PW #6 Statistical estimation (II)

Estimation of proportions

6.1. Estimation of proportion

In today's lesson we continue our discussion about statistical estimation, learning how to estimate the proportion of a population and the associated confidence intervals. Recall that the purpose of a confidence interval is to use a sample proportion to construct an interval of values that we can be reasonably confident contains the true population proportion.

The basic idea is summarized here:

When we select a random sample from the population of interest, we expect the sample proportion to be a good estimate of the population proportion. But we also know that sample proportions vary, so we expect some error. (Remember that the error here is due to chance. It is not due to a mistake that anyone made.)

For a given sample proportion, we will not know the amount of error, so we use the standard error as an estimate for the average amount of error we expect in sample proportions. (Recall that the standard error is the expected standard deviation of sample proportions when we take many, many random samples.)

If a normal model is a good fit for the sampling distribution, then about 95% of sample proportions estimate the population proportion within 2 standard errors. We say that we are 95% confident that the following interval contains the population proportion.

$$p \pm \text{margin of error}$$
$$p \pm 2(\text{standard error})$$
$$p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

The population proportion, will be estimated to be found in the following intervals, depending on the confidence level:

- **68% Confidence interval**: [p - SE; p + SE]
- **95% Confidence interval**: [p - 2*SE; Mean + 2*SE]
- **99.7% Confidence interval**: [p - 3*SE; Mean + 3*SE]

You may realize that this formula for the confidence interval is a bit odd, since our goal in calculating the confidence interval is to estimate the population proportion p. Yet the formula requires that we know p. This is not the usual way statisticians estimate the standard error, but it

captured the main idea and allowed us to practice finding and interpreting confidence intervals. Now, we develop a different way to estimate standard error that is commonly used in statistical practice.

6.2. Example

According to a 2010 report from the American Council on Education, females make up 57% of the college population in the United States. Students in a statistics class at Tallahassee Community College want to determine the proportion of female students at TCC. They select a random sample of 135 TCC students and find that 72 are female, which is a sample proportion of 72 / 135 ≈ 0.533. So 53.3% of the students in the sample are female.

What can they conclude about the proportion of females at the college? How confident can they be in their estimate?

To answer these questions, we need to find a confidence interval.

Checking conditions:

A confidence interval comes from a normal model of the sampling distribution. Let's first make sure that a normal model is appropriate here. Recall the two conditions for using a normal model for sample proportions:

- The sample must be random.
- The expected number of successes in the sample, np, and the expected number of failures, $n(1 - p)$, are both greater than or equal to 10. In symbols, this is $np \geq 10$ and $n(1 - p) \geq 10$. Recall that *success* doesn't mean good and *failure* doesn't mean bad. A success is just what we are counting.

When we try to check these conditions, we have a problem. We do not know p, the population proportion. In fact, p is what we are trying to estimate! So we cannot determine the expected number of successes and failures. Our solution to this problem is to adjust these conditions. Advanced theory tells us that if the *actual* number of successes and failures in the sample are greater than or equal to 10, then a normal model is still a good fit.

This sample contains 72 successes (female students) and 63 failures (male students). Both are greater than 10. We therefore use the normal model for the sampling distribution.

Finding the margin of error:

We know that a sample proportion is only an estimate for the population proportion. We do not expect the sample proportion to equal the population proportion, so there is some error due to random chance. We use the standard deviation of the sample proportions to describe the amount of error we can expect in random samples. We call this the standard error.

The standard error of the sample proportion depends on the population proportion and sample size. Here is the formula for the standard error:

$$\sqrt{\frac{p(1-p)}{n}}$$

Now let's calculate the margin of error for the TCC estimate of 53.3%. Notice that we have the same problem we had earlier. We don't know p, the population proportion. So we can't calculate the margin of error! Our solution to this problem is to estimate the standard error using the sample proportion in place of p. We call this the estimated standard error, and the formula is:

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

For this example, the estimated standard error is:

$$\sqrt{\frac{0.533(1-0.533)}{135}} \approx 0.043$$

So the margin of error for the 95% confidence interval is:

$$2\sqrt{\frac{0.533(1-0.533)}{135}} \approx 2(0.043) = 0.086$$

Finding the confidence interval:

We can interpret the margin of error by saying we are 95% confident that the proportion of all students at TCC who are female is within 0.086 of our sample proportion of 0.533. We can then write the interval in the following form:

$$\hat{p} \pm \text{margin of error} = 0.533 \pm 0.086$$

When we add and subtract the margin of error from the sample proportion, the confidence interval is 0.447 to 0.619.

Conclusion:

We are 95% confident that the proportion of all TCC students who are female is between 0.447 and 0.619. We can also make this statement using percentages. We are 95% confident that the percentage of all TCC students who are female is between 44.7% and 61.9%.