1

Polling 101

Review

1. Why is the Agresti-Coull estimator not often used in polling analysis?
   Solution: *This question gets at the inherent tension between the value of a biased estimator and that of a precise estimator. The Agresti-Coull estimator is biased, but has greater precision than the usual sample proportion estimator. Thus, to a statistician, it is the better estimator.*

   However, it is rather difficult to make the case to use a biased estimator when an unbiased one is easily calculated. Furthermore, the sample proportion is easily understood, while the Agresti-Coull is not. Finally, unless the sample size is incredibly small (less than 10), the improvement on precision is rather minor. Thus, it is an improvement, but not much of one.

   For these reasons, the Agresti-Coull estimator is rarely used.

2. How would you convince a person that an estimator with a lower MSE is preferred to one that is unbiased?
   Solution: *Great question. I have yet to find a way of doing this successfully. The closest I've come is to emphasize the precision and the likelihood of being closer to the right answer (with the lower MSE estimator) rather than being right "on average." The "on average" requires performing the experiment many times in order to reap the benefit.*

3. What are the main differences between a confidence interval and a credible interval? When should you use each?
   Solution: *The main difference is that a credible interval is based on probabilities. That is, one can state "the probability of π being in the interval is 95%." For a confidence interval, such a statement is not true. All that can be said is "95% of the time, π is in the*

interval." We don't know if this is one of those 95% or one of the 5%.

I advocate using credible intervals at all times, because it provides more information. However, some hold that Bayesian analysis is inherently weak because it bases itself on an assumption about the distribution of π.

4. List several advantages and disadvantages to Bayesian analysis.
   Solution: *Bayesian analysis provides a way of incorporating prior information (or lack thereof) in the analysis. With that prior information, and the additional information from the collected data, one is able to calculate the distribution of the population parameter (or of its uncertainty).*

   The main drawback is that the results are based on the assumption that the prior distribution is correct. If the prior is wrong, then the posterior distribution is wrong.

   However, I do not see this as a weakness. Frequentist analysis also makes an assumption about the distribution of the parameter. It just is not as explicit about it.

5. Why is the conservative margin of error approximately $1/\sqrt{n}$?
   Solution: *The conservative margin of error is based on assuming the population parameter is* $\pi = 0.500$. *Using this as the estimate of* π *produces the widest confidence intervals.*

6. What does the thick line at the bottom of Figure 1.2, page 12, represent?
   Solution: *This line represents the 95% confidence interval. A total of 95% of the binomial distribution occurs within that interval (47 to 53%).*

7. Why is the beta distribution called the "conjugate prior" for the binomial distribution?
   Solution: *The beta distribution is the conjugate prior to the binomial distribution because if we make the prior distribution a beta, then the posterior distribution will also be a beta.*

8. What is “coverage,” and how can a biased estimator have actual coverage closer to the claimed coverage than an unbiased estimator? Do all biased estimators have better coverage than unbiased estimators?
   Solution: *Coverage is the proportion of the time that the confidence interval will contain (or cover) the parameter (see Figure 1.6 for an illustration).*

   While a part of the confidence interval, the parameter estimator is not the whole story. The endpoints contain both the estimator and *the standard error (standard deviation of the sampling distribution). Thus, the unbiased estimator may have a standard error that is much greater than “it should be,” which causes the coverage to be much higher than 95%.*

   No, not all biased estimators have better coverage. However, it does open a new field of study in trying to determine correct confidence intervals, even if the estimator is biased.

Conceptual Extensions

1. There were four counties (local authorities) in which the 2014 Scottish independence referendum received more than 50% of the votes cast (Figure 1.1). What do they have in common?
   Solution: *I don’t know. Some ideas would include that they are all urban areas, that they had the same economic base, that the English avoided those areas. In reality, there may be nothing connecting these four counties beyond how they voted.*

2. The turnout for the 2014 Scottish independence referendum was not 100%. Assuming that those who did not vote in each county were similar to those who did, would a higher turnout have helped independence or not?
   Solution: *All four of the counties that voted in favor of independence had lower-than-average turnout. So, increasing their turnout would help Scottish independence. However, there would need to be an increase in over 250 thousand votes in these four counties. Now, consider that increasing the turnout in the other counties would make it harder for independence to pass. Thus, increasing turnout would not help.*

   Given that we have the actual numbers, we see that increasing the *turnout to 100% across Scotland, while keeping the same support level would result in 1,922,398 in favor and 2,361,687 against independence. This is 44.9% in favor... not much of a shift in voting results.*

```
dt = read.csv("sct2014referendum.csv")
attach(dt)

registered = Valid/Turnout

pi = Yes/Valid

sum(pi*registered)
sum((1-pi)*registered)
```

Above is an example of how to use R for this question.

3. The Survation poll estimated independence support to be 47%, which was 2.3% too high. Was this poll “too far off”? This question gets at your understanding of the confidence interval.
Solution: *Ultimately, it depends. Depending on the significance level of the confidence interval, 47% could be within the range. Recall if 47% is within the range, so between the "reasonable" and "unreasonable" values, it will be considered reasonable.*

4. In the beta distribution, one can think of $a+b$ as the effective sample size, n. With this, explain why $\mathbb{E}\left[\ Y\ \right] = \frac{a}{a+b}$ makes sense. Also, compare the variance of a beta distribution with that of the sample proportion, $\frac{\pi(1-\pi)}{n}$.
Solution: *The easiest way to think of this is recalling the sample proportion of π. Remember it was $\P = \frac{X}{n}$, where X was the parameter of interest and n was the sample size. This was an unbiased estimator for π. For a similar logic, we are looking for an estimator for μ. In the case of the Beta, a is our parameter of interest, and $a+b$ is the sample size. Therefore, $\mathbb{E}\left[\ Y\ \right] = \frac{a}{a+b}$ makes sense.*

   Comparing the variance of the beta to that of the sample proportion is a bit odd, but also quite similar. Recall the variance for the beta distribution is $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$. As usual, the denominator is the sample size, plus an extra 1 for paramaterization. The numerator will be both parameters of the beta, similar to the numerator of the sample proportion equation.

5. Using words, answer these two questions:

   (a) How do we know that the binomial distribution is equivalent to the hypergeometric distribution when $n = 1$?
   Solution: *Look back to table 1.1. The expected values are already the same, the variances are different; but, notice what happens when we substitute* 1 *in for* n*. The exponent for the hypergeometric distribution reduces to one, and the variances are the same. Therefore, the two distributions are equivalent at* $n = 1$*.*

   (b) How do we know that the binomial distribution is equivalent to the hypergeometric distribution when $N = \infty$?
   Solution: *Back again at table 1.1, think about it like so: in the numerator and denominator of the hypergeometric variance, we have a super huge number over a super huge number. Therefore, the exponential once again cancels out, and we are left with the binomial and hypergeometric distributions being equivalent.*

6. Using mathematical proofs, answer these two questions:

   (a) How do we know that the binomial distribution is equivalent to the hypergeometric distribution when $n = 1$?
   Solution:
   $n\pi(1-\pi)^{\frac{N-n}{N-1}}$
   $n\pi(1-\pi)^{\frac{N-1}{N-1}}$
   $n\pi(1-\pi)$

   (b) How do we know that the binomial distribution is equivalent to the hypergeometric distribution when $N = \infty$?
   Solution:
   $n\pi(1-\pi)^{\frac{N-n}{N-1}}$
   $n\pi(1-\pi)^{\frac{\infty-1}{\infty-1}}$
   $n\pi(1-\pi)$

Computational Extensions

1. Using Defn. 1.3, prove MSE $[\ P\ ] = \text{bias}^2[P] + \mathbb{V}[\ P\ ]$.
   Solution:

   Recall two things:

   $$\begin{aligned} biasP &= \mathbb{E}[\ P - \pi\ ] = \mathbb{E}[\ P\ ] - \pi \\ \mathbb{V}[\ P\ ] &= \mathbb{E}[\ P^2\ ] - \mathbb{E}[\ P\ ]^2 \end{aligned}$$

   With this, we have:

   $$\begin{aligned} MSE(P) &= \mathbb{E}[\ (P-\pi)^2\ ] \\ &= \mathbb{E}[\ P^2 - 2p\pi + \pi^2\ ] \\ &= \mathbb{E}[\ P^2\ ] - 2\,\mathbb{E}[\ p\pi\ ] + \mathbb{E}[\ \pi^2\ ] \\ &= \mathbb{E}[\ P^2\ ] - 2\,\mathbb{E}[\ p\ ]\,\pi + \mathbb{E}[\ \pi^2\ ] \\ &= \mathbb{E}[\ P^2\ ] - \mathbb{E}[\ P\ ]^2 + \mathbb{E}[\ P\ ]^2 - 2\,\mathbb{E}[\ p\ ]\,\pi + \mathbb{E}[\ \pi^2\ ] \\ \\ &= \mathbb{V}[\ P\ ] + \mathbb{E}[\ P\ ]^2 - 2\,\mathbb{E}[\ p\ ]\,\pi + \mathbb{E}[\ \pi^2\ ] \\ &= \mathbb{V}[\ P\ ] + (\ \mathbb{E}[\ P\ ] - \pi)^2 \\ &= \mathbb{V}[\ P\ ] + (bias(P))^2 \end{aligned}$$

2. Repeat all of the Survation calculations for this chapter for a poll in which 256 out of 500 people supported Scottish independence. Do the conceptual conclusions significantly differ?
   Solution: *Starting off with the quick example in page 7, the sample proportion for the estimate for the support would be:*

   $P = \frac{x}{n} = \frac{256}{500} = 0.512$. *Next, in the quick example on page 10, assuming the support was 50%, the calculations would be the same. Next up, with the quick example on page 13, we would get:*

   $256/500 - 1.96\sqrt{\frac{0.500(1-0.500)}{756}} = 47.64$

   to

   $256/500 + 1.96\sqrt{\frac{0.500(1-0.500)}{756}} = 54.76$

   Notice these results are different by almost 5% on each interval. This is likely due to the change in sample size.

3. In several places, we focused exclusively on the kernel of the distribution and completely ignored the normalizing constants. Why can we do that? Are distributions *uniquely* determined by their kernel?
   Solution: *When we are doing problems like this, we are focusing on a particular parameter of estimation. The normalizing constant (keyword being* constant*) does not have much to do with the parameter of estimation, so we are focused on the kernel to tell us the information we need. No two are the same. There are times that it is close, and distributions are very similar, but the kernels are unique. That is the heart of how they preform, but there are many similarities, like between the binomial and the beta, just for one example.*

2

Polling 399

Review

1. What information is needed to properly use stratified sampling?
Solution: *In order to conduct stratified sampling, we will need the point estimate,* $P = \frac{X}{n}$*, the proper weight (especially important for an unbiased estimate), and the stratified estimate we saw on page 38.*

2. How would you convince a person that an estimator with a lower MSE is preferred to one that is unbiased?
Solution: *Unbiased estimators are incredibly unrealistic in real life applications, so typically if you have a really greate estimator with a low MSE compared to an ubiased estimator that has a potential to have a large MSE or be tailored to being unbiased, you will want to take the estimator that gets you closest to reality. In other words, unbias is great, but a lot of times, biases can give insight to the data. MSE is important because we can actually see how far off our model is from reality, which is crucial regardless of bias.*

3. What is the source of bias in the stratified estimator?
Solution: *The source of bias in a stratified estimator is when* $w_g \neq \frac{N_g}{N}$*. In other words, when the weights are incorrect, then the estimator will be biased.*

4. What proportion of the US population agrees with the statement "Women should be treated the same as men in the military"?
Solution: *The population support is equal to 50%.*

5. What was it about the gender variable that made it desirable as

the grouping variable?
Solution: *Usually, the estimate will naturally bias towards the gender that has the largest weight, so grouping it to essentially smooth things out will beneficial.*

6. What were some important conclusions from the stratified sampling experiments (pages **??**–**??**)?
Solution: *In my opinion, some of the most important conclusions are what happens in the case of a biased and unbiased estimator, and what happens if our weights are very close to being perfect versus incredibly off from what they should be. This tells us how models using stratified sampling actually behave and how we can come to conclusions in the real world.*

7. What was it about the first stratified sampling experiment (page **??**) that produced results identical to the simple random sampling experiment?
Solution: *The support of the candidate was constant (π_i), so there was no advantage to using stratified over simple random. This tells us that when we need to do grouping or work with weights, we will typically prefer one to the other (usually determined by what is needed and the MSE).*

8. In the formula for the stratified confidence interval, where did the term under the square root sign come from?
Solution: *That is our standard error for stratified sampling, where the w_g^2 term is the weight for the group of interest, the $\hat{\pi_g}(1 - \hat{\pi_g})$ is the estimator for the group, and n_g is the sample size for that particular group.*

Conceptual Extensions

1. What is it about the "Democratic Support" map (Figure **??**) that suggests polling should also take state into consideration? Why is state usually not taken into consideration in national polls?
Solution: *The outcome of the states in the 2016 election was in*

favor of Clinton, and the map represents this because most large states have a high democratic support, with a reasonable amount of others in between. Therefore, the state should be taken into account because the majority was Clinton, but Trump still won.

2. In lieu of stratifying on the state, one could stratify on the region. Break the country into six good strata based on the state. Why is your stratification the best?
Solution: *There are many answers for this, but I would break it up into the West, the South, the North, the mid-West, South-East, and East. This is the best because it is pretty even across the map and targets the six main sections of the United States. There could be a lot of important relationships that lie withiin these sections, so I believe this grouping is appropriate.*

3. Why is the mean squared error stratified estimator the same as that for the simple random sample estimator in the example on page **??**?
Solution: *The MSE is the same due to the consistency in the stratified estimator. Because there is no variability in weights or in proportions, there is no change from simple random sampling, so it will be the same.*

4. Why is the margin of error for the simple random sample estimator ($1/\sqrt{n} = 0.0316$) smaller than that of the stratified estimator in the example of Section **??**, page **??**?
Solution: *It could be smaller because we are not concerned with as many variables as we are when we do stratified sampling. For example, we have the weight and the difference in estimator to be concerned with, so, in most cases, it makes for a more accurate estimator, even if it is a little higher. Another practical way to think about this is as the complexity of sampling and polling increases, there is more room for error (especially in the case of real-life exmaples). It is important to know smaller error does not always mean best, especially when we are maximizing information learned.*

Computational Extensions

1. Derive the formula for the confidence interval, Eqn. **??**.
   Solution: *We know that the formula for a confidence interval is defined as:*

   $$lowerbound = point\ estimate - Z_{\alpha/2} standard\ error$$

   $$upperbound = point\ estimate + Z_{\alpha/2} standard\ error$$

   Recall when we are working with stratified sampling, the point estimate is defined as:
   $\sum_{g=1}^{G} w_g \bar{X}_g$
   Which is the sum of the weights multiplied by the stratified sampling estimator. Our $Z_{\alpha/2}$ *will depend on what level of confidence we are testing at. The typical level of confidence is 95%, so* $Z_{\alpha/2} = 1.96$. *The standard error is just the square root of the mean-squared error, so:*
   $\sqrt{\sum_{g=1}^{G} w_g^2 \frac{\hat{\pi}(1-\hat{\pi})}{n_g}}$
   Putting all we know about stratified sampling, we get:

   $$endpoints = \sum_{g=1}^{G} w_g \bar{X}_g \pm Z_{\alpha/2} \sqrt{\sum_{g=1}^{G} w_g^2 \frac{\hat{\pi}(1-\hat{\pi})}{n_g}}$$

2. Determine if the simple random sample estimator has a smaller MSE than the stratified estimator under these conditions: Five strata with support $\pi_g = \{0.50, 0.30, 0.20, 0.90, 0.60\}$. The five groups are equally represented in the population and in your sample of 6000.
   Solution: *Below is the R code for this problem:*

```
###I will set the seed so you get the same results
    as me
set.seed(737157)
###Stratified sample
##Population information

pi = c(0.50, 0.30, 0.20, 0.90, 0.60) ##Real
    subpopulation support
pg = c(1/5, 1/5, 1/5, 1/5, 1/5) ##Real population
    weights
p = sum(pi*pg) ##Total population support

##Sample info

ng = c(6000, 6000, 6000, 6000, 6000) ##Sub-sample
    sizes
wg = c(1/5, 1/5, 1/5, 1/5, 1/5) ##Claimed weights

##Generating our polls
```

```
x1 = rbinom(1e6, size=ng[1], prob=pi[1])
x2 = rbinom(1e6, size=ng[2], prob=pi[2])
x3 = rbinom(1e6, size=ng[3], prob=pi[3])
x4 = rbinom(1e6, size=ng[4], prob=pi[4])
x5 = rbinom(1e6, size=ng[5], prob=pi[5])

##Calculating thestratified estimator
estStS = wg[1]*x1/ng[1] + wg[2]*x2/ng[2] + wg[3]*x3/
    ng[3] + wg[4]*x4/ng[4] + wg[5]*x4/ng[5]

##Finally MSE

mean( (estStS-p)^2 ) ##We get an answer of 0.0036

#####Simple random sample
p = 0.50 ; n = 6000

x = rbinom(1e6, size=n, prob=p)
estSRS = x/n
mean( (estSRS-p)^2 ) ### We get an answer of 4.166e
    -5
```

Make sure you carefully read the code and make sure you know what it does. We see that the simple random sample estimator has a smaller error than the stratified.

3. Determine if the simple random sample estimator has a smaller MSE than the stratified estimator under these conditions: Six strata with support $\pi_g = \{0.50, 0.30, 0.20, 0.90, 0.40, 0.70\}$. The six groups are equally represented in the population and in your sample of 6000.
Solution: *Below is the code I used for this problem.*

```
###I will set the seed so you get the same results
    as me
set.seed(737157)
###Stratified sample
##Population information

pi = c(0.50, 0.30, 0.20, 0.90, 0.40, 0.70) ##Real
    subpopulation support
pg = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6) ##Real
    population weights
p = sum(pi*pg) ##Total population support

##Sample info

ng = c(6000, 6000, 6000, 6000, 6000, 6000) ##Sub-
    sample sizes
```

```
wg = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6) ##Claimed
    weights

##Generating our polls

x1 = rbinom(1e6, size=ng[1], prob=pi[1])
x2 = rbinom(1e6, size=ng[2], prob=pi[2])
x3 = rbinom(1e6, size=ng[3], prob=pi[3])
x4 = rbinom(1e6, size=ng[4], prob=pi[4])
x5 = rbinom(1e6, size=ng[5], prob=pi[5])
x6 = rbinom(1e6, size=ng[6], prob=pi[6])

##Calculating thestratified estimator
estStS = wg[1]*x1/ng[1] + wg[2]*x2/ng[2] + wg[3]*x3/
    ng[3] + wg[4]*x4/ng[4] + wg[5]*x4/ng[5] + wg[6]*
    x4/ng[6]

##Finally MSE

mean( (estStS-p)^2 ) ##We get an answer of 0.0136

#####Simple random sample
p = 0.50 ; n = 6000

x = rbinom(1e6, size=n, prob=p)
estSRS = x/n
mean( (estSRS-p)^2 ) ### We get an answer of 4.166e
    -5
```

Make sure to read over it carefully. The simple random sample estimator has a smaller MSE than the stratified.

4. Using the Table **??** gender information for the actual voting results, estimate a 95% confidence interval for Clinton's support in the general population.
Solution: *Using the gender information, we will find the estimated Clinton support for the general population is:*

$(0.53)(0.53) + (0.47)(0.47) = 50.18\%$

With the margin of error: $1.96\sqrt{0.53^2\frac{0.5(1-0.50)}{410} + 0.47^2\frac{0.50(1-0.50)}{392}} = \pm 3.5\%$

5. Using the Table **??** party affiliation information for the actual voting results, estimate a 95% confidence interval for Clinton's support in the general population.
Solution: *Using the party information, we will find the estimated Clinton support for the general population is:*

$(0.37)(0.34) + (0.31)(0.38) + (0.28)(0.33) = 34\%$

With the margin of error: $1.96\sqrt{0.34^2\frac{0.5(1-0.50)}{267} + 0.38^2\frac{0.50(1-0.50)}{298} + 0.28^2\frac{0.50(1-0.50)}{231}} = \pm 3.5\%$

3

Combining Polls

Review

1. Our first attempts at averaging polls was to simply average the proportions. In that case, how did we determine that averaging the sample sizes was better than adding them?
Solution: *On page 60, we ran an experiment that generated results from five polls at a fixed value of π. We then found the coverage rate of the experiment (how often π was actually in our interval). When we did this, we found for adding samples π was only present 55% of the time. Because it does not get us close to our claimed coverage, compared to averaging the sample sizes, we determined averaging was the way to go.*

2. Why does a weighted average make more sense than a simple average?
Solution: *When we use the weighted average, we are using the number of people in the poll who support a certain canidate; in other words, we are treating multiple polls as just one large poll, which makes calculations more efficient and the confidence intervals are more practical.*

3. What is the difference between accuracy and precision?
Solution: *Accuracy tells us how close a measurement is to the true value, whereas precision refers to how close the measurements of the same experiments are to one another. Another way to think about precision is how repeatable a measurement is amongst experiments.*

4. How did we measure accuracy in Section **??**?
Solution: *An example of measuring accucacy is on page 61, where we ran the experiment and determined how close the claimed cov-*

erage was to the actual coverage.

5. How did we measure precision in Section **??**?
   Solution: *We measured precision in the first expiriment when we ran multiple polling expiriments to determine whether or not π was in our confidence interval. This is an example of precision because we are repeating the experiment and trying to see how often we come across π.*

6. Why is it important to take into consideration the age of the poll?
   Solution: *It is important to consider the age of a poll because there could be things that affect a candadite's approval rate. For example, what if there is a huge scandal in April and their rate shoots downward, but in May, it was revealed to be a false accusation? This would cause, potentially, pretty large changes across the board.*

7. Which of the three weighting functions described in the text is the best? How would we know?
   Solution: *Really, it depends. The rectangular function has a potential of weighting heavily if, during that range of time, a candidate has a decreasing or increasing amount of support because it emphasizes recent polls; but, the width of the rectangular function can be tampered with. The triangular function reduces the weight on the poll as time passes, so it will still weight the newer polls more heavy. The gaussian function weights more-recent polls heavier than the older polls, but still gives a positive weight to everything.*

   The functions that weight newer polls more have confidence intervals with higher endpoints. With this being said, the narrower the interval, the more precise. So, in the case of South Korea and using Table 3.1, I would say the triangular and gaussian function would be our best bets because they seem the most narrow and the best distribution of weights from old and new polls, whereas the rectangualr function takes too much weight on the newer polls.

8. Why did we use linear regression in this chapter? When would it be appropriate to use? When would it *not* be appropriate?
   Solution: *We used weighted least squares regreession (WLS) to*

make predictions for support levels of candidates. Regression is appropriate to use when the proper requirements are met, such as normality of the residuals, constant expected value of the residuals, and constant variance of the residuals. It is not appropriate when making future estimations or if our data doesn't have the proper distribution we need to preform OLS or WLS.

Conceptual Extensions

1. Using just the maps of Figure **??**, can you tell which candidate is ahead? What additional information would you need? Why are maps like these misleading for determining which candidate is ahead?
   Solution: *It is really difficult to tell who is ahead; one may notice more lighter splotches on the map for Hong compared to Ahn, yet there are many more darker splotches on Hong's graphic than Ahn's. It may be useful to have population information. So, if one district has a small population and not a lot of votes compared to a mich larger support rate for a much larger population, this will not affect much. The maps are misleading because we are missing information (dates, population numbers, where the capitol is, etc).*

2. The basic assumption for this chapter (page **??**) is that the polls are estimating the same population. Why does this assumption need to be stated?
   Solution: *This assumption needs to be stated because determining likely voters and non-likely voters is up to the polling stations. Although when populations tend to be similar and errors are minor, this could cause the wrong winner. Futhermore, the thought process (and statistics) would change if we did not operate under this assumption.*

3. What is a "likely voter," and how would you determine one?
   Solution: *A likely voter is someone who has indicated strong intentions to vote on election day. Determining a likely voter could be done in a few ways, but one is a survey sent out by the polling houses looking for interest in likely voters.*

4. Why is it more satisfying to weight polls with larger sample sizes more than those with smaller?
Solution: *The larger the sample size, the more information we can get. When we have more information, it can make predictions and confidence intervals more accurate and precise.*

5. Why do we expect the coverage rates in Section **??** to be close to 95%?
Solution: *We believe that the paramater, π, will be in the confidence interval 95% of the time. So, because we are working with this prediction at a 95% confidence interval, we can expect the coverage rate to be pretty close to 95%.*

6. What would happen to the interval widths if the confidence level is changed from 95% to 90%? What would happen if we change them to 100%?
Solution: *If we make our confidence interval lower, then we will expect the widths to get wider since we are not as precise. I would not recommend doing a 100% confidence interval, because that is practically saying there is no way your prediction is wrong, but the widths would become very narrow since there is such certainty.*

7. How should we determine the appropriate weighting function (Figure **??**)?
Solution: *The appropriate weighting function can be determined based off where you want the weight; for example, notice the rectangular function seems to wait fairly heavy on just recent polls, whereas the triangular and Gaussian use the old polls as well.*

8. Why were the estimates so good for Moon and Hong, but so wrong for Ahn?
Solution: *Using regression assumes that current trends continue, and there is nothing in the data that suggests why Ahn's polls are falling. So, the regression did not really take into account things like die-hard supporters, increase of support if a scandal was found to be false, media, and many other factors. Because the regression just took that decline in data and ran with it, it caused things to go*

wrong for Ahn.

Computational Extensions

1. Estimate the Hong support level on election day using other weighting functions. Which of the weighting functions of Table **??** produces the best estimate of his actual support level?
   Solution: *Below is the R code for this problem. Make sure you understand it, as well as the explanation at the end.*

```
###### This is just getting the functions ready
     ########

## Rectangular function

rectangleWeight <- function(daysAgo, theta, n) {
  weight = ifelse(daysAgo>theta, 0, 1)
return(n*weight)
}

## Gaussian weight

gaussianWeight <- function(daysAgo, theta, n) {
  weight = dnorm(daysAgo/theta*1.17741/dnorm(0))
return(n*weight)
}

###### Now we are going to load the data
library(readr)
korData <- read_csv("C:/Users/18153/Downloads/
    kor2017pres-polls.csv")

attach(korData)

###Rectangle function##############33

pHong = ‘Liberty Korea‘
nHong = round(pHong*n)
allPoll = rectangleWeight(129-doy,7,n)
allHong = pHong*allPoll

N = (sum(allPoll))
X = ( sum(allHong))

wald.test(X,N)

############Guassian function####################

allPoll1 = gaussianWeight(129-doy,7,n)
allHong1 = pHong*allPoll1
```

```
N1 = (sum(allPoll1))
X1 = ( sum(allHong1))

wald.test(X1,N1)
```

When using the rectangular function, we get a 95% confidence interval of 18.0% to 19.2%, which a little larger than the triangular function. When using the Gaussian function, we get a 95% confidence interval of 10.1% to 26.0%. This is larger than both the rectangular weight and triangular weight function, but is the only interval that contains Hong's support level (which is 24.03%).

2. Devise an experiment to determine when the Wald procedure is superior to the exact Binomial procedure in estimating the population proportion.
Solution: *Recall in chapter 1 how we determined when the Agresti-Coull procedure was superior to the sample proportion. We ran an experiment in R, and we will do something similar here:*

```
## Initialization
n = 15
pi = 0.700
confLevel = 0.95


# Set aside internal memory
coverB = coverW = numeric()
clB = clW = numeric()


# ## Begin experiment
for(i in 1:1e4) {

sample = rbinom(1 , size=n , prob=pi) ## The sample

# Binomial
clB = binom.test(x=sample,n=n)$conf.int
coverB[i] = ( clB[1]<=pi && pi<=clB[2] )


# Wald
clW = wald.test(x=sample,n=n)$conf.int
coverW[i] = ( clW[1]<pi & pi<clW[2])
}

# End experiment


# ## Results
mean( coverB ) ## Coverage using binom.test
mean( coverW ) ## Coverage using wald.test
```

Notice that the Wald coverage is 95% and the Binomial coverage is 98%. When the sample size is very large, both will become about the same. As the sample size gets smaller, the Binomal is the winner. Play with these paramaters and see when the Wald test beats the Binom test in other situations.

4

Digit Tests

Review

1. Why would it be easier for counting fraud to take place in the 2009 Afghan election than a typical Ghanaian election?
   Solution: *Recall on page 100, Ghana sends the ballots to the national electoral commission. They record all the votes for the polling stations and have candidate agents sign off on the legitamacy. They are then forwarded to the Electoral Commission, they are counted, and the winner is announced. In Afghanistan, the ballot papers are sent to a central counting facility. Because of the centralizd counting, it is much easier to have fraud since one has to control such a small group.*

2. What conditions are necessary for an election to be democratic?
   Solution: *An election needs to be free and fair. In terms of free, each citizen has a right to vote. For fair, this vote actually has to matter; this means that the votes all have the same probability of counting and one demographic doesn't have a higher probability than the other.*

3. How did Newcomb discover the Benford distribution?
   Solution: *The Benford distribution was discovered by looking at the wear and tear of logarithmic tables in old mathematical books. He concluded that one could distinguish between these digits in nature and their logarithms. Benford went onward and made the same discovery, but took it a step further with his research.*

4. What is the difference between 'natural' and 'unnatural' numbers according to Benford?
   Solution: *Benford believes that natural numbers are those that have deviations from random behavior. Unnatural numbers are tampered*

with. As they are more randomized by humans rather than nature, the vote-count fraud will produce a distribution different from the Benford distribution, in turn, making it detectable for election fraud.

5. What is the purpose of the Chi-square test?
   Solution: *Kind of hinting to the above question, the Chi-square distribution is useful to detect for election fraud when the vote-count follows a distribution that isn't from the Benford distribution.*

6. Why does ignoring information in the data tend to lead to a less-powerful test?
   Solution: *In the case of the Benford test, it is superior to the normal approximation we have seen in previous sections. The Benford test uses the entire distribution of leading digits, and because more information is used, it will be more powerful. This being said, ignoring information and not using all of it could lead to a bad test and no detection of counting fraud.*

Conceptual Extensions

1. Flip a fair coin 1000 times and create a histogram of run length. Estimate the average run length. Now, simulate those 1000 flips by hand. That is, randomly write out H and T values as you *think* would arise from flipping a coin. Create a histogram of run length and determine the average run length in your fake data. Compare the results.
   Solution: *We will use R for this.*

```
headAndtail <- function(n, x, p){

  sample = rbinom(1000, 1, 0.5)

  runs <- diff(sample)
  n = 0
  for(i in 1:1000){
    if(runs[i]==0){
      n = n+1
    }
  }
  print(n)
  hist(n)
  print(mean(n))
}
```

Make sure you know what the code is doing.

2. Why is the expected distribution of initial digits not uniform? What is the intuition behind the Benford distribution?
   Solution: *The intuition behind the Benford distribution was the logarithm function, as that is how the pattern of the leading digits was recognized. Because of the log function that is being applied, like on page 105, the probability of the distributions decacy. If it was a uniform distribution, then the probabilities would be the same and the graphic would be much more ‘box‘ like as opposed to the downwards slope.*

3. What is the Central Limit Theorem, and how is it used in the quick example on page **??**?
   Solution: *The Central Limit Theorem (CLT) is araaguably the most imporrtant theorem in statistics. Let X be a random variable with a mean μ and finite variance σ^2. Let us draw a random sample of size* n *from this distribution. Then, as* n *gets larger, the distribution of the sample sums converge to normal. In this example, it rapidly comes into play because of the large sample size* ($\mathrm{n} = 34$).

4. Why is 10 the usual base of the logarithm used in the Benford test? Would using a different base work equally well?
   Solution: *Using a different base other than ten doesn't matter, just so long as you are consistent with that base in all your calculations. Statisticans, mathmaticians, and other logarithm users typicall go with the generic base 10, just out of habit. Plus, most programs have the log function set at base 10, so you would need to change it and add a bit more work for yourself if you wanted a different base.*

5. What is the “Two Policeman and a Drunk” theorem and why is it useful in proofs?

Solution: *The Two Policeman and a Drunk theorem (otherwise known as the Squeeze theorem) is regarding the limit of a function that is between two other functions. It is a visual to imagine two policemen walking a drunk man to a prison cell. If the officers are going to the cell, then (even if the drunk man is wobbling about) he will also end up in the cell. It is useful in proofs because it can give us insight to the limit, simplifying the problem without much hassle.*

6. In Figure **??**, at what points is the mean digit the greatest?
Solution: *The mean is the greatest at the discrete digits 1 through ten (so 1, 2, 3, ... , 10).*

7. Why would one expect the likelihood simulation method of Section **??** to produce a less-powerful test than either of the two multinomial averaging methods of Section **??**?
Solution: *We have more information in the multinomial averaging than just a likelihood simulation. Because we are averaging more Benford distributions or thetas (using more expected leading digit frequencies), so both of the multinomial methofs are superior than the likelihood simulation.*

8. Which of the two multinomial averaging methods of Section **??** would you expect to be more powerful?
Solution: *I would expect the second multinomial method to be more powerful. It averages up multiple Benford distributions which really optimizes the information gain. But, there will be cases in which the two multinomial methods are strikingly close.*

9. What conclusions can be drawn from Table **??**?
Solution: *Using a significance level of 0.001, we can conclude that there is no evidence of counting fraud in the elections. We are also able to compare MAI and MAII to determine how different (and similar) both of these tests really are.*

Computational Extensions

1. Find the populations of all the countries in the world. Write down the initial digit of those populations. Determine if that initial digit follows the Benford distribution sufficiently closely.
   Solution: *I am just going to find countries with leading digits 1 through 9 and show the distribution. If you feel bold, go ahead and find them all.*

Country	*Population*	*Leading digit*	*Benford*
China	*1,439,323,776*	*1*	*Yes*
United States	*331,002,651*	*3*	*Yes*
Nigeria	*206,139,589*	*2*	*Yes*
Argentina	*45,195,774*	*4*	*Yes*
Servia	*8,737,371*	*8*	*No*
South Korea	*51,269,185*	*5*	*Yes*
Sierra Leone	*7,976,983*	*7*	*No*
Belarus	*9,449,323*	*9*	*No*
France	*65,273,511*	*6*	*Yes*

   After 6, the leading digits quit following the Beford Distribution.

2. Find the populations of $n = 100$ cities in the world. Write down the initial digit of those populations. Determine if that initial digit follows the Benford distribution sufficiently closely.
   Solution: *Like the above example, we can conclude which digits will and will not follow the Benford distribution.*

3. Some researchers claim that the "Second digit Benford test" is better for detecting problematic election counts. The "Second digit Benford distribution," BENF_2, has the same assumptions underlying the first digit test, it is just applied to the second digits. Determine the theoretical distribution of second digits. Test the Afghan data using BENF_2.
   Solution: *For this problem, we will need to update some of our R functions:*

```
getSecondDigit <- function(count) {
  as.numeric(substring(count,2,2))
}

getSecondDigitDistribution <- function(count) {
  ld = getSecondDigit(count)
  tabulate(ld,9)
}

SecondBenfordTest <- function(counts) {
  SecDist = getSecondDigitDistribution(counts)
chisq.test(SecDist, log10(1+1/(1:9)))
}
```

Then, we will run the functions like before:

```
SecondBenfordTest(d[,10])
SecondBenfordTest(d[,13])
SecondBenfordTest(d[,20])
SecondBenfordTest(d[,31])
```

We will get something like this:

Candidate	*p-value*
Ashraf Ghani Ahmadzai	*0.3045*
Dr. Abdullah Abdullah	*0.2705*
Hamed Karzai	*0.3045*
Ramazan Bashardost	*0.2867*

We can determine that there is no sinificant evidence of fraud in this electiion, even using the second Benford test. Notice that the p-values did decrease, so the second Benford distribution could be useful.

5

Differential Invalidation

Review

1. Compare and contrast the four regression methods discussed in this chapter.
Solution: *We looked at four different types of regression: Ordinary Least Squares (OLS), Weighted Least Squares (WLS), Binomial Regression (GLM), and Beta-Binomial Regression (VGLM). The objective of OLS is to minimize the sum of the squared errors. We also need to make asumtions about the residuals: they follow a normal distribution, they have a constant mean, and they have a constant variance. In many real-life applications, the variances are heteroskedastic, meaning they are not constant. In this case, a typical solution is WLS as a way to estimate the slopes, and the new objective is to minimize the weighted sum of the squared errors. In the case of normality violation and when the dependent variable follows a binomial distribution, we can use a GLM. This helps to improve WLS. Last but not least, we looked at a VGLM. Recall we did this when the data was overdispered; for example, people that live in the same area typically share the same political views, which makes the data 'clumpy'. Using a VGLM can help adjust to this idea.*

2. What is differential invalidation?
Solution: *Differential invalidation is when votes for one demographic is invalidated at a higher rate than the other.*

3. What is ballot box stuffing?
Solution: *Ballot box stuffing is when one person submits many ballots during a vote in which only a signle ballot per person is legal.*

4. How does differential invalidation affect the electoral divisions displayed on an invalidation plot.

Solution: *When the ballots are systematically invalisated based on the candidate they were cast for, then there is a significant slope that becomes distinct in the plot.*

5. Compare and contrast statistical significance and practical significance.
   Solution: *Statistical signifigance arises from modeling and p-values and the hypothosesis, whereas practical significane is the ability to provide a more reasonable bound. A way I like to think about this is when detecting fraud, we should be testing with a* very *small singificance level. Accusing someone of electoral fraud is a huge deal, so we need to be sure, which is where practical signifigance comes from. We need to view patterns, how much things are changing, and why they are changing. Statistics is a tool to help us get to these conclusions.*

Conceptual Extensions

1. Why do the conclusions in this chapter never state that an election was unfair?
   Solution: *This brings back the question regarding statistical and practical significance; we can have statistical evidence that there is differential invalidation, but this is indication and not solid truth. Like stated, accusing a candidate of fraud is a pretty big deal, so we need to be clear between statistical significance and practical significance.*

2. How does ballot box stuffing affect the electoral divisions displayed on an invalidation plot.
   Solution: *Ballot box stuffing will look about the same as differential invalidation; as more votes are cast for one candidate, then the invalidation rate will decrease. Therefore, in the case of ballot box stuffing, the slope will go from insignificant to significant an have a slope.*

3. What does a p-value indicate?

Solution: *The p-value tells us how likely our null hypothesis is to occur; a very small p-value tells us to reject our null hypothesis (as it has a very small probability of actually occuring) and a large p-value tells us to accept the null hypotesis (as it has a very large probability of occuring).*

4. If you is concerned with practical significance, should you pay attention to the confidence interval or the p-value?
Solution: *When concerned with practical significance, for example, we may want to be paying attention to how much the invalidation rate is increasing. This is easier to do with confidence intervals as opposed to p-values. It is true that a p-value going from 0.05 to 0.0005 may give us a little more insight into what is happening election to election, but confidence intervals give us a solid range. Really, it is up to the scientist.*

Computational Extensions

1. Perform the generalized Benford test on the Côte d'Ivoire data to see if there is evidence of vote count-fraud in the 2010 election.
Solution: *Below is the R code I used for this problem:*

```
dt <- read.csv("https://ews.kvasaheim.com/data/
    civ2010pres2cei2.csv")
attach(dt)
BenfordTest(dt[,6])
BenfordTest(dt[,7])
```

When ran, the p-value for Gbagbo was 0.3564 and the p-value for Outtara was 0.5332. Thus, we do not have significant evidence of vote count fraud in the 2010 Côte d'Ivoir election.

6

Considering Geography

Review

1. What is differential invalidation?
Solution: *Differential invalidation is when votes for one demographic is invalidated at a higher rate than the other.*

2. How does one detect differential invalidation?
Solution: *Differential invalidation can be detetected using regression and by using spatial analysis.*

3. Why does one typically use distance measures instead of contiguity measures for **W**?
Solution: *Contiguity measures are important for allowing us to determine whether or not spatial correlation can be detected; however, using distances allows the* **W** *matrix to have a large enough sample size so the Central Limit Theorem can help.*

4. What is the **J** matrix? Let **A** be a 3×3 matrix. What are **JA** and **AJ**? In other words, what deos pre- and post-multiplying by **J** do?
Solution: *Recall that* **J** *is a square matrix of all 1's, so it is* $n \times n$ *in size.* **JA** *will be a* 3×3 *matrix of all 3's, as will* **AJ**. *Pre and post multiplying by* **J** *does the same thing, since* **J** *is an* $n \times n$ *matrix of all 1's.*

5. Why is `quasibinomial` used instead of `binomial`?
Solution: *We use the* `quasibinomial` *in order to adjust for overdispersion. It is no surprise when citizens from the same area vote similarly, so the* `quasibinomial` *helps adjust for this 'clumpiness'.*

6. What does the `quasi` mean in `quasibinomial`?
Solution: *`Quasi` is a suffix that says "having a resemblance of" or "close to" something. By using `quasi` in front of binomial, it tells us that the data follows a distriubtion similar to the binomial, but needs a bit of correction.*

7. How important is it to select the correct weighting function (Table **??**)?
Solution: *It isn't necassarily the kernel function that is important, but the choice if the bandwidth. So, the choice of the weighting function kernel is rather arbitrary, but the bandwidth should be prioritized.*

8. Compare and contrast all four geographic methods covered in the chapter. What were their strengths and their weaknesses?
Solution: *The first model we looked at was the Spatial Lag Model (SLM). THe SLM includes a temporally lagged dependent variable as an independent varaible. This lagged deoendent variable is a nieghborhood average. What's interesting is the neighbor affect, ρ, can be determined as biased or unbiased by using the SLM. Although the SLM is effective in reducing spatial correlation, it makes a bold assumption that the parameter effect is constant. We then looked at the Casetti's Spatial Expansion Model (SEM). The SEM model differs because it models spatially varying effects; therefore, it uses a function of β_1. The major drawback of the SEM is instead of using a linear fit, it uses a polynomial. A high degree polynomial will allow the estimated effect function to match true variation, but it will reduce the degrees of freedom, which infaltes the p-values. We then looked at the Geographically Weighted Regression model, which uses weighted regression functions. Because of this, it does not use the typical measure for degrees of freedom. There are not many drawbacks to a GWR, other than the fact the test should not be used until the degrees of freedom can be determined. The last model e looked at was the Spatial Lagged Expansion Method (SLEM). THe SLEM model is a combination of the spatial lag model and the expansion method to allow flexibility for geographically weighted regression. The drawback to this model is it was created really for explatory analysis; but, it produces a better model than GWR.*

Conceptual Extensions

1. Why might candidate support be spatially correlated?
   Solution: *Candidate support can be spatially correlated because people near one another tend to vote similarly.*

2. What do Matruh and New Valley (*El Wadi El Gedid*) governorates have in common?
   Solution: *Matruh and New Valley are both to the West of Egypt, where support for the 2011 Referendum was very high. In these governorates, the invalidation rate was fairly low.*

3. What do Cairo and Giza governorates have in common?
   Solution: *In Giza and Cairo, the support for the 2011 Referendum was low. In turn, the invalidation rate for these two governorates was higher than that of the governorates that did support the Referendum.*

4. Referring to page **??**, what would $\mathbf{W}^3$ represent? How many zero entries would it have?
   Solution: *The* $\mathbf{W}^3$ *represents a third-order contiguity. So, we are concerened with neighbors of neighbors of neighbors. Therefore, we would have 24 zeros in our matrix (assuming we are using rook conginuity).*

5. How would one determine the functional form in Casetti's SEM?
   Solution: *We can determine how many extreme a map has (minimum or maximum) and use the correct polynomial to find the right function. This could be an issue the higher we go in polynomial from because we risk losing degrees of freedom.*

6. What methods are available for testing hypotheses using geographically-weighted regression (GWR)?
   Solution: *We could take the GWR a step further and use a SLEM model, or we can test significance using a t-test.*

Computational Extensions

1. Fit the model of Section **??** using linear effects.
   Solution: *Below is the R code used for this problem:*

```
source("https://ews.kvasaheim.com/Rfctns/
    electionTesting.R")
install.packages("spgwr")
library(spgwr)

de <- read.csv("https://ews.kvasaheim.com/data/
    egy2011referendum.csv")
attach(de)
View(de)

pInv = INVALID/TOTAL
pSup = VALID/INVALID

dists <- as.matrix(dist(cbind(north, west)))
dists.inv <- 1/dists
diag(dists.inv) <- 0
w =dists.inv

J = matrix( rep(1, 27^2), ncol=27)
wstar =  (w) / (w \%*\% J)
contagion = wstar \%*\% pInv

modSLEMlinear <- lm(pInv ~ west*north*contagion + I(
    west) + I(north) + west*north*pSup + I(west)*
    pSup + I(north)*pSup)
summary(modSLEMlinear)
```

   Notice we take away the squared u and v variables (west and north) to use the linear effect. Below is the regression table.

   Notice the changes. Our adjusted R^2 also has decreased.

	Estimate	Std. Error	t-value	p-value
Nominal Effect				
Intercept	5.031	5.877	0.856	0.405
North	-0.1771	0.1963	-0.902	0.381
West	0.1428	0.1811	0.788	0.443+65
N × W	-0.005080	0.006058	-0.839	0.415
Neighbor Effect				
Intercept	-517.3	636.7	-0.813	0.429
North	17.52	21.42	-0.818	0.426
West	-15.02	19.72	-0.762	0.458
N × W	0.5098	0.6643	0.767	0.455
Referendum Effect				
Intercept	0.3669	1.532	0.239	0.814
North	-0.003254	0.05220	-0.602	0.951
West	0.1456	0.04887	0.298	0.770
N × W	0.0001901	0.001670	-0.114	0.911