Chapter 6

Inferential Methods for a Single Numerical Variable

Section 6.1: Comparing Dependent Samples, i.e. Test on Differences

**Example 6.1** (Example 2.5, Rosner) A common symptom of otitis media (ear infection) in young children is the prolonged presence of fluid in the middle ear. The hypothesis has been proposed that babies who are breast-fed for at least 1 month may build up some immunity against the effects of the disease. A small study of 24 pairs of babies is set up, where the babies are matched on a one-to-one basis according to age, sex, socioeconomic status, and type of medications taken. One member of the matched pair is a breast-fed baby and the other was bottle-fed.

Research Question: Is there a (statistically) significant difference in the *average* duration length of ear infection between the breast-fed and bottle-fed babies?

|  |  |
| --- | --- |
| Chapter 2: Identified which didbetter | Chapter 6: Measuring the amount of difference |
|  |  |

When we analyzed this data in Chapter 2, we used the binomial distribution which allowed us to have only two outcomes. With this analysis, we simply identified which baby did better (i.e. the breast-fed or bottle-fed) for each pair.

Here, we will measure the amount of difference for each pair and use this in our analysis. In particular, we can consider the size of the difference which is useful because a difference of -2 is VERY different than a difference of -24.

Summary Statistics for the Differences…





Does the observed data tend to support or refute the research question of differences existing? Discuss.

Identifying and removing the outliers…

* Boxplot



Descriptive statistics without the 3 significant outliers…



Discuss any differences here from what was observed with all 24 observations.

Recall, the research question, “Is there a (statistically) significant difference in the *average* duration length of ear infection between the breast-fed and bottle-fed babies?” The average for our data (without the outliers) is -6.3. This value will certainly change over repeated samples and to answer this question, we would like to know the average stays away from 0 (i.e. the no difference situation).



Tinkerplots Investigation

Getting the appropriate data for taking repeated samples.

|  |  |
| --- | --- |
| Original data with Average Difference = -6.3 | Adjusted data with Average Difference = 0Note: Adjusted Data = Original Data + 6.3333 |

Setting up “spinner” or hat in Tinkerplots using adjusted data (to simulate no difference)



Taking 100 repeated samples produces the following graph. The statistic of interest here is the average; thus, on each sample the average difference is obtained.



Research Question: Is there a (statistically) significant difference in the *average* duration length of ear infection between the breast-fed and bottle-fed babies?

The following graph would be used to obtain the p-value (for a two-side test).



The number of dots that exceed -6.3 or positive 6.3 would be included in the calculation of the p-value. Here the p-value would be given by 6/100 = 0.06.

Formal Statistical Test in Excel

Research Question: Is there a (statistical) significant difference in the average of ear infection between the breast-fed and bottle-fed babies?

Completing the test in Excel…



Determine p-value and make the statistical decision.



The Decision Rule: If the p-value is less than the error rate, then the data is said to support the research question.

P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistical Decision: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Conclusion – Writing a final statement in laymen’s terms

**Confidence Interval**

Concept of 95% confidence interval…



Wiki Entry for Confidence Interval for Mean



Identify the parts of this interval and obtain the endpoints of the 95% confidence interval.

* Lower Endpoint:
* Upper Endpoint:

Sketch this interval on the line below.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Interpret the meaning of this 95% confidence interval.

Getting this done in Excel



Example 6.1.2 For this example, consider the Vitamin Intake dataset on our course website. This purpose of this investigation is to compare the actual intake of calories (and certain vitamins/minerals) against their daily recommended intake.

|  |  |
| --- | --- |
| Data : Difference for Energy = Actual - DRI | CDF Plot |
|  |

 

Answer the following research question using the data from the course website.

Research Question: Is there enough statistical evidence to suggest, on average, there is a difference between the actual energy intake against their daily recommended intake?

The Decision Rule: If the p-value is less than the error rate, then the data is said to support the research question.

P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Statistical Decision:

* Data provides enough evidence for research question
* Data does not provide enough evidence for research question

Conclusion – Writing a final statement in laymen’s terms

Construct an approximate 95% confidence interval using the following formula.

$$Lower Endpoint: Average Difference-\left(2\*\frac{Standard Deviation}{\sqrt{Count}}\right)$$

$$Upper Endpoint: Average Difference+\left(2\*\frac{Standard Deviation}{\sqrt{Count}}\right)$$

Provide, in context, an interpretation of this 95% confidence interval.

Section 6.2 Comparing a Single Numerical Variable Across Two Groups

Let us continue with the previous example, but now the desire is to compare the difference in calorie intake between females and males.

Research Question: Do a statistical difference exist between genders in the deficiency of caloric intake?

|  |  |
| --- | --- |
| Data in Spreadsheet | The Two Buckets Concept |

Summary statistics by gender





Conceptual Understanding of Repeated Sampling with Two Buckets



Conceptual Understanding of Repeated Sampling under the No Difference Situation





Looking at difference in average over repeated sampling (under the situation of no difference)



Getting the p-value in Excel…



|  |  |
| --- | --- |
| Data | Using the T-Test Function to get the p-value |

Research Question: Is there enough statistical evidence to suggest, on average, there is a difference between genders in their actual energy intake against their daily recommended intake?

Obtain P-value:

P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Decision:

The Decision Rule: If the p-value is less than the error rate, then the data is said to support the research question.

* Data provides enough evidence for the research question
* Data does not provide enough evidence for the research question

Conclusion: Write a conclusion using everyday language for this research question.

Using a 95% Confidence Interval to measure the amount of difference between the two averages.



JMP software output for 95% Confidence Interval



Interpret the 95% confidence interval using everyday language.



Next, consider an evaluation between genders of their actual intake of calcium against their daily recommended intake.

Research Question: Is there enough statistical evidence to suggest, on average, there is a difference between genders in their actual calcium intake against their daily recommended intake?



Obtain P-value:

P-value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Decision:

The Decision Rule: If the p-value is less than the error rate, then the data is said to support the research question.

* Data provides enough evidence for the research question
* Data does not provide enough evidence for the research question

Conclusion: Write a conclusion using everyday language for this research question.

The following output provides the 95% confidence interval for this analysis.



Interpret this interval in context.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
AverageF – AverageM