Inspiring Greatness

INDEPENDENCE SCHOOL DISTRICT

Math Virtual Learning

AP Statistics

Combining sampling distributions

April 22, 2020

Lesson: April 22, 2020

Objective/Learning Target:
Students will be able to describe the sampling distribution of the difference of two proportions and two means

Review #1

How does the mean and standard deviation of the binomial distribution relate to the mean and standard error of the p-hat sampling distribution? (hint: recall linear transformations from chapter 2 and 6). Rephrased, how do we go from the equations on the left to the ones on the right, and why does that make sense?

Binomial Distribution

$$\mu_x = np$$

$$\sigma_x = \sqrt{np(1-p)}$$

p-hat sampling distribution

$$\mu_{\hat{p}} = p$$

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

Review #2

The average height of a man is 66 inches with a standard deviation of 9 inches. The average height of a woman is 62 inches with a standard deviation of 7 inches.

What is the mean and standard deviation of the difference between the heights of men and women?

Review #1 Answer

The binomial distribution records the number of successes. The sampling distribution records the sample proportion. These distributions are very closely related. If we take the number of successes in an occurrence we can easily convert it to a proportion of successes by dividing by n. This should hint at the transformation of the whole distribution. If we divide the formula for the binomial mean we get the p-hat mean. Likewise, if we divide the formula for the binomial standard deviation by n, we get the formula for the p-hat standard deviation. See page 436 for an example of this process.

Review #2 - answer

From chapter 6, recall that we can add or subtract means directly and no matter how we combine the means we always add the variances of the two distributions. We always add variances, because combining the two distributions is going to result in more possible outcomes thus more variability.

$$\mu = \mu_{male} - \mu_{female} = 66 - 62 = 4 \; inches$$

$$\sigma = \sqrt{{\sigma_{male}}^2 + {\sigma_{female}}^2} = \sqrt{9^2 + 7^2} = \sqrt{130} \approx 11.4$$

Comparing samples

We have spent a bit of time dealing with a single sample. We have estimated population parameters using confidence intervals, and tested a sample against a know population parameter. Although useful, they do not always answer interesting questions. Real world questions often require comparing two (or later on more) things.

We would like to be able to answer questions like…

-Does the short blade helicopter really drop faster than the long?

- Does medicine A result in better outcomes than medicine B?

- Is this politician really have increased poll ratings from last year? Or is that just sampling error?

Sampling distribution

We used a sampling distribution to define what “should” be happening in our one sample and one proportion procedures. Likewise, we need to define a sampling distribution for two proportions or two means. But how do we account for two samples? If given the hypotheses:

$$H_0: p_1 = p_2$$

$$H_a: p_1 < p_2$$

We might think that we are going to create two separate sampling distributions. However, the math proves to be rather complicated and unnecessary to find the p-values..

Sampling distribution

With a little math, we can rearrange our hypotheses:

$$H_0: p_1 - p_2 = 0$$

$$H_a: p_1 - p_2 < 0$$

Take a moment to convince yourself that these are equivalent. Why are these easier to prove than the previous set of hypotheses? Well they imply one distribution, not two. Furthermore, we have already covered the math to create this new distribution!

p_1-p_2 Sampling distribution

The distribution of the first proportion will be

- An approximately normal distribution (it must meet the np condition)
- The mean of the distribution will be defined by the equation:

$$\mu_{p1} = P_1$$

- The standard deviation of the distribution will be defined as:

$$\sigma_{p1} = \sqrt{\frac{P_1(1-P_1)}{n_1}}$$

The distribution of the second proportion will be

- An approximately normal distribution (it must meet the np condition)
- The mean of the distribution will be defined by the equation:

$$\mu_{p2} = P_2$$

- The standard deviation of the distribution will be defined as:

$$\sigma_{p2} = \sqrt{\frac{P_2(1-P_2)}{n_2}}$$

Note that capital P is used to denote the population proportion

p_1-p_2 Sampling distribution

Both the distributions look very similar here. They really only differ by their true value of P and their sample size. We can combine these two distributions the same way we learned to combine distributions in chapter 6!

To find the mean difference of the two distributions we simply subtract the two means:

$$\mu_{p1-p2} = \mu_{p1} - \mu_{p2} = P_1 - P_2$$

Note that if the population proportions are the same the result will be 0, if P_1 is larger it will be positive, and if P_2 is larger it will be negative.

p_1-p_2 Sampling distribution

Now the standard deviation. For simplicity let's first calculate the variance.

$$\sigma_{p1-p2}{}^2 = \sqrt{\frac{P_1\,(1-P_1)}{n_1}}^2 + \sqrt{\frac{P_2(1-P_2)}{n_2}}^2 = \frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$$

We squared each standard deviation equation and summed the resulting values. This is the equation for the variance of the P_1-P_2 sampling distribution. We then square root to find the standard deviation.

$$\sigma_{p1-p2} = \sqrt{\frac{P_1(1-P_{1)}}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

p_1-p_2 Sampling distribution

We can now describe the resulting sampling distribution:

Center: It will have a mean p-hat of:

$$\mu_{p1-p2} = \mu_{p1} - \mu_{p2} = P_1 - P_2$$

Spread: The standard deviation of the distribution is:

$$\sigma_{p1-p2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

Normal: We know that when we combine two normal distributions the resulting distribution is also normal!

Sampling distribution for the difference of two means

If we want to test to see if there is a difference between two means, we might have a hypothesis like:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

Again, we are looking at the difference of the two, instead of a direct inequality. This allows us to consider a single distribution instead of multiple interacting distributions.

Sampling distribution for the difference of two means

Each distribution will be similar in shape, center, and spread and can be defined in the follow way for both x_1 and x_2.

Center: $\mu = \mu_{x-bar}$

Spread: $\sigma_{x-bar} = \frac{\sigma}{\sqrt{n}}$

Normality: Each will meet a condition of the central limit theorem

Sampling distribution for the difference of two means

Therefore,

Center: $\mu_{x_1 - x_2} = \mu_{x_1} - \mu_{x_2}$

Spread: variance - $\sigma_{x_1 - x_2}^2 = \left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}$

Standard deviation - $\sigma_{x_1 - x_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Normal: Since both sampling distributions are normal the combination of the two are also normal!

Implications

We can now describe a distribution for the difference of two sample means, or the difference of two sample proportions. This allows us to make predictions about the difference of these statistics!

Once we have these distributions, we can conduct inference in much the same way as we did for a single mean or proportion.

You try!

We take a sample of 40 brown bears from mainland Alaska and find the mean weight to be approximately 858 pounds with a standard deviation of 100 pounds. A random sample of 38 brown bears from kodiak island is has a mean of 955 pounds with a standard deviation of 55 pounds. We want to know the difference in weights between mainland and kodiak bears. Describe the sampling distribution of interest.

Answer

Shape: Both samples have sizes greater than 30. Therefore, when we combine the two distributions, we still have a normal distribution.

Center: $\mu_{x_1 - x_2} = \mu_{x_1} - \mu_{x_2} = 955 - 858 = 97$

Spread: $\sigma_{x_1 - x_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} = \sqrt{\frac{55^2}{38} + \frac{100^2}{40}} = 18.155$

Try again!

According to a CNN poll of 1024 randomly selected U.S. adults on September 2nd, 2010, 50% approved of Obama's job performance. A CNN poll of 1010 randomly selected U.S. adults U.S. adults on August 30th, 2009, showed that 53% percent approved of Obama's job performance. Describe the distribution of the difference between the two approval ratings.

Answer

This is dealing with proportions, so we will use the appropriate methods.

Shape: both distributions meet the np and n(1-p) conditions for normality, so the resulting distribution is also normal

Center: $\mu_{p1-p2} = \mu_{p1} - \mu_{p2} = P_1 - P_2 = 0.53 - 0.5 = 0.03$

Spread: $\sigma_{p1-p2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} = \sqrt{\frac{0.53(1-0.53)}{1010} + \frac{0.5(1-0.5)}{1024}} \approx 0.0222$

Notice that the center is within 2 standard deviations of zero. What does that suggest?

Conclusion

Now that we can describe the sampling distribution of the difference between two proportions or the difference between two means, we have an idea about how these statistics should be working.

In the next two lessons, we are going to start using these distributions to develop a method of making inferences about two proportions and two means.