Name: 

Notes:

• You may unstaple the midterm while completing it and restaple it upon submission as long as you preserve the page order.
After voting for him in large numbers in 2008 and 2012, young Americans are souring on President Obama.

According to a new Harvard University Institute of Politics poll, just 41 percent of millennials — adults ages 18-29 — approve of Obama's job performance, his lowest-ever standing among the group and an 11-point drop from April.

Obama's signature health care law is also unpopular among millennials. Fifty-seven percent of those surveyed said they disapprove of Obamacare, compared with 38 percent who said they approve.
A majority of respondents also said they disapprove of the way Obama is handling the economy, Syria, Iran and the budget deficit.

The results reflect a similar downward trend among the public at large. Recent polls ranging from Gallup to CNN show Obama's approval rating hovering around 40 percent, while disapproval of the health care law is in the mid-to-high 50s.

"Millennials are starting to look a lot more like their older brothers and sisters, parents and grandparents," IOP polling director John Della Volpe said in a conference call with reporters Wednesday.

The online survey of 2,089 adults was conducted from Oct. 30 to Nov. 11, just weeks after the federal government shutdown ended and the problems surrounding the implementation of the Affordable Care Act began to take center stage. The poll's margin of error was plus or minus 2.1 percentage points.

Fifty-five percent of the survey's respondents said they voted for Obama in the last presidential election, while 33 percent said they voted for Republican Mitt Romney. If the election were held again, Obama would still come out on top, but by a tighter 46 to 35 percent vote; 13 percent said they would vote for someone else.

According to the Pew Research Center, 66 percent of 18- to 29-year-olds voted for Obama in 2008, and 60 percent voted for his re-election in 2012.

Harvard's poll found millennials, like the rest of the public, aren’t happy with Congress either. Just 19 percent of respondents said they approve of congressional Republicans, while 35 percent approve of their Democratic counterparts. Both figures are single-digit drops from April. Forty-five percent also said they would "recall and replace" their member of Congress if they had the option.
1 Polling

(Preface: The use of this example should not be interpreted as an endorsement of President Obama). The previous two pages are an NPR article reporting the results of a Harvard University Institute of Politics poll conducted in between October 30 and November 11, 2013 of millennials' (adults aged 18-29) approval of President Obama’s job performance. Please start writing all your responses where indicated below.

a) Who is the study population?

b) In this study, what does the value of true population proportion \( p \) correspond to?

c) Say we had access to infinite resources, what would be the best way to measure \( p \)?

d) What condition must be met for the sample of \( n \) young Americans to be representative of the population?

e) In this example, what is the sample proportion \( \hat{p} \), the point estimate of the population proportion \( p \)? Give the numerical value.

f) Recall that the standard error of \( \hat{p} \) based on samples of size \( n \) is

\[
\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}
\]

and that when we don’t know \( p \) as is typically the case, we use the plug-in estimate \( \hat{p} \) of \( p \), resulting in

\[
\text{SE}_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Compute the standard error for this poll.

g) What is the margin of error for this poll? Show all your work.

h) Construct a 95% confidence interval for the proportion of all young Americans who approved of Obama’s job performance. Assume all conditions for inference are met.

Please start writing all your responses here:
2 Sampling

Consider the following hypothetical extension of the previous question on the poll of Obama’s approval ratings among young Americans in 2013. First, say that you are an all knowing “higher power” and you know that \( p = 0.425 \). Furthermore, the following 1000 polling companies conduct polls on the same dates as the Harvard poll, using the same sampling methodology and the same sample size \( n \). You observe the following results:

- Poll 1: Gallup finds that out of their sample of size \( n \), 844 young Americans approve of Obama.
- Poll 2: Ipsos finds that out of their sample of size \( n \), 857 young Americans approve of Obama.
- ...
- Poll 1000: Monmouth University finds that out of their sample of size \( n \), 871 young Americans approve of Obama.

Based on these 1000 polls based on samples of size \( n \), you compute 1000 values of \( \hat{p} \). You plot these in a histogram:

![Histogram of \( \hat{p} \) values]

Hint: Think back to:
Please start writing all your responses where indicated below.

a) The above normal-shaped distribution is called the **X** distribution of the sample proportion \( \hat{p} \). It describes how different values of \( \hat{p} \) vary from sample to sample due to **Y**. What are **X** and **Y**?

b) What is the numerical value of the red line on the x-axis i.e. the center of this distribution?

c) The standard deviation of the above distribution of 1000 values of \( \hat{p} \) is called what?

d) Between what two values will 95% of sample proportions based on samples of size \( n \) occur? Show your work.

e) Mark these two values on the plot on the previous page.

f) What would happen to this histogram if the sample size were \( n = 5000 \)?

g) What does this change in sample size to \( n = 5000 \) mean in terms of the quality of the polling results?

h) Bringing things back to real life and focusing on what we would do in practice, we would **not** take 1000 different samples of size \( n = 2089 \) and compute 1000 different values of \( \hat{p} \), but instead take only a single sample of \( n = 2089 \) and compute a single value of \( \hat{p} \). What was the point of the above exercise then? Respond in two sentences or less.

i) Say instead of estimating population proportions \( p \) with sample proportions, we want to estimate population means \( \mu \) with sample means \( \bar{x} \). The Central Limit Theorem guarantees that as we consider samples of size \( n \) that get larger and larger, the distribution of different values of \( \bar{x} \) based on different samples of size \( n \) behaves more and more **X** and gets **Y**. What are **X** and **Y**?

j) Which is the correct form of the standard error of \( \bar{x} \): A) \( \sqrt{s^2} \) or B) \( \sqrt{\frac{s^2}{n}} \)? How do you know this?

Please start writing all your responses here:
3 Hypothesis Testing

SDM Exercises 19.2 and 19.4: A friend of yours claims to be psychic. You are skeptical. To test this you take a stack of \( n = 100 \) playing cards and have your friend try to identify the suit, either hearts/diamonds/clubs/spades, without looking. They get 45 out of 100 right. **Please start writing all your responses where indicated below.**

a) Define a hypothesis test in non-statistical terms where \( H_0: \) Your friend is not psychic.

b) Define the corresponding hypothesis test in statistical terms. Hint: it should involve \( p \).

c) What is the standard error of \( \hat{p} \) used for hypothesis testing in general?

d) What is the standard error of \( \hat{p} \) used for this particular hypothesis test?

e) Draw the distribution of how different values of \( \hat{p} \) based on different samples of size \( n = 100 \) will behave from sample-to-sample. In particular focus on 1) its shape, 2) its center, and 3) where the middle 95% of values of \( \hat{p} \) will lie. Be as precise as possible, in other words do not use any “rules of thumb.”

f) In the plot above, mark with a dashed vertical line where the observed value of \( \hat{p} \) lies.

g) Even though you don’t have access to a computer, make a guess about the conclusion of the hypothesis test. State your conclusions both statistically and in terms of a statement on your friend being psychic.

h) **BONUS** Say you did have access to a computer right now. In order to compute the p-value above exactly, what values of \( A, B, \) and \( C \) must you input?

```r
library(mosaic)
xpnorm(A, mean = B, sd = C)
```

**Please start writing all your responses here:**
4 Inference for Regression

Recall our professor evaluations dataset based on the study from the University of Texas in Austin. In particular, we were interested in explaining a professor’s teaching evaluation score using their gender and age as explanatory variables. Here is a random sample of 5 rows out of the \( n = 463 \) professors in dataset:

<table>
<thead>
<tr>
<th>#</th>
<th>score</th>
<th>gender</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>3.9</td>
<td>female</td>
<td>40</td>
</tr>
<tr>
<td>207</td>
<td>4.7</td>
<td>male</td>
<td>62</td>
</tr>
<tr>
<td>432</td>
<td>3.3</td>
<td>male</td>
<td>62</td>
</tr>
<tr>
<td>297</td>
<td>4.1</td>
<td>female</td>
<td>42</td>
</tr>
<tr>
<td>445</td>
<td>4.9</td>
<td>female</td>
<td>52</td>
</tr>
</tbody>
</table>

Recall we fit the following regression model with an interaction term:

\[
\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 \\
\hat{\text{score}} = b_0 + b_{\text{age}} \text{age} + b_{\text{male}} \mathbb{1}[\text{is male}] + b_{\text{age, male}} \text{age} \mathbb{1}[\text{is male}]
\]

Recall the visual representation of the our model. Hint: look at this closely.

![Jittered) Scatterplot of Teaching Evaluations](image)

Finally, recall the results of the regression with confidence intervals

```r
library(broom)
lm(score ~ age * gender, data=evals) %>%
tidy(conf.int=TRUE)
```

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
<th>conf.low</th>
<th>conf.high</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.883</td>
<td>0.205</td>
<td>23.8</td>
<td>0.000</td>
<td>4.480</td>
<td>5.286</td>
</tr>
<tr>
<td>age</td>
<td>-0.018</td>
<td>0.004</td>
<td>-3.9</td>
<td>0.000</td>
<td>-0.026</td>
<td>-0.009</td>
</tr>
<tr>
<td>gendermale</td>
<td>-0.446</td>
<td>0.265</td>
<td>-1.7</td>
<td>0.094</td>
<td>-0.968</td>
<td>0.076</td>
</tr>
<tr>
<td>age:gendermale</td>
<td>0.014</td>
<td>0.006</td>
<td>2.4</td>
<td>0.015</td>
<td>0.003</td>
<td>0.024</td>
</tr>
</tbody>
</table>
Please start writing all your responses where indicated below.

a) The table reports a p-value of 0 in the age row. Write down the corresponding hypothesis $H_0$ vs $H_A$ in terms of the $\beta_{age}$, the true population associated effect of age on teaching score.

b) The p-value mentioned in part a) is 0. Report what this means for the hypothesis test corresponding to the two hypotheses above. Report this both in 1) statistical terms and 2) language that non-statisticians can understand.

c) Based on these results, among male professors at the University of Austin for every year increase in age, there is an associated X of on average Y units in teaching score. What are X and Y?

d) What conclusion is suggested by the 95% confidence interval for $\beta_{age:gendermale}$ of (0.003, 0.024)? Justify your conclusion.

e) Say we relaxed the gender categorical variable to allow for the following three levels: female, male, and other, and furthermore say some professors selected the new “other” option. Describe precisely how the above plot would change.

f) BONUS 1 Describe precisely how the shape of the above regression table would change.

g) BONUS 2 The 95% confidence interval for $\beta_{gendermale}$ is (-0.968, 0.076). Based on values in the table, write down your best guess of the formula that R uses to compute the left end point of -0.968. Your formula and the reported left endpoint of -0.968 should match up to 2 decimal places.

Please start writing all your responses here: