Sample Averages as Random Variables

Ryan Miller

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*
- The act of data collection is itself a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group

Introduction

- Lately we've been discussing **random variables**, which are used to represent the numeric outcome of a *random process*
- The act of data collection is itself a *random process*
  - We don't know which cases from the population will be sampled
  - We don't know which study participants will be randomized to the treatment/control group
- This means that *any summary measure* (means, proportions, correlations, etc.) in our sample data is the observed value of a random variable

The Sample Average as a Random Variable

- The *sample average* is a particularly useful summary measure, it can used to describe the *center* of the distribution of a quantitative variable
- For a sample of n cases from a population, the sample average is calculated:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Proportions are Averages

- Now, consider a *binary categorical* variable, we've already seen how we can express the two categories using 1 and 0 (remember how we used random variables to represent Wins/Losses last week)

Proportions are Averages

- Now, consider a *binary categorical* variable, we've already seen how we can express the two categories using 1 and 0 (remember how we used random variables to represent Wins/Losses last week)
- This means that sample proportions are also sample averages

$$\hat{p} = \frac{1+0+1+1+0+...+1}{n}$$

- Sample averages have *theoretical properties* that make them attractive random variable for statisticians to focus on

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates

The Distribution of the Sample Proportion

- According to the US Census, 27.5% of the adult population are college graduates
- Randomly sampling n adults represents a *random process*
  - The proportion of college graduates in this sample is a *random variable*
  - Let's use explore some different outcomes of this random variable for sampling protocols: random samples of size $n = 10$, and random samples of size $n = 100$

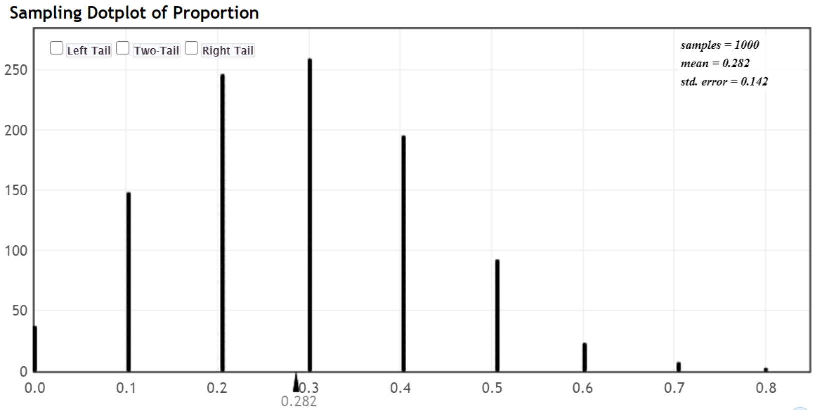
Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that might be observed
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$

Random Samples of size $n = 10$

- For a single random sample of size $n = 10$, there are exactly 11 different sample proportions that might be observed
  - Thus, the sample space is: $\{0/10, 1/10, 2/10, \ldots, 10/10\}$
- Rather than trying to perform probability calculations, we'll instead look at repeatedly drawing different random samples (of size $n = 10$) to judge the likelihood of each of these outcomes

Random Samples of size $n = 10$



- Each dot represents the proportion of college graduates in a different random sample of size $n = 10$

Random Samples of size $n = 10$

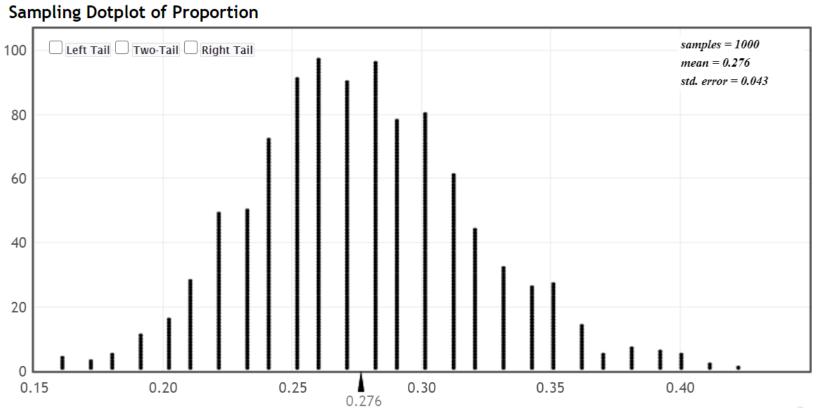
- Due to the relatively small number of discrete outcomes, it's reasonable to use a table to convey a probability model for the sample proportion:

Sample Proportion (n = 10)	Probability
0/10	$40/1000 = 0.04$
1/10	$150/1000 = 0.15$
2/10	$250/1000 = 0.25$
3/10	$270/1000 = 0.27$
4/10	$190/1000 = 0.19$
...	...
10/10	$0/1000 = 0$

Random Samples of size $n = 100$

- For a random sample of $n = 100$, there are now 101 discrete outcomes that could be observed for the sample proportion $\{0/100, 1/100, 2/100, \dots, 100/100\}$
  - It is impractical to write-out a probability for each of them, instead it makes more sense to treat the sample proportion as a *continuous random variable*

Random Samples of size $n = 100$



- Notice this distribution is roughly *bell-shaped*, it's *centered* at the population proportion (approximately), and has a spread described by the *standard error*

A Normal Model?

- You might be thinking that we can apply a Normal model here, but getting the proper Normal distribution requires us to get the center and spread correct
  - StatKey reports these values, but we'll get into where they come from in the next presentation

Plausible Values

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*) of a given summary measure
  - Sampling variability is quantified by the **standard error**, which describes the average distance of sample estimates from their expected value

Plausible Values

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*) of a given summary measure
  - Sampling variability is quantified by the **standard error**, which describes the average distance of sample estimates from their expected value
- For example, random samples of US adults of size $n = 100$ yield sample proportions that are on average 0.043 off from their expected value of 0.275

Plausible Values

- The **sampling distribution** is useful to statisticians because it expresses the *sampling variability* (sometimes called *sampling error*) of a given summary measure
    - Sampling variability is quantified by the **standard error**, which describes the average distance of sample estimates from their expected value
- For example, random samples of US adults of size $n = 100$ yield sample proportions that are on average 0.043 off from their expected value of 0.275
    - Random samples of size $n = 10$ yield sample proportions that are on average 0.142 off from their expected value of 0.275
    - This should make sense, larger samples contain more information about the population and therefore provide estimates that are more reliable (ie: tend to have less sampling error)

Conclusion

- This presentation introduced the idea of the *sample average* as a random variable
  - Proportions are averages of 0's and 1's, therefore the sample proportion is also a random variable

Conclusion

- This presentation introduced the idea of the *sample average* as a random variable
  - Proportions are averages of 0's and 1's, therefore the sample proportion is also a random variable
- The probability distribution of the sample average is called the **sampling distribution**, and it is useful in understanding *sampling variability* or *sampling error*
- *Standard error* describes the sampling variability of a particular summary measure using a specific sampling procedure
  - For example, the variability of sample proportions of college graduates in random samples of size $n = 10$