Introduction to Data Analytics 1

Ze Wang

2019-07-30

# Part 1: Variables, Hypothesis, Designs

*Title:* Offshore outsourcing: Its advantages, disadvantages, and effect on the American economy

*Abstract*: The United States has trained some of the world’s best computer programmers and technology experts. Despite all of this training, many businesses do not have a full understanding of information technology. As the importance of technology in the business world grows, many companies are wasting money on extensive technology projects. When problems arise, they expect that further investment will solve these issues. To prevent such problems, many companies have begun to outsource these functions in an effort to reduce costs and improve performance. The majority of these outsourced information technology and call center jobs are going to low-wage countries, such as India and China where English-speaking college graduates are being hired at substantially lower wages. The purpose of this study is to evaluate the positive and negative aspects of offshore outsourcing with a focus on the outsourcing markets in India and China, arguably the two most popular destinations for outsourcers. The cost savings associated with offshore outsourcing will be evaluated in relation to the security risks and other weakness of offshore outsourcing. In addition, an analysis of the number of jobs sent overseas versus the number of jobs created in the United States will be used to assess the effects that outsourcing is having on the American economy and job market. Finally, the value of jobs lost from the American economy will be compared to the value of jobs created. The goal of these analyses is to create a clear picture of this increasingly popular business strategy.

Answer the following questions about the abstract above:

1. What is a potential hypothesis of the researchers?

One of the potential hypotheses could be that the outsourcing strategy adopted by many American companies lead to higher unemployment rate in the information technology industry in the US.

1. What is one of the independent variables?

One of the independent variables could be the numbers of jobs in information technology industry being sent overseas from the US every year.

* 1. What type of variable is the independent variable?

It is a discrete variable.

1. What is one of the dependent variables?

One of the dependent variables could be the unemployment rate of informational technology industry in the US

* 1. What type of variable is the dependent variable?

It is a discrete variable.

1. What might cause some measurement error in this experiment?

Measurement error might come from the unclear definition of the information technology industry. Sometimes it is hard to precisely determine which company is in information technology field and which is not, especially when the company apply information technology as one of their means to achieve goals in other industry such as healthcare.

Also, it is hard to get a very accurate number when measuring the unemployment rate. As we know, lots of people are freelancer programmers in the information technology industry and their employment status will also be hard to define, thus affecting the real number of the unemployment rate.

1. What type of research design is the experiment?

It will be correlational research design

* 1. Why?

[Correlational research](https://www.questionpro.com/blog/correlational-research/) is a non-experimental research design which helps researchers to establish a relationship between two closely connected variables.

1. How might you measure the reliability of your dependent variable?

One of the methods would be using test-retest method, which can estimate a measurements consistency by evaluate the result after repeat the test

1. Is this study ecologically valid?

Yes. It is ecologically valid because all the data are from real world and the result can be applied to real-word settings.

1. Can this study claim cause and effect?
   1. Why/why not?

No. This study can be claimed cause and effect only when we can identify the close relationship between the two variables, the time-order of the two variables and also rule out other possibilities.

1. What type of data collection did the researchers use (please note that #5 is a different question)?

It is administrative data.

# Part 2: Use the assessment scores dataset (03\_lab.csv) to answer these questions.

The provided dataset includes the following information created to match the abstract:

* Jobs: the percent of outsourced jobs for a call center.
* Cost: one calculation of the cost savings for the business.
* Cost2: a separate way to calculate cost savings for the business.
* Business: an ID number for each business.
* Where: where the jobs were outsourced to.

Calculate the following information:

1. Create a frequency table of the percent of outsourced jobs.
2. Create histograms of the two types of cost savings. You will want to add the breaks argument to the hist() function. This argument adds more bars to the histogram, which makes it easier to answer the following questions:

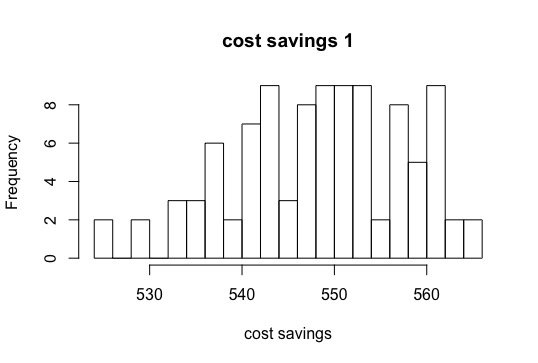
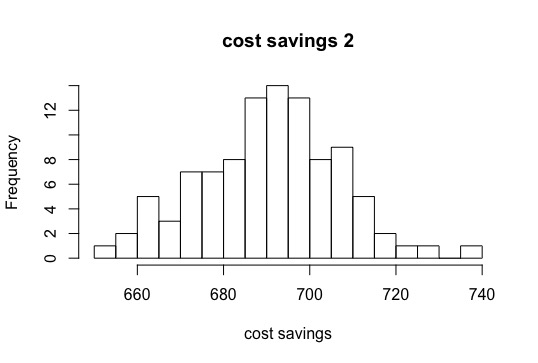
hist(dataset$column, breaks = 15)

15 is a great number to pick, but it can be any number. For this assignment, try 15 to see a medium number of bars.

Below is the code I used to make the table and histogram:

> *table(jobsdata$jobs)*

*> hist(jobsdata$cost, main = "cost savings 1", xlab = "cost savings", breaks = 15)*

*> hist(jobsdata$cost2, main = "cost savings 2", xlab = "cost savings", breaks = 15)*

1. Examine these histograms to answer the following questions:
   1. Which cost savings appears the most normal?

Based on the histogram above, histogram of cost saving 2 apparently seems to be most normal.

* 1. Which cost savings data is multimodal?

Based on the histogram above, histogram of cost saving 1 apparently is multimodal.

* 1. Which cost savings data looks the most skewed (and in which direction positive or negative)?

Based on the histogram above, histogram of cost saving 1 is skewed to the left.

* 1. Which cost savings data looks the most kurtotic?

Based on the histogram above, histogram of cost saving 2 is more kurtotic.

1. Calculate the z-scores for each cost savings, so they are all on the same scale.

I used following code to calculated the z-score for each cost savings.

*> scale(jobsdata$cost)*

*> scale(jobsdata$cost2)*

1. How many of the cost saving scores were more extreme than 95% of the data (i.e., number of z-scores at a *p* < .05)?

Following is the data I used to get the number cost savings z-scores that are more extreme

than 95% of the data

*> jobsdata <- jobsdata %>%*

*+ mutate(z\_score = (cost - mean(cost)) / sd(cost))*

*> View(jobsdata)*

*> jobsdata <- jobsdata %>%*

*+ mutate(z\_score2 = (cost2 - mean(cost2)) / sd(cost2))*

*> View(jobsdata)*

*> filter(jobsdata, z\_score > 1.645)*

# A tibble: 2 x 7

jobs cost cost2 business where z\_score z\_score2

<dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>

1 80.5 566 688 50 India 1.85 -0.198

2 79.2 565 685 77 China 1.74 -0.383

*> filter(jobsdata, z\_score2 > 1.645)*

*# A tibble: 5 x 7*

jobs cost cost2 business where z\_score z\_score2

<dbl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>

1 75.1 555 718 1 China 0.661 1.65

2 87.0 538 740 20 India -1.18 3.01

3 78.4 550 730 73 China 0.119 2.39

4 79.5 547 722 76 China -0.206 1.90

5 81.7 543 718 94 China -0.639 1.65

* 1. Cost Savings 1: 2 scores are more extreme than 95% of the data
  2. Cost Savings 2: 5 scores are more extreme than 95% of the data

1. Which business had:

Following is the code I applied to fing the ID number of the business with highest cost saving, lowest cost saving.

*>which.max(jobsdata$cost)*

*>which.max(jobsdata$cost2)*

*> which.min(jobsdata$cost)*

*> which.min(jobsdata$cost2)*

* 1. the highest cost savings?

No. 50 business has the highest cost 1 saving.

No. 20 business has the highest cost 2 saving.

* 1. the the lowest cost savings?

No. 67 business has the lowest cost 1 saving.

No. 5 business has the lowest cost 2 saving.

* 1. Use both cost savings columns and find the ID number of the business with the lowest and highest z-score.

The result is the same from the previous 2 questions.

No. 50 business has the highest cost 1 z-score

No. 20 business has the highest cost 2 z-score

No. 67 business has the lowest cost 1 z-score

No. 5 business has the lowest cost 2 z-score