The author evaluates the effectiveness of a six-month program designed to improve the English oral language skills of Limited English Speakers (LEPs).
We hypothesized that English oral language proficiency of LEP students would increase significantly after the six-month after-school intervention. The pre-test was administered to a sample of 30 LEP students during the first week of school in June. The mean number of correct items for the group was 23.4. The post-test was administered to the same 30 LEP students during the first week in December (after completion of the program). The mean number of correct items for the group was 55.1.

(a) Let $x =$ the mean number of correct answers on the English oral language skill post-test. Write the formulas for the null and alternative hypotheses of the author.

The null hypothesis is $H_0 : x \leq 23.4$

$H_0$: The mean number of correct answers on the English post-test was no greater than (i.e., equal to or less than) the mean number of correct items in the pretest.

$H_A : x > 23.4$

$H_A$: The mean number of correct answers on the English post-test was greater than the mean number of correct items in the pretest.

(b) Suppose the author computed the test statistic for this hypothesis and obtained a p-value of .045. Interpret this result in one or two clear sentences in a way that someone with no statistical training could understand it.

This means that the probability of mean post-test score being 55.1 or greater by chance alone (assuming the after-school intervention program had no effect on student performance) is 0.045. This difference in pre- and post-test mean scores is thus something that we would predict to happen by chance only one out of twenty times—which is “statistically” significant by most rules of thumb.

(c) Is this result “significant at the .01 level?” Why or why not? Draw a picture if this helps you answer the question.

No, this result is not significant at the .01 level. It is only significant at the .045, or to round off, at the .05 level. Significance at the .01 level would mean that there is a 1% chance of the mean score being at least 55.1, if the program had no effect.
3. Among all water scarce countries (defined as countries with a per capita water consumption of less than 100 liters per day), average daily per capita fresh water consumption is distributed normally with a mean of 33.3 liters and a standard deviation of 24.9 liters. (From P. Gleick, 2000, “The human right to water,” in Water Policy 1(15))

(a) What is the probability that a randomly selected water scarce country will have an average daily per capita water consumption value of more than 80 liters per day?

\[ Z_{80} = \frac{80 - 33.3}{24.9} = 1.86. \]  From the normal tables, the P-value is 0.4686. Therefore, the probability that a randomly selected country will have an average daily consumption per capita of more than 80 liters per day would be 0.5 - 0.4686 = 0.0314.

(b) The country of Cambodia has an average per capita water consumption value of 6 liters. How does Cambodia’s water consumption compare with the population of water scarce countries? (Note that we discussed many different ways to express this comparison in class—you may choose any approach that is appropriate for the question.)

\[ z_6 = \frac{6 - 33.3}{24.9} = -1.10. \]  The corresponding p-value from the normal table is 0.3643. Therefore (0.5 - 0.3643)*100 = 13.57% of countries have per capita water consumption levels lower than that of Cambodia. Alternatively, 86.43% of countries have water consumption levels greater than that of Cambodia.

(c) Suppose the United Nations wanted to provide assistance to those countries whose water situation is the most dire. With limited funding, however, the UN can only assist a subset of “water scarce” nations—in particular, the 25% of countries with the lowest average per capita water consumption. Determine the cutoff water consumption value that would allow the UN to identify the countries suffering the greatest water scarcity problems.

We want to know the z-score associated with the 25th percentile, or with a p-value of 0.25. From the normal table, we find the z-score to be -0.675. (Note that the value is negative because we are talking about countries that have below average water consumption.) So the cutoff water consumption value would be:

\[ z\text{-score}\times\text{std. Deviation} + \text{mean} = -0.675 \times 24.9 + 33.3 = 16.49 \text{ liters.} \]

(d) Is it correct to say that you have identified the 25th percentile value for water scarce countries in part c? If not, why not?

Yes. It is true that 25% of the countries have a water consumption value lower than the value we obtained in part (c).
4. The following table was produced from the University of Michigan’s General Social Survey and contains data from a SRS of Americans during 1986.

<table>
<thead>
<tr>
<th>Number of children</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>399</td>
</tr>
<tr>
<td>1</td>
<td>214</td>
</tr>
<tr>
<td>2</td>
<td>357</td>
</tr>
<tr>
<td>3</td>
<td>218</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>8 or more</td>
<td>28</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>1468</strong></td>
</tr>
</tbody>
</table>

(a) Suppose you wanted to draw a sub-sample of 50 households from the 1468 households included in the GSS survey in order to conduct more in-depth interviews with them. Moreover, suppose you wanted to stratify this sub-sample on number of children in the household. Explain in one or two simple sentences what “stratifying on number of children” means and what you want to accomplish by doing it. Then show how you would determine the number of households that should be drawn from each stratum to form your sub-sample of 50 households.

The purpose of “stratifying on number of children” is to ensure that we obtain a sub-sample of 50 households that is representative of the full sample (1468 households) with respect to one characteristic of interest—the number of children per household. (Note that we are deliberately choosing not to follow a simple random sampling procedure in order to be certain that our sample is representative in this regard.) The procedure is as follows:

1. Classify the data into strata (this has already been done for you).
2. Determine the size of the sample you want (in this case, 50)
3. Determine the proportion of the sample to be obtained in each stratum. The computations are as follows (note that you do have to round off, as you are dealing with households which must be sampled in whole numbers):

<table>
<thead>
<tr>
<th>Number of children</th>
<th>Frequency</th>
<th>Proportion in stratum</th>
<th>Proportion X 50</th>
<th>Stratum sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>399</td>
<td>399/1468=0.272</td>
<td>13.6</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>214</td>
<td>0.146</td>
<td>7.3</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>357</td>
<td>0.243</td>
<td>12.2</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>218</td>
<td>0.149</td>
<td>7.4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
<td>0.087</td>
<td>4.4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0.042</td>
<td>2.1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>0.033</td>
<td>1.7</td>
<td>2</td>
</tr>
</tbody>
</table>
(d) Draw a simple random sample in each strata to obtain the number of observations required. For example, we would take an SRS of 14 households with no children (the first stratum).
(e) Pool the observations to create a final sample.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>14</td>
<td>0.01</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>8 or more</td>
<td>28</td>
<td>0.019</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1468</td>
<td>1</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

(b) Compute the mean number of children for the original sample (n=1468). Use the value 8 for households reporting “8 or more” children. Show your work.

The mean number of children for the sample households is

\[
(0\cdot 399 + 1\cdot 214 + 2\cdot 357 + 3\cdot 218 + 4\cdot 128 + 5\cdot 61 + 6\cdot 49 + 7\cdot 14 + 8\cdot 28)/1468 = 2.05
\]

(c) Devise a test of the hypothesis that the mean number of children is 2.6. (The appropriate standard deviation value to use is 1.89.) Show your work: write the equation for your hypothesis and the computations used in obtaining your test statistic. Interpret the results you obtain in one or two sentences. State clearly whether you reject your hypothesis or not and why.

\[
H_0 : \text{Mean is equal to } 2.6 \\
H_A : \text{Mean is not equal to } 2.6 \\
\]

We use the t-statistic. This is the same as using the z-score for a normal distribution as n is greater than 30.

The t-statistic

\[
t = \frac{\bar{X} - \mu_0}{s.e.} = \frac{2.05 - 2.6}{0.049} = -11.21
\]

The associated p-value for the two-tailed test (which this is), is vanishingly small. We can reject the null hypothesis and conclude that the mean number of children is significantly different from 2.6. Note that, when devising tests of your own, if you have an hypothesis about the direction of a difference, make sure you construct an appropriate one-tailed test.

If you used the confidence interval approach:
\begin{align*}
&\bar{x} - t\left(\frac{\alpha}{2}\right) \frac{s}{\sqrt{n}} < \mu < \bar{x} + t\left(\frac{\alpha}{2}\right) \frac{s}{\sqrt{n}} \\
&\bar{x} \pm t\left(\frac{\alpha}{2}\right) \frac{s}{\sqrt{n}} \\
&2.05 \pm 1.96 \times \frac{1.89}{\sqrt{1468}} \\
&2.05 \pm 0.097 \\
&1.95, 2.15
\end{align*}

[4] (d) Do you think the mean number of children is an appropriate measure of central
tendency to use for these data? Why or why not? (No need to do more computations
here; just a couple of sentences will do.)

\textit{We would probably not use mean number of children as the preferred measure of
central tendency for these data. Just looking at the frequency distribution table, we
can see that the data are skewed, with a large number of observations in the 0, 1, and
2 categories.}

5. The following graphic comes from \textit{USA Today}:

[2] (a) With reference to Tufte’s good graphic presentation principles, do you have any
critiques of this graphic?

\textit{This graph certainly violates the principle “The representation of numbers, as
physically measured on the surface of the graph, should be directly proportion to the
numerical quantities represented.” The arrow depicting 84.2 should be 2.7 times
longer than the one depicting 30.8. Perhaps you found other issues to discuss.}

[6] (b) Suppose the data used to develop this graphic were obtained from a simple random
sample of 1,180 Americans. Construct a 95\% confidence interval for the proportion of
Americans who commute to work alone in their cars. Explain in one clear sentence
how to interpret this confidence interval to someone with no training in statistics.
From the sample, the proportion of Americans commuting to work alone in their cars is 
\[ \hat{p} = \frac{84.2}{84.2 + 30.8} = 0.732 \]
\[ n = 1180 \]

Now we form the confidence interval: 
\[ \hat{p} - E < p < \hat{p} + E \]
\[ 0.732 - E < p < 0.732 + E \]

Now \[ E = z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \]

For 95% confidence interval, \[ \frac{\alpha}{2} = 0.025 \]
\[ E = z(0.025) \sqrt{\frac{0.732(1-0.732)}{1180}} = 1.96 \times 0.01289 = 0.025 \]

Therefore the 95% confidence interval for the proportion of commuting Americans who go to work alone in their cars is:
\[ 0.732 - 0.025 < p < 0.732 + 0.025, \text{ or} \]
\[ 0.707 < p < 0.757 \]

We can interpret this interval by saying that we are 95% confident that between 70.7% and 75.7% of Americans who commute to work do so alone in their cars.

(c) Suppose the data used to develop this graphic were obtained from the US census (i.e., from all Americans). How would your construction of a 95% confidence interval for the proportion of Americans who commute to work alone in their cars change and why?

If this data were obtained from a census, then we would be able to compute a value for the population parameter, \( p \). We would have no need to use \( \hat{p} \) to estimate \( p \), and would thus have no need for a confidence interval. We would simply report the value of \( p \) for the entire American population, and would do so at a 100% confidence level.

7. In the attached Goolsbee article “In a World Without Borders: The Impact of Taxes on Internet Commerce”, read section III and answer the following questions.

(a) What type of research design does Goolsbee use to carry out his study? Please write a couple of sentences in support of your answer?

This is a quasi-experimental research design. It is a cross sectional study that collects data after the experimental group has been exposed to a treatment, as shown below.

\[ X \quad O_e \]
X is the treatment, sales tax in this case, and \( O_e \) is the outcome after the treatment, internet purchasing behavior in this case. The study also tries to incorporate some elements of experimental design. It is not a managed or a laboratory experiment, but in some sense it is a natural experiment, where there are different rates of sales taxes and different rates of online purchases in different jurisdictions. The study tries to isolate the effect of sales tax by isolating its effect from other factors.

(b) Do you think the research design is internally valid? Why or why not? What are the different controls that Goolsbee uses and why?

We look at internal validity using the following four criteria.

1) Time order: In this case it is clear that sales tax differences have not come about after the internet came into being. So it is more or less clear that the direction of causation runs from taxes to online purchasing. One could, of course, say that online purchasing may have caused new differences in sales taxes, but this is doubtful, and there is no convincing theoretical explanation for why this should be so.

2) Covariation: This is also present as the study shows that higher sales taxes are coincident with higher online purchasing behavior.

3) Nonspuriousness: This is more problematic. The study does control for various other explanatory factors such as age, income, education, race, some city-level and individual-level differences. The city-level controls that he uses are procedures to assign tax rates, internet access and infrastructure and cost of living or housing prices. The individual level controls he uses are automobile ownership, place of residence, and technological sophistication. But one cannot say that one has exhausted all possible explanatory variables. This is one of the main problems with quasi-experimental designs. Nevertheless, the study does control for many factors that could possibly affect online buying behavior. In that sense, it meets this criteria of nonspuriousness relatively well.

4) Theory: The study also has a reasonable theory that higher taxes will lead to online buying to avoid these taxes.

In this sense, the study is reasonably internally valid.

(c) Do you think the research design is externally valid? Why or why not?

The study is not conducted in a laboratory setting, but conducted in the real world with all its effects. In this sense, the study has high external validity. Also, the study itself does not influence the observations, as the people who buy online, or the taxes are not affected by the study (at least not before it was published). One could extrapolate the findings from the sample used in the study to the rest of the country with reasonable confidence. But if one wanted to extrapolate to other parts of the world, say Europe, then there could be problems of external validity, as there could be other contextual factors such as culture that could influence buying behavior in these places.