Hypothesizing My Way to a Somewhat Destination

Posted by Andrew Wong on March 9, 2019

This is a record of my Flatiron School Data Science Bootcamp - Module Two Final Project Journey. It is most about hypothesis testing (hence the title!), and lots of statistics (I read a few very thick statistic books, some are really interesting).

I had a few false starts. First, I jumped into too quickly to run SQL query. Stupid mistake, I thought I can wing it. Second, I start without a good grasp of statistical understanding - I am still unclear about null and alternative hypothesis. Basic rookie mistake. Third, I don't have a map to guide me. As a result, I am lost for a few days.

So, lesson learnt indeed for me, and here's some improvement that I have made.

FIRST

I re-visited some of my past work on exploratory data analyysis, and I remembered (I should have!) on some framework that can guide me. Here's the whole lot of it.

THE DATA SCIENCE FRAMEWORK AND THE SCIENTIFIC METHOD
There are a few popular exploratory data analysis frameworks such as:

Harvard Data Science Workflow: This workflow centers on business questions to drive data analysis. By asking powerful questions, this model drive better data context.

KDD Workflow: The Knowledge Discovery in Databases (KDD): workflow centers on complext data discovery, especially in legacy data warehouse systems.

CRISP-DM Workflow: The Cross-Industry Standard Process for Data Mining (CRISP-DM) workflow centers on iterative data analysis. It is based on a simple approach of bringing analytics to production in a business-oriented and systematic way.

Data Journalism Workflow: This is a very interesting workflow because it centers around storytelling, and sharing of information through visual, and always thinks about the readers context.

OSEMN Workflow: This workflow centers on a rapid, iterative data discovery, scrubbing, exploration, modelling, and interpretation.

I have used the OSEMN workflow in the last module. And, I have decided to use the Harvard Data Science Workflow this time around. Why you may ask?

This research project is heavy on hypothesis testing, therefore it will be beneficial to start asking interesting questions (e.g. what is the scientic goal?, what do you want to predict or estimate (in my context here, what initial hypothesis questions that I can think of?). This is the FIRST PHASE of the process.

This research project is based on a well-known relational SQL database. The data sample is ready for me to access and connect. This is the SECOND PHASE of the process.

The bulk of my time will be to understand the relationship between different tables in the relational database, and drawing connections between tables to find what are the data in respective tables, and the relationship between them. This is the THIRD PHASE of the process.

After a clear understanding of what data I can use, and relationships between tables, I will start to form insightful hypothesis (NOTE: My definition of insightful hypothesis means provoking Aha moments). After forming null and alternative hypothesis, I will start to query data through SQL. This will be an iterative exercise as I practice honing my skill in SQL to model the data. Once this is completed, I will start to run statistical analysis such as t-test, welch's t-test, and cohen'd (and perhaps some other statistical tests). This is the FOURTH PHASE of the process. (NOTE: See Diagram 2: The Scientic Method for more description about hypothesis testing)

After all the effort, I hope to present business-centric insights on Northwind Traders. This is the FIFTH and LAST PHASE of the process.

SECOND

I have started to review the Northwind Traders database schema, and start to draw out my own interpretation. By drawing and seeing the linkages between different tables, I feel more confident and closer to the data.

THIRD

Start to ask some high-level questions.

EMPLOYEE

Who work for Northwind - their demographic, psychographic, employement profile? Which regions and territories with employees with highest (and lowest) order transactions? SALES

Who are Northwind's customer profiles? What are the customers product categories preference? Which regions or territories have the highest (and lowest) growth? SUPPLIERS OF NORTHWIND

Who are the suppliers - their business relationship profile with Northwind? What are the relationship between Northwind orders, products, and product categories, and their suppliers?

FOURTH

Get a good understanding of basic statistical concepts. For this particular module, I read a fair bit of hypothesis testing. Here's an excerpt of it:

Constructing a Testable Hypothesis STATISTIC REFRESHER

The following paragraphs taken from the book (as a reminder to myself, and good reference for the reader of this research project): Experimental Design and Analysis by Howard J. Seltman (edition: July 11, 2018; accessed on the Internet on 23rd February 2019).

Any hypothesis must allow for different possible conclusions or it is pointless. For an exploratory goal, the different possible conclusions may be only vaguely specified. In contrast, much of statistical theory focuses on a specific, so-called "null hypothesis" (e.g., reaction time is not affected by background noise) which often represents "nothing interesting going on" usually in terms of some effect being exactly equal to zero, as opposed to a more general, "alternative hypothesis" (e.g., reaction time changes as the level of background noise changes), which encompasses any amount of change other than zero.

Statistical analysis of experiments starts with graphical and non-graphical exploratory data analysis (EDA). EDA is useful for:

1. Detection of mistakes
2. Checking of assumptions
3. Determining relationships among the explanatory variables
4. Assessing the direction and rough size of relationships between explanatory and outcome variables, and
5. Preliminary selection of appropriate models of the relationship between an outcome variable and one or more explanatory variables.
6. EDA always precedes formal (confirmatory) data analysis.

FIFTH

Here's an excerpt of the first hypothesis that I have written with statistical analysis.

HYPOTHESIS 1 - DETERMINING SIGNIFICANT OF DISCOUNTED PRODUCTS ON QUANTITY ORDER

H0 - There is no statistically significant difference in the quantity ordered of discounted vs. non-discounted products. H0: $\mu 1 = \mu 2$

H1 - There is a statistically significant difference in the quantity ordered of discounted vs. non-discounted products. HA: $\mu 1 \neq \mu 2$

This hypothesis is interesting for Northwind Trader because based on statistical analysis (that's hard, solid data!), they can determine whether discount given to customers will eventual yield greater revenue or not.

INTERPRETING HYPOTHESIS 1 T-TEST

In performing a hypothesis test in statistics, a p-value determine the significance of results. If the p-value is a number between 0 and 1 and interpreted in the following way: A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

The Hypothesis 1 T-Test below indicates a high p-value. Therefore we ACCEPT ALTERNATIVE HYPOTHESIS that there is a statistically significant difference in the quantity ordered of discounted vs. non-discounted products. The results are statistically significant at alpha = 0.05

stats.ttestind(discount.Quantity.values, full.Quantity.values)
TtestindResult(statistic=6.4785631962949015, pvalue=1.1440924523215966e-10)

WELCH'S T-TEST STATISTIC REFRESHER

In statistics, Welch's t-test, or unequal variances t-test, is a two-sample location test which is used to test the hypothesis that two populations have equal means. Welch's t-test is an adaptation of Student's t-test, and is more reliable when the two samples have unequal variances and/or unequal sample sizes.

The Hypothesis 1 Welch's T-Test below indicates a high p-value. Therefore we ACCEPT ALTERNATIVE HYPOTHESIS that there is a statistically significant difference in the quantity ordered of discounted vs. non-discounted products.

stats.ttestind(discount.Quantity.values, full.Quantity.values, equalvar=False)
TtestindResult(statistic=6.239069142123973, pvalue=5.65641429030433e-10)

EFFECT SIZE (COHEN'S d) STATISTIC REFRESHER

Cohen's d is an effect size used to indicate the standardised difference between two means. It can be used, for example, to accompany reporting of t-test and ANOVA results. It is also widely used in meta-analysis. Cohen's d is an appropriate effect size for the comparison between two means.

Cohen suggested that: d=0.2 be considered a 'small' effect size d=0.5 represents a 'medium' effect size d=0.8 a 'large' effect size.

This means that if two groups' means don't differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically signficant.

The Hypothesis 1 Effect Size (Cohen's d) below indicates a slightly than 0.2 d-value. Therefore, we can conclude with confidence (not high confidence) that the effect size is slightly more than a 'small' effect size.

← PREVIOUS POST (/A_NOTE_ON_WRITING_ON_JUPYTER_NOTEBOOK)

NEXT POST → (/FROM_WOBBLY_DATA_SCIENCE_PROJECT_TO_EFFICIENT_DATAOPS)


(https://learn.co/andrewwongls)


(https://github.com/andrewwongls)