Introductory Statistics – Day 28

Difference of Means

At this point in the course, the basic structure of a hypothesis test should be well engrained. Once you have your research question, you engage in something like the following plan:

1. Look at your data (if you start with raw data rather than summary statistics). Create plots. TinkerPlots is helpful for this. What do you notice? Is there anything odd or suprising about the data? Are there missing or bizarre data points?
2. Write appropriate null and alternative hypotheses.
3. Make a list of the descriptive statistics that you would need for this test.
4. Look up the appropriate standard error formula for your test.
5. If you are using the t distribution, look up the formula for the degrees of freedom.
6. Run the test, find a p-value, and state your conclusion.
7. If you reject your null hypothesis, create a confidence interval (usually 95%) for your parameter of interest.

So far we have conducted hypothesis tests for one proportion and two proportion problems, as well as one mean and paired means problems. In this section we investigate scenarios where we compare two means (not paired). To do this, we will need formulas for standard error and degrees of freedom.

Confidence Interval:

- The sample means are: $\bar{x}_1$ and $\bar{x}_2$
- We are interested in: $\bar{x}_1 - \bar{x}_2$
- Degrees of Freedom: $df = \min\{n_1 - 1, n_2 - 1\}$
- Standard Error: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Confidence Interval: $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times SE$

Hypothesis Test:

- Null Hypothesis: $H_0 : (\mu_1 - \mu_2) = 0$
- Alternate Hypothesis: $H_A : (\mu_1 - \mu_2) \neq, <, or > 0$
- Standard Error: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Test Statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$

Listening for Parkinson's

Parkinson's is a disease of the nervous system which affects muscle control, movement, and speech. In a study at the University of Oxford, Dr. Max Little investigated the vocal patterns of research participants with and without Parkinson's disease. He measured attributes including vocal frequency (in Hz), jitter (variations in frequency), and shimmer (variations in amplitude). A subset of this data is included in the file ParkinsonsSpeech (data set from UCI Machine Learning Repository).

Source: Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

- Open up the data set in Excel and copy the columns into TinkerPlots. Explore the data. What do you notice about frequency, jitter, and shimmer for the Parkinson's and non-Parkinson's patients?
- From your graphs, you might notice that vocal jitter looks somewhat higher on the Parkinson's group than the non-Parkinson's group. Conduct a hypothesis test to determine if the Parkinson's group has a higher average level of jitter in speech. Note: This is a one tailed test because of our prior knowledge of Parkinson's as a disease that affects muscle control in speech.
- Conduct a hypothesis test to determine if the Parkinson's group has a higher average level of vocal shimmer.
- Both vocal jitter and shimmer are statistically higher in the Parkinson's group, so it's appropriate to follow-up with a 95% confidence interval about the difference between the Parkinson's group and the non-Parkinson's group. State your confidence intervals as complete sentences related to the context.
- Do individuals with Parkinson's disease always have higher levels of vocal jitter and shimmer than individuals without Parkinson's disease?

Download the VideoGamesXBoxPS3 data set from Moodle.

Is there a difference in average sales per video game between the XBox360 and PS3 platforms? If you were a game designer, would one platform tend to lead to more profits than the other? We're going to explore this question for different geographic areas. For each of the following, conduct a hypothesis test using $\alpha = 0.05$. Note, the columns for sales are measured in millions of dollars.

1. Is there a difference in averages sales per video game between XBox360 and PS3 for the Global Sales column?
2. Is there a difference in averages sales per video game between XBox360 and PS3 for just North America (NA sales column)?
3. Is there a difference in averages sales per video game between XBox360 and PS3 for just Japan (JP sales column)?
4. Looking at your answers for the previous two questions, what is surprising about your results?
5. Concept check: Globally, XBox 360 games had a noticeably higher standard deviation than PS3 games. What does that mean in practical terms for video game sales?

Recall the basic process for a hypothesis test:

1. Explore the data visually.
2. Write appropriate null and alternative hypotheses
3. Make a list of the descriptive statistics that you would need for this test. Organize these carefully in your Excel sheet.
4. Look up the appropriate standard error formula for your test
5. If you are using the t distribution, look up the formula for the degrees of freedom.
6. Run the test, find a p-value, and state your conclusion. Assume that $\alpha = 0.05$.
7. If you reject your null hypothesis, follow-up with a 95% confidence interval.