DATA SCIENCE My Class work in Google Classroom 2/21/22

**LESSON 6**

**Why do you need to take a sample?**

 We need samples to narrow down large data (data sets)  by choosing a subset of the population. For example, big data, is to large and not cost-effective to sample, therefore identifying and analyzing a representative sample narrowing it into a subset is sufficient to collect data. There are however pros and cons to sampling.

***The number of people in your sample is called the sample size***

**Describe the most reliable form of sampling.**

Simple random sampling saves time and resources, it is reliable because every single member is chosen randomly and by chance.

**Explain convenience sampling and its reliability**

Convenience sampling uses respondents who are often convenient to the researcher. This is not reliable because of its lack of representation of the population.

Often forming a bias.

**Explain the difference between a null and an alternative hypothesis.**

In statistical hypothesis testing the null hypothesis of a test always predicts no effect or no relationship between variables, while the alternative states research predictions of an effect or relationship.

**Explain the difference between a Type I and a Type II error**

A Type I error (False-Positive) occurs if an investigator rejects a null hypothesis that is actually true in the population. (It happens when the null is True but you reject it anyway)   A Type II error (False-Negative/error of omission) occurs if the investigator fails to reject the null hypothesis that is actually false in the population.

<https://www.google.com/search?q=Explain+the+difference+between+a+type+1+and+a+Type+2+error&oq=Explain+the+difference+between+a+type+1+and+a+Type+2+error&aqs=chrome..69i57j33i160j33i299l2.30400j1j15&sourceid=chrome&ie=UTF-8>

**LESSON 7**

Why might you want to recode data to be categorical and binomial?

So the machine can learn from the data and the model to which it should apply. Categorical variables must have the numbers encoded before the data model is concluded.

What is one of the properties of a t-distribution

The means of a distribution is equal to 0.

Give an example of pre- and post-test paired data

Pre-test (Before treatment)  data will have a controlled group and another group. Which will be tested two different times. And a Post-test(after treatment)  paired group the same but afterward. And controlled groups may be a placebo treatment( no treatment at all)

<https://www.real-statistics.com/analysis-of-covariance-ancova/pretest-posttest-design/>

Explain effect size, and why you might use it instead of a p-value. 

A P-value is not enough therefore a p-value tells us that an intervention works, where as the effect size explains how much it works.

<https://www.google.com/search?q=Explain+effect+size%2C+and+why+you+might+use+it+instead+of+a+p-value&oq=Explain+effect+size%2C+and+why+you+might+use+it+instead+of+a+p-value&aqs=chrome..69i57.21868j0j15&sourceid=chrome&ie=UTF-8>

<https://www.latimes.com/business/technology/la-diversity-right-thing-snap-story.html>

CHAPTER 6 QUIZES

***STASTICAL INTERFERENCE***

***2. Population and Samples***

1. A Marketing firm in Austin Texas is researching a local auto parts chain based in Dallas. They have identified 72,500 households within a 10 miles of any of their stores. They send out surveys to 2,200 homes in the Dallas area. Of those 2,200 surveys the marketing firm gets some responses from 168 of them.

***What is the sample in this example?***

B. It is the 168 households that responded to the survey.

*Since the sample is the smaller part and only the portion you have data for, the 168 households are the sample.*

A Marketing firm in Austin Texas is researching a local auto parts chain based in Dallas. They have identified 72,500 households within a 10 miles of any of their stores. They send out surveys to 2,200 homes in the Dallas area. Of those 2,200 surveys the marketing firm gets some responses from 168 of them.

***What is the population on this example?***

C. All 72,500 households in the Dallas area?

***STASTICAL INTERFERENCE***

***3.SIMPLE RANDOM SAMPLING***

1. A Simple Random Sample means that every person/object has an equal chance of being selected.

1. TRUE

***STASTICAL INTERFERENCE***

***4. CLUSTER SAMPLING***

1. A Financial Auditor has to audit 6,000 accounts receivables. The accounts cover 12 years’ worth of business, which comes down to about 10 accounts per week on average. So the financial auditor decides to randomly choose 5 different weeks during those 12 years and review all the registered accounts for each of those weeks, totaling 57 accounts.

What type of sampling is this?

Cluster Sampling

1. The data is broken down into chunks, or clusters, or weeks. So by choosing everything in certain weeks, that is the cluster sampling.

2. What is the difference between cluster and simple random sampling?

B. Cluster sampling randomly selects groups simple random sampling randomly selects individuals.

***STASTICAL INTERFERENCE***

***5.Stratified Sampling***

An insurance company is reviewing policies to see if they can recommend a better coverage option for their clients. They have hundreds of thousands of policyholders and don't have time to review them all. One employee suggests they categorize the policyholders into the following groups: Auto only, auto and home, motorcycle only, and recreational vehicles only. From each of these groups, they randomly select a small portion of the total policyholders in that category. What type of sampling does this example show?

Graphical user interface, text, application

Description automatically generated

SYStematic Sampling

A financial auditor has 6,000 accounts receivable that need auditing. She wants to audit 50 of the accounts, so she calculates that 50 / 6,000 is 120. She randomly selects a starting point of 85 and then selects every 120th account after that. Thus, her total sample size is 50 accounts. What type of sampling is this?

Graphical user interface, text, application, email

Description automatically generated

Convience Sampling

A financial auditor has to audit 6,000 accounts receivable. She has all of the account files on a single sub-directory on her computer, so she sorts them from smallest file to largest file. Experience tells her that the smaller files are usually less messy. She then takes the 50 smallest files and audits those because she has better things to do than this audit. What type of sampling is this?

Graphical user interface, text, application, email

Description automatically generated

8. Choosing a Sample Method

Graphical user interface, text, application, email

Description automatically generated

9. Sampling Revisted

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

***10. Hypothesis Testing***

1. Which hypothesis always contains the “=” sign

B. The Null Hypothesis

2. Which o the following is NOT a form of the alternative hypothesis can take?

C. The same as (“The same as” would be a form of the null, not the alternative hypothesis

3. Ha is the alternative of the symbol for the alternative hypothesis?

A. TRUE

4. What is p value?

C. It is the probability of having a greater than or less than the test statistics

5. What does Alpha level .01 mean?

D. Both A and B (There is a 99% chance of accuracy and there is a 1% chance of error)

.05 95% chance of accuracy 5% of error

.10 90% chance of accuracy 10% of error

6. If the p-value is less then alpha what should you do?

D. Reject the Null Hypothesis

***STASTICAL INTERFERENCE***

***11. TYPE I and TYPE II ERORR***

1.What is a Type I error?

D. It happens when the null is True, but you reject it anyway

2. What is a Type II error?

C. It happens when the null is False, but you fail to reject it anyway.

***STASTICAL INTERFERENCE***

***12. COURT ROOM ANALOGY***

1. If the defendant is guilty but the jury has found him not guilty,

B. that is a Type II Error

(It happens when the null is False, but you fail to reject it anyway)

1. If a defendant is not guilty but the jury finds him guilty is this an example of Type I or Type II error
2. Type I

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

BASIC SAMPLING REVIEW WORKSHOP VIDEO

https://vimeo.com/410440398