Voting

1 Predicting a vote

Lots of things are put to a vote. When a new toothbrush is marketed, consumers vote if they like it with their wallets. A new movie is released and people vote to go or not. In politics, a candidate is elected based on a vote. Clearly voting is important as is being able to predict the outcome of a vote. These predictions or made based on focus groups, phone surveys, street interviews, online surveys, anecdotal evidence or other means.

What we look at in this project, is how to interpret the results of a survey. For simplicity, our surveys ask a sample of people if they are "for" or "against" something. To simplify, they are one or the other. There are no (real) issues of non-response, question bias etc. In particular we can ask lots of questions about a survey: What does the survey predict? How well does it predict it? What assumptions do you need to know about a survey to make inferences? What assumptions are tacit, what should be explicit?

? Question 1: A survey was taken from likely voters about their choice of candidate in an upcoming election. Candidate A received 40% of the yes votes, Candidate B received 60%. Which candidate will win? Why? How do you know? What more would you like to know about the survey?

? Question 2: A person driving around her town notices that candidate A has twice as many signs as candidate B. Does this mean candidate A is likely to receive twice the votes come election day? Why or why not?

? Question 3: How much more "accurate" do you expect a survey of 1000 people to be compared to one of 100 people. Be precise in your response.

? Question 4: Online polls are very common. A new one appears daily on `cnn.com`. The results are accompanied by this disclaimer:

> "This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole."

What does the word "Scientific" refer to? When does a survey "reflect" the opinions of a larger population than that which is surveyed? Why do you think they write this?

2 Simulating a survey

To get a better understanding of a survey, we are going to *simulate* a survey where we know exactly what the voters will do. Suppose there are 1000 voters, 420 "for" and 580 "against". We will "survey" 100 voters and find a sample percentage of those "for".

```
> people = c(rep(1, 420), rep(0, 580))
> no = sum(sample(people, 100))
> no

[1] 43
```

In this *sample* we got 43% (43 out of 100) were "for".

We first defined the population. If you look at `people` you will see 420 "1"'s and 580 "0"'s. This is done using `rep` to *repeat* the values. The `sample()` function selects 100 values of `people` at random, and the `sum()` adds up just the 1's which we take to be those who will vote "for".

Our *simulation* will repeat this sampling 1000 times to see what the distribution is of the results of a single sample. To do so uses a *for loop*.

```
> res = c()
> for (i in 1:1000) res[i] = sum(sample(people, 100))
```

That's it. For each value of `1:1000`, the `for()` loop assigns `i` this value and then evaluates the command. In this case it samples and then *assigns* the value to the ith component of our results.

Now that we have a sample lets take a look at it. Remember, each sample is a *random* value, so the results are a *distribution* of values. We want to know things like the mean, or the variation of the results. First, lets look at a histogram and density

```
> hist(res, probability = TRUE)
> lines(density(res))
```

Notice in figure 2, the distribution looks bell-shaped, symmetric – not skewed, and unimodal. Also notice that there are some values more than 50. That is sometimes the "for" people win the survey, yet they will lose the actual vote by 160 votes.

How often is the sample percentage 50% or more? We simply need to take a proportion

```
> prop.more = sum(res >= 50)/length(res)
> prop.more

[1] 0.06
```

About 6% of the time. We conclude that there is a fair amount of variation in this sampling.

Yet, the average value of the results is more or less exactly as we expect

```
> summary(res)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.00   39.00   42.00   42.02   45.00   57.00
```

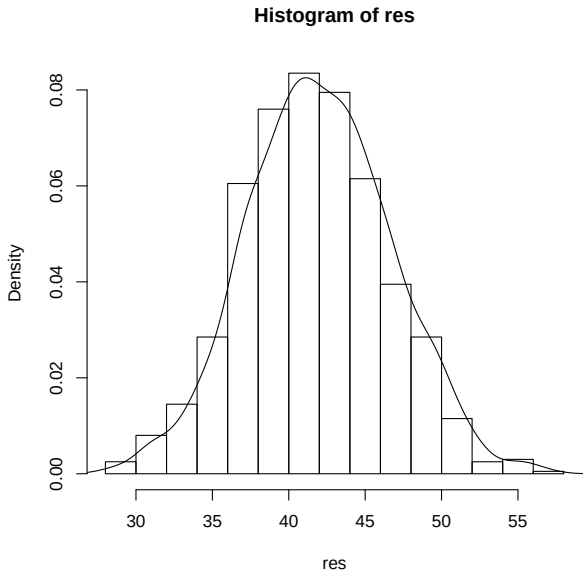

Figure 1: Histogram of 1000 random samples of size 100.

The average we see is 42%, and our answer is exactly 42% (from 420/1000). Looking at the `summary()` output we see that 1/2 the time the sample is between 39% and 45%. We might say that 1/2 the time the sample proportion is within 600 percentage points of the true proportion. That is the difference between the 0.25 quantile (Q_1) and the 0.75 quantile (Q_3). Usually though, when you hear a value of 600 percentage points error, you don't say "1/2 the time". Rather, it is implied that with high probability (or most of the time) it is so. Often implied is 95% of the time. That is, between the 0.025 quantile and 0.975 quantile if we want symmetry. From the sample, this is

```
> quantile(res, c(0.025, 0.975))

 2.5% 97.5%
   33    51
```

Or a "margin of error" of 18 percentage points.

Finally, a note about the sampling. The `sample` command does a *sample without replacement*. That is once we ask a person, there is no chance of asking them again. If we allow for this chance, then we *sample with replacement* and then a familiar probability model can be used. Thus, in this case the number of "for"'s has a *binomial distribution* as each person asked in the survey has a Bernoulli trials response: the answer has only two possibilities, and each time the responses are equally likely and independent of the others.

It seems intuitive, that there isn't much difference between sampling with or without replacement if the size of the sample is "small" compared to the size of the population. Let's do a simulation like the above to compare in this specific case.

```
> res1 = c()
> for (i in 1:1000) res1[i] = sum(sample(people, 100, replace = T))
> summary(res1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   39.00   42.00   42.21   45.00   57.00
```

The two are different as we'd expect as these are random samples, but not much so. You can compare the two distributions further with side-by-side boxplots or a q-q plot (Figure 2.)

```
> qqplot(res, res1, main = "compare sampling w/ and w/o replacement")
```


```
> qqplot(res, res1, main = "compare sampling w/ and w/o replacement")
```


Figure 2: Comparing sampling with and without replacement

Question 5: Do your own simulation of sampling without replacement. What is the mean of your results?

Question 6: From your sample, what is the margin of error if we want to be correct 95% of the time? How about 90% of the time? Is there some relationship between the two?

Question 7: Repeat the simulation, only this time use a sample of just size 10 (that is use `sample(people,10)` at the appropriate point). How does this new distribution of numbers compare to your previous one?

Question 8: By sampling with replacement you ensure that the random sample is like a Bernoulli trial. Explain why allowing the possibility of picking a person more than once makes the responses independent and identically distributed.

3 Using the binomial model to analyze

Now, suppose we don't know the answer, but we convince ourselves that we can do a survey where each person selected chosen from the population with equal probability and is chosen independently of the others.

For concreteness, suppose the true percentage of people "for" is called p. We take a sample of size n, and we have a proportion of people "for" called $\hat{p}$ =(number "for")/(number questioned).

We have the following. The *population proportion* is an unknown parameter p is what we are trying to find out. It is a number between 0 and 1. The *sample proportion* is a random number $\hat{p}$. We know this value and it is between 0 and 1. We would like what $\hat{p}$ tells us about p.

In our simulation, we could tell alot of how $\hat{p}$ tells us about p as we could see the sample distribution of $\hat{p}$ from our simulation of 1000 values. In this case, we only have one value of $\hat{p}$. However, we do know it's sampling distribution. If the assumptions are met then

The number "for" has a Binomial distribution with parameters n and p.

The number for is $n\hat{p}$ and is a binomial. As $\hat{p}$ is binomial divided by n, the distribution of $\hat{p}$ has a mean of the binomial number divided by n or $(np/n) = p$. and a standard deviation which is the standard deviation of the binomial number divided by n or

$$\mu(\hat{p}) = p \quad \mathbf{SD}(\hat{p}) = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}.$$

Furthermore, if np and $n(1-p)$ are large enough, we know that the binomial distribution is approximated by the normal distribution, and so we can conclude under the assumption of a large enough sample that $n\hat{p}$ is approximately normal with mean np and standard deviation $\sqrt{np(1-p)}$ and so

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately normal with mean 0 and standard deviation 1 (a standard normal).

? Question 9: Show algebraically that the formula for Z comes from standardizing the Binomial random variable $n\hat{p}$.

? Question 10: The standard deviation of $\hat{p}$ depends on n. As n gets larger what happens to $\hat{p}$'s standard deviation?

? Question 11: A n gets larger what happens to the mean of $\hat{p}$?

? Question 12: Explain how $n\hat{p}$ (the number "for" in a sample) can be approximately normal when its vale is between 0 and n, yet a normal can take any value between $-\infty$ and $+\infty$?

4 Confidence intervals

The value of Z above is the standardized sample proportion and if our assumptions are valid it has a standard normal distribution. From this, we know the following: 68% of the time Z is between -1 and 1, 95% of the time Z is between -2 and 2, and 99.8% of the time it is between -3 and 3.

Let's focus on the 95% of the time case. What this means is the probability that Z is in the interval $[-2, 2]$ is 0.95. Or in formulas

$$0.95 = P(-2 \le Z \le 2) = P(-2 \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le 2) = P(a \le p \le b),$$

where a and b could be figured out using the quadratic equation. We can simplify the algebra at the expense of a slight approximation. We expect that $\hat{p}$ is "close" to p, so if we replace the standard deviation in the denominator ($\sqrt{p(1-p)/n}$) by the *standard error* defined by

$$\mathbf{SE}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} \quad \text{compare to } \mathbf{SD}(\hat{p}) = \sqrt{[(1-p)/n}.$$

Then we can solve easily to get

$$0.95 = P(\hat{p} - 2 \cdot \mathbf{SE} \leq p \leq \hat{p} + 2 \cdot \mathbf{SE}).$$

That is, 95% of the time, the true value of p is within 2 standard errors of the sample proportion. We say that 2 standard errors is the *margin of error*, the interval $[\hat{p} - 2 \cdot \mathbf{SE}, \hat{p} + 2 \cdot \mathbf{SE}]$ is a *confidence interval* and the *confidence level* is 95%.

More generally, if α and z^* are related by the equation

$$1 - \alpha = P(-z^* \leq Z \leq z*)$$

then the margin of error is $z^* \cdot \mathbf{SE}$, the confidence interval is found from $\hat{p} \pm z^* \cdot \mathbf{SE}$ and the confidence level is $(1 - \alpha) \cdot 100\%$.

Often z^* is written $z_{\alpha/2}$ from how it is found as (why?)

$$1 - \frac{\alpha}{2} = P(Z \leq z^*).$$

This is how it is found with R: (notice, z is the $1 - \alpha/2$ quantile for the distribution of Z)

```
> alpha = 0.05
> qnorm(1 - alpha/2)

[1] 1.959964

> alpha = c(0.01, 0.05, 0.1)
> qnorm(1 - alpha/2)

[1] 2.575829 1.959964 1.644854
```

To do an example, suppose 1142 likely voters are asked if they would vote for candidate A or B and 613 said they would vote for A. If this was a random sampling, what is the 95% confidence interval for the population proportion? What is the 90% confidence interval?

```
> n = 1142
> phat = 613/n
> SE = sqrt(phat * (1 - phat)/n)
> alpha = 0.05
> z = qnorm(1 - alpha/2)
> phat

[1] 0.5367776

> c(phat - z * SE, phat + z * SE)

[1] 0.5078570 0.5656982
```

```
> alpha = 0.1
> z = qnorm(1 - alpha/2)
> c(phat - z * SE, phat + z * SE)

[1] 0.5125067 0.5610485
```

We see that $\hat{p}$ is 0.54 and the 95% confidence interval is $[0.5, 0.6]$. The 90% confidence interval is narrower. In either case, p is bigger than 0.50 so it appears likely that candidate A will win.

There is an easier way to get these numbers[1] using the built-in function `prop.test`. To illustrate

```
> prop.test(613, 1142)

        1-sample proportions test with continuity correction

data:  613 out of 1142, null probability 0.5
X-squared = 6.0324, df = 1, p-value = 0.01405
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5073437 0.5659611
sample estimates:
        p
0.5367776
```

Look carefully at the bottom of the output and you will find the 95 percent confidence interval and the value of $\hat{p}$ (which rounds to 0.5368). To find the 90% confidence interval requires changing the default as follows

```
> prop.test(613, 1142, conf.level = 0.9)

        1-sample proportions test with continuity correction

data:  613 out of 1142, null probability 0.5
X-squared = 6.0324, df = 1, p-value = 0.01405
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.5120099 0.5613682
sample estimates:
        p
0.5367776
```

[1] Actually, these numbers aren't *exactly* the same as they use **SD** instead of **SE** in the calculation

Question 13: Use a table in a book to find z^* when $\alpha = 0.15$. Compare to the result of `qnorm(1-0.15/2)`.

Question 14: A survey of size 1000 is taken and $\hat{p} = .5$ What is the margin of error for a 95% confidence level? Compare this to a survey of size 100.

Question 15: In the American press, usually only $\hat{p}$ and the margin of error are reported. What is missing from you making an accurate inference?

Question 16: Typically, in the American press sample sizes are about 1100. If the margin of error is 3 percentage points, and $\hat{p}$ is 0.5, what is the confidence level? (You might need `pnorm` instead of `qnorm()`)

Question 17: How big a survey would you need for a 95% confidence interval to have margin of error no bigger than 2 percentage points?

Question 18: In Bush vs. Gore, the general election results for Florida had Bush receiving 2,912,790 votes and Gore receiving 2,912,253 votes. What was the true value of p (based on these "official" counts)? How large would n need to be for a survey to have a 95% confidence interval of length 10^{-4} when $\hat{p} = .5$.