Madeline Kuehn
Math 634

Reflection #6: Statistical inference using hypothesis testing: Inference about means and proportions with two populations.

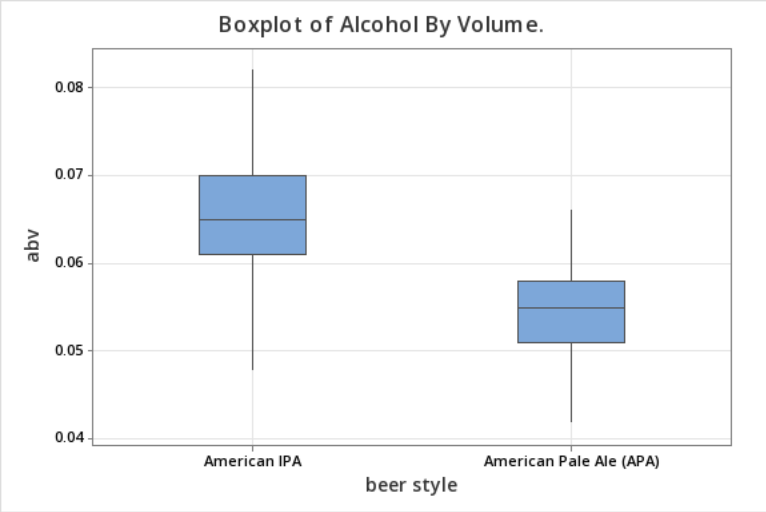
Part 1: Statistical inference using hypothesis about two population means

Different types of beers have different ingredients, flavors, and alcohol amounts. Two popular beer styles in the US are the American Pale Ale and the American IPA. The IPAs tend to have a stronger flavor and come in a variety of colors whereas the Pale Ales tend to be lighter in flavor and in color.

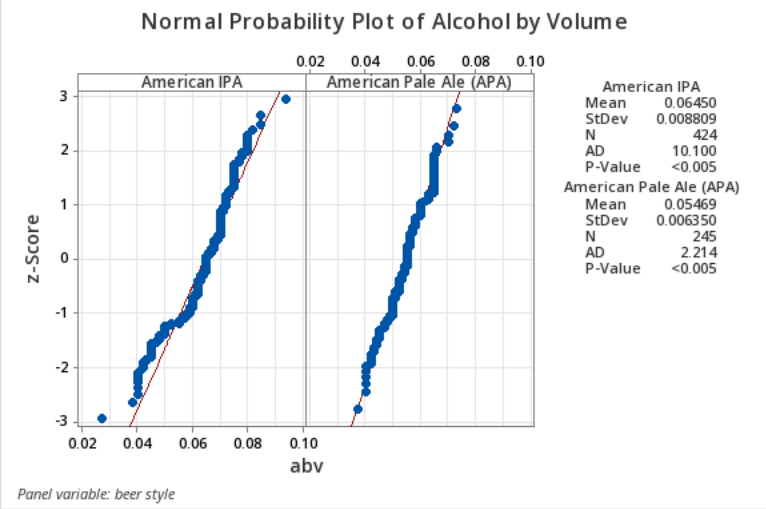
Using a statistical package, test to determine if the mean alcohol by volume for the American IPA is the same as the mean alcohol by volume for the American Pale Ale.

1. Graph the alcohol volumes for the two types of beer styles using appropriate graphs and calculate statistics appropriate for this type of data.

   The first graph that I plotted was a box plot comparing the ABV or AmIPA to APA. This allows us to look at the difference in the IQR to see if the variances are approximately equal.



   I also graphed the normal probability plot, and it does not show any severe departure from normality.



Madeline Kuehn
Math 634

Population 1:	American IPA	Population 2:	American Pale Ale
Sample Size	$n_1 = 424$	Sample Size	$n_2 = 245$
Sample Mean	$\overline{x_1} = 0.0645$	Sample Mean	$\overline{x_2} = 0.0547$
Sample Variance	$s_1^2 = 0.0000776$	Sample Variance	$s_2^2 = 0.0000403$
Sample Std. Deviation	$s_1 = 0.00881$	Sample Std. Deviation	$s_2 = 0.00635$
IQR	0.009	IQR	0.007

2. Conduct the appropriate hypothesis test using the following steps.
   a. Determine the null and alternative hypotheses.
   Null: H_0: $\mu_1 - \mu_2 = 0$. There is no difference in the mean ABV for American IPA and American Pale Ale.
   Alternative: H_a: $\mu_1 - \mu_2 \neq 0$. There is a difference in the mean ABV for the two groups.

   b. Use a significance level of α = 0.05.

   c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value. Using the graphs created, determine if you should be conducting a two-sample test of the mean with equal or unequal variances.

   Assumptions:
   1. The samples are independent of one another. APA beers are independent of AmIPA beers.
   2. Both samples are simple random samples.
   3. Our populations are both normally distributed and the samples are both large ($424 > 30, 245 > 30$).
   4. Both populations have approximately equal, but unknown std. deviations. This is due to the fact that our box plot shows a very similar spread using the IQR 0.009 vs. 0.007.

   First, we need to find the pooled standard deviation, s_p.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$
$$= \sqrt{\frac{(423)(0.0000776) + (244)(0.0000403)}{667}}$$
$$\approx 0.007997$$

   The standard error is $s_{\overline{x_1} - \overline{x_2}}$.

Madeline Kuehn
Math 634

$$s_{\overline{x_1}-\overline{x_2}} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$
$$= 0.007997\sqrt{\frac{1}{424}+\frac{1}{245}}$$
$$\approx .0006418$$

So, the observable test statistic, t_0 is:

$$t_0 = \frac{\overline{x_1}-\overline{x_2}}{s_{\overline{x_1}-\overline{x_2}}}$$
$$= \frac{.0645-.0547}{.0006418}$$
$$\approx 15.27$$

d. Determine the P-value.

Because our alternative hypothesis is not equal, we know we are doing a 2-tailed t-distribution with 667 degrees of freedom. Using the excel formula, $= T.DIST.2T(15.27, 667)$ gives us a P-value of $2.28806 \times 10^{-45} \approx 0$.

e. Make a decision to reject or fail to reject the null hypothesis, H_0.

Because our p-value is less than $\alpha = 0.05$, we reject the null hypothesis, H_0.

f. State the conclusion in terms of the original question.

There is sufficient evidence to conclude that there is a difference in the average Alcohol by Volume amounts for American IPA is different from the average Alcohol by Volume amounts for American Pale Ale.

3. Calculate the 95% confidence interval for the difference between the two means. Does this confidence interval support your results from the hypothesis test? Why or why not?

The formula for the confidence interval is $(\overline{x_1}-\overline{x_2}) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$.

Computing $t_{\alpha/2}$with $n_1+n_2-2 = 424+245-2 = 667$ degrees of freedom is found by using the t-distribution. We know that we are finding $\frac{\alpha}{2} = \frac{0.05}{2} = .025$. Using Excel, $= T.INV(0.025, 667)$ gives us a value of 1.964 since we are in the upper tail.

The pooled sample variance is:

$$s_p^2 = (0.007997)^2 \approx 0.00006397$$

So,

$$(\overline{x_1}-\overline{x_2}) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

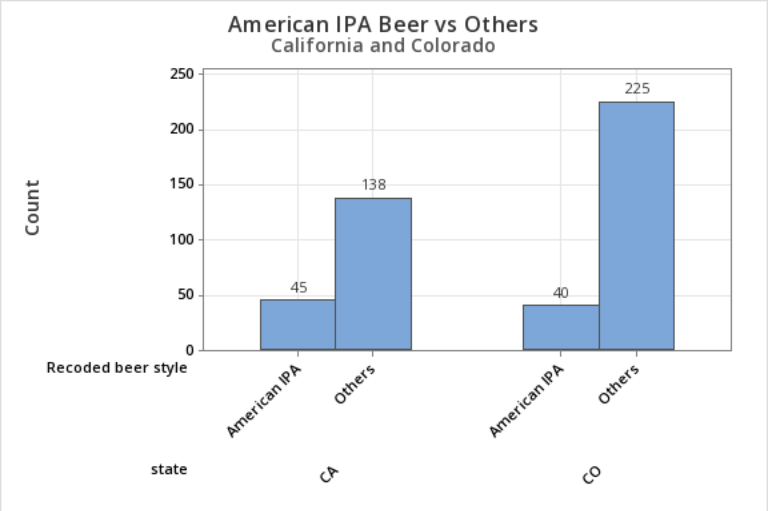
Madeline Kuehn
Math 634

$$= (.0645 - .0547) \pm 1.964\sqrt{0.00006397\left(\frac{1}{424} + \frac{1}{245}\right)}$$
$$= .0098 \pm .00126$$

The 95% Confidence interval goes from 0.00854 to 0.01106. This interval does support our results from the hypothesis test that on average, the American IPA have between 0.00854 to 0.01106 more alcohol by volume than American Pale Ale.

Part 2: Statistical inference using hypothesis about two population proportions
Colorado and California are huge producers of beers with many microbreweries in each state. Both states produce a variety of different types of beers (Beer styles) as well. Is the proportion of American IPA's compared to all other types of beers (Beer styles) the same in both California and Colorado? (Note: You are comparing American IPA versus others)

1. Graph the American IPAs as compared to all other types of beer styles for each state (California and Colorado) using appropriate graphs and calculate statistics appropriate for this type of data.



Population 1	California	Population 2	Colorado
Number of American IPAs	$x_1 = 45$	Number of American IPAs	$x_2 = 40$
Sample Size	$n_1 = 183$	Sample Size	$n_2 = 265$
Sample Proportion of American IPAs	$\widehat{p_1} = \frac{45}{183} \approx 0.246$	Sample Proportion of American IPAs	$\widehat{p_2} = \frac{40}{265} \approx 0.151$

Madeline Kuehn
Math 634

2. Conduct the appropriate hypothesis test using the following steps.

a. Determine the null and alternative hypotheses.

Null: H_0: $p_1 - p_2 = 0$. There is no difference in the proportion of production of American IPA vs. other beers in California compared to Colorado.

Alternative: H_a: $\mu_1 - \mu_2 \neq 0$. There is a difference in the proportion of production of American IPA vs. other beers in California compared to Colorado.

b. Use a significance level of α = 0.05.

c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value.

Assumptions:

1. The sample proportion are from two simple random samples, looking at just California vs. Colorado
2. The samples are independent.
3. The samples are large enough such that there are at least 10 successes and 10 failures for each sample:
    a. $n_1\widehat{p_1} = 183(0.246) = 45.018 \geq 10$
    b. $n_1(1 - \widehat{p_1}) = 183(.754) = 137.982 \geq 10$
    c. $n_2\widehat{p_2} = 265(.151) = 40.015 \geq 10$
    d. $n_2(1 - \widehat{p_2}) = 265(.849) = 224.985 \geq 10$

First, we need to find the pooled sample proportion, $\bar{p}$.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{45 + 40}{183 + 265} = \frac{85}{448} \approx 0.1897$$

So, our observed test statistic, z_0 is:

$$\begin{aligned} z_0 &= \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.246 - 0.151}{\sqrt{0.1897(.8103)\left(\frac{1}{183} + \frac{1}{265}\right)}} \\ &= \frac{.095}{\sqrt{0.0014200}} \\ &\approx 2.520 \end{aligned}$$

d. Determine the P-value.

Madeline Kuehn
Math 634

We need to use the z-distribution to find the P-value. Because our alternative hypothesis is not equal, we know we are doing a 2-tailed distribution.

So, p-value $= 2\big(1 - P(z > |z_0|)\big) = 2\big(1 - P(z < -2.520)\big) = 2(0.005872) \approx 0.0117$

e. Make a decision to reject or fail to reject the null hypothesis, H_0.
Because our p-value is less than $\alpha = 0.05$, we reject the null hypothesis, H_0.

f. State the conclusion in terms of the original question.

There is sufficient evidence to conclude that there is a difference in the production of American IPA vs. other beers in California compared to Colorado.

3. Calculate the 95% confidence interval for the difference between the two proportions. Does this confidence interval support your results from the hypothesis test? Why or why not?

The formula for the confidence interval is $(\widehat{p_1} - \widehat{p_2}) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$.
Computing $z_{\alpha/2}$. First, we find $\frac{\alpha}{2} = \frac{0.05}{2} = .025$. Using Excel, $= NORM.S.INV(0.025)$ gives us a value of 1.960 since we are in the upper tail.

So,

$$(\widehat{p_1} - \widehat{p_2}) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

$$= (0.246 - 0.151) \pm 1.960\sqrt{\frac{0.246(1-0.246)}{183} + \frac{0.151(1-0.151)}{265}}$$

$$= 0.095 \pm 0.07584$$

The 95% Confidence interval goes from 0.0191 to 0.1708. This interval does support our results from the hypothesis test that we are 95% confident that the true difference in production of American IPA vs other beers between California and Colorado is between 0.0191 to 0.1708. Because it doesn't include 0, it supports rejecting the null hypothesis.