Introductory Statistics – Day 17

Sampling distributions
and
The Central Limit Theorem

Facial prototyping and reasoning with proportions

(adapted from a presentation by Dr. Allan Rossman)

Researchers investigating facial prototyping have asked whether people tend to associate names with faces for people they have not actually met. The following image is taken from a research paper by Lea, Thomas, Lamkin, & Bell (2007). Research participants were asked who is on the left, Bob or Tim?

Who is on the left, Bob or Tim? Image from Lea, Thomas, Lamkin, & Bell, 2007

In this scenario, our null hypothesis is
that there is no such thing as facial prototyping.

In statistical terms, we expect 50% of particpants to choose Tim on the left and 50% to choose Bob on the left. i.e.

$$H_0 : p = 0.5$$

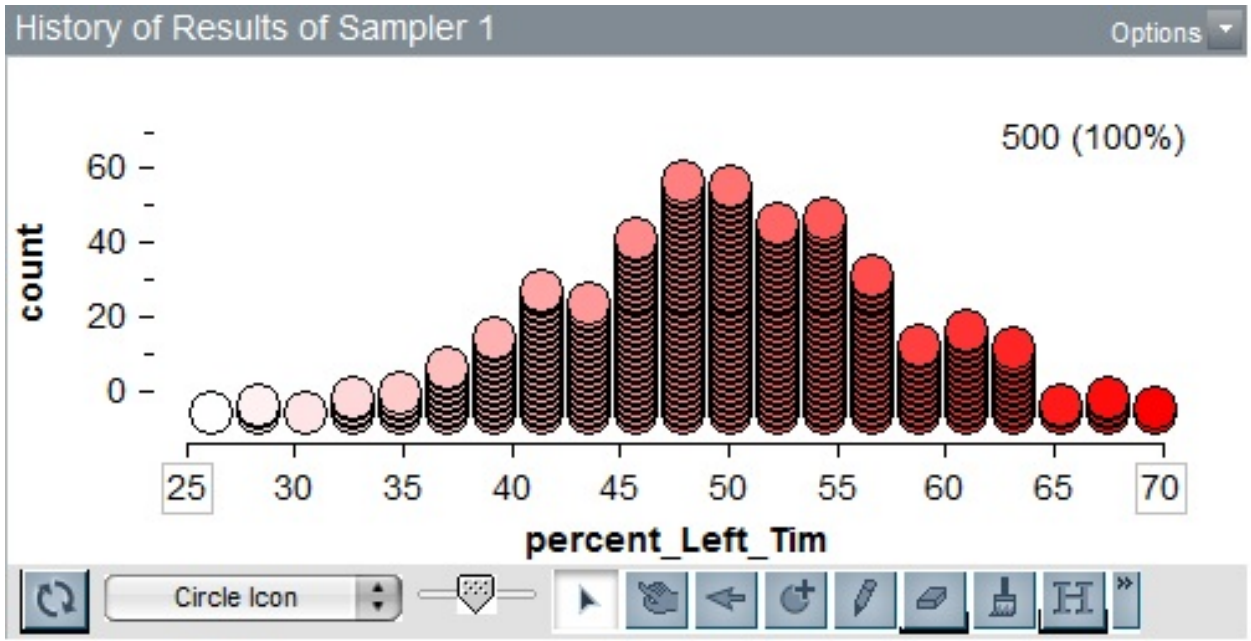
The alternate hypothesis is
that facial prototyping is a genuine phenomenon and that a noticeable majority of the participants will agree on which face is on the left. i.e.

$$H_A : p \neq 0.5$$

Example: In a class of 46 students, 36 said that Tim was on the left and only 10 said that Bob was on the left. In other words, 78% of the class thought the left face would be Tim and only 22% thought it would be Bob.

- Is this evidence of facial prototyping?
- If there is no such thing as facial prototyping, how many votes for Tim on the left would you expect?
- Create a model of the situation if there is no such thing as facial prototyping. Build a sampling distribution in TinkerPlots.
- How rare is such an extreme sample result (if facial prototyping is not happening)?

Below is a sampling distribution for the null hypothesis that there is no such thing as facial prototyping ($p = 0.5$). The sampling distribution is built with 500 randomly generated samples of size 46 with $p = 0.50$. It is not unusal to get between 40% and 60% of participants selecting Tim on the left. However, not a single sample out of 500 samples was as extreme as the 78% (or 22%) we saw in the sample data. Therefore our p-value is less than $\frac{1}{500}$, i.e. p-value < 0.002.



Sampling distribution, with spinner at p=0.5 & 500 samples of size 46.

In other words, it would be *very* unusual to get a result as extreme as 78% of people spontaneously agreeing that Tim is on the left if people do not engage in facial prototyping. This provides evidence that the researchers can use to back up their theory that facial prototyping is a real phenomenon.

Now for a bit of theory

- Simulations and sampling distributions $\Longrightarrow$ what is reasonable and what is extremely rare.
- Today: connect sampling distributions and normal distributions.
- What is the mean and standard deviation for a sampling distribution?

The standard deviation for the sampling distribution is called the standard error or SE for the sampling distribution.

The formula for the standard error will depend on the scenario.

Definition

Central Limit Theorem for Proportions

Given a population with a proportion p, the set of all possible samples of size n forms a sampling distribution that approaches a normal distribution with a

$$\text{mean of } p \text{ and}$$
$$\text{a standard error of } SE = \sqrt{\frac{p(1-p)}{n}}.$$

As the sample size, n, increases, the sampling distribution gets closer and closer to a normal distribution.

If we do not know the population proportion p, we can approximate the standard error using the sample proportion $\hat{p}$ instead of p. In this case $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

The Fine Print - (Conditions to check)

- The sample observations must be independent and generally not more than 10% of the overall population
- There should be at least 10 successes and 10 failures in our sample. i.e. If we ask a yes/no question, we should have at least 10 yes and 10 no responses in our sample. This is often written as $np \geq 10$ and $n(1-p) \geq 10$.

Activity 1: Facial prototyping, with a larger n

When researchers notice a phenomenon with a small sample size, they often conduct a follow-up study with a larger sample size. Let's suppose our facial prototyping researchers decide to conduct a follow-up study with 300 participants. They use the same Tim and Bob images as the prior activity and find that 192 participants select Tim as the photo on the left.

Rather than building a simulation, this time we will use the formulas from the Central Limit Theorem to describe the bell curve that we would have created if we had done a simulation.

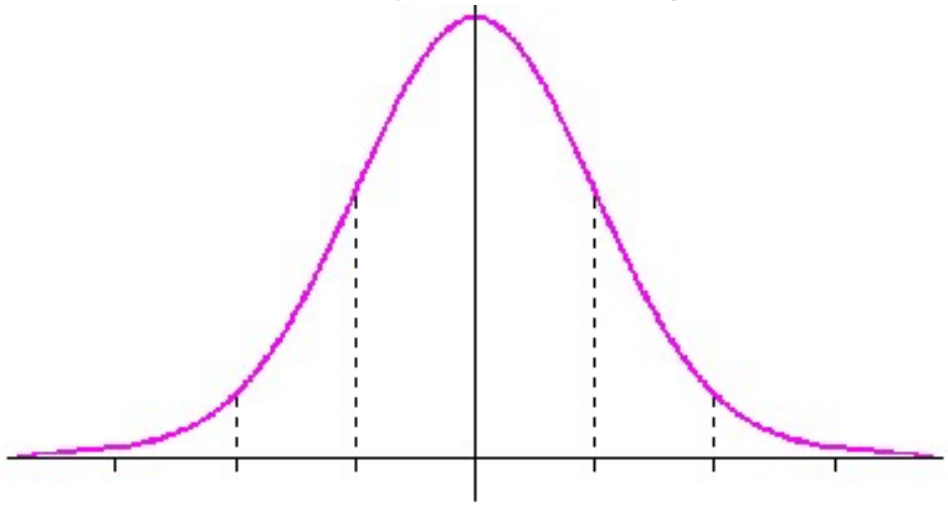
A. State the null hypothesis.

B. State the alternative hypothesis.

C. If the null hypothesis is true, what is the mean and standard error for the sampling distribution?

Activity 1, continued :

D. Draw a picture of your sampling distribution.



E. Use Excel to find the probability of a sample as *extreme* as the data above, given the mean and standard error you just found.

F. What is your conclusion?

Activity 2: Two students wanted to try this out with their own names: Heather and Jennifer. They use their own pictures and find 300 volunteer participants. They find that 162 of the 300 participants assigned the left picture to Heather and the other 138 selected Jennifer.

A. State the null and alternative hypotheses.

B. Will the mean and standard error change? If so, find the new mean and standard error. If not, explain why they will stay the same as the Bob and Tim example.

C. Draw a picture of your sampling distribution.

Activity 2, continued:

D. Use Excel to find the probability of a sample as extreme as 162 out of 300, i.e. $\hat{p} = 0.54$.

E. How does this p-value compare with the p-value you did with the Tim and Bob version of the experiment? Why might this have happened? Which of the following sound reasonable?

- Facial prototyping is actually not a real phenomenon and the researchers just got lucky with Tim and Bob.
- The phenomonon might be real, but it only works for the names Tim and Bob.
- The names Jennifer and Heather are associated with similar facial prototypes.
- Other reasons or confounding variables. What is your explanation?

F. Suppose the researchers descide to classify two names as "significantly different profile types" if testing them provides a certain p-value. What p-value would you select as the cut-off? Explain your reasoning in complete sentences.

Activity 3: For each of the scenarios, determine whether the sampling distribution will be basically a normal distribution. If so, state the mean and standard error for the sampling distribution.

According to the CDC, approximately 9.3% of people in the US have diabetes. Random samples of 500 people are taken to determine the rates of diabetes.
- Does this meet the requirements for a normal distribution? (Explain why)
If so, state the mean and standard error for the sampling distribution.

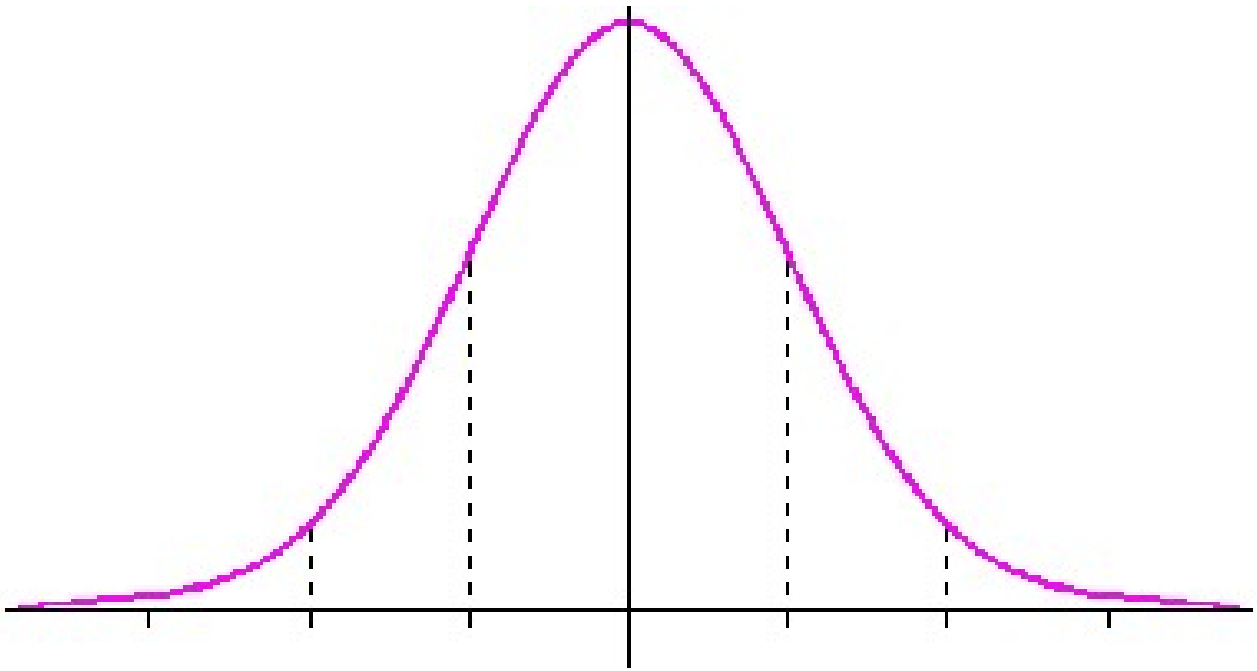
According to the ASPCA, approximately 44% of households in the US have dogs. Random samples of 80 households are taken to determine dog ownership.
- Does this meet the requirements for a normal distribution? (Explain why)
If so, state the mean and standard error for the sampling distribution.

The odds of finding a four-leaf clover are quite low. While exact figures are hard to find, one researcher claims that only 1 in 1000 clovers will be four-leaf clovers. A sample 2000 clovers from one field are selected in a search for an accurate rate of the appearance of four-leaf clovers.
- Does this meet the requirements for a normal distribution? (Explain why)
If so, state the mean and standard error for the sampling distribution.

Activity 4: Finding p-values

According to the CDC, approximately 9.3% of people in the US have diabetes. A researcher is investigating whether this rate is lower in a community of recent immigrants from Africa. She gathers a random sample of 500 recent immigrants and finds that the rate of diabetes in her sample is 8.2%.

1. State the null and alternate hypotheses.
2. Find the p-value. Use the mean and standard error you computed in the prior activity rather than creating a simulation in TinkerPlots.
3. State your conclusions as a complete sentence.



Activity 5: According to the ASPCA, approximately 44% of households in the US have dogs. A researcher wants to know if the rate of dog ownership is higher in rural areas. He takes a random samples of 80 rural households and finds that 51 had a dog.

1. State the null and alternate hypotheses.
2. Find the p-value. Use the mean and standard error you computed in the prior activity rather than creating a simulation in TinkerPlots.
3. State your conclusions as a complete sentence.