Mahoney, Mike

DA 6823

Kilger

Module 1: Part #1 (50 points)

**The Power of Statistics + the Levels of Measurement + the Different Classes of Variables and Determining Appropriate Statistical Technique + Basic Descriptive Measures**

**General Instructions:** In your own words, answer each of the following questions - don’t copy (e.g. cut and paste) some definition out of a book word for word. This is not a group project – you are expected to complete this module on your own. You may refer to text books, online or other sources but not your fellow classmates. If you don’t understand the question, feel free to ask the instructor in class, in office hours or in an email.

1. Provide a short definition for dependent variable. (3 points)

The dependent variable is the target variable, or the variable you are trying to predict or understand. The outcome of the DV is dependent on the values represented in the independent variable. It would be the “Y” in an equation such as Y = mX + b

1. Provide a short definition for independent variable. (3 points)

The independent variable is the predictor variable(s) and is a variable used to predict the dependent variable or target variable. The value of the independent variable would not change with a change of the dependent variable (assuming multiple IV’s or Error/Bias). The IV would be “X” in the equation Y= mX + b

1. Provide a short definition for control variable. (3 points)

A control variable is a constant and would not change as a result in a change of the IV or DV. As you change one of your predictor variables, you would hold the others constant, to ensure the change in the Target variable is indeed as a result in the change of the one predictor (IV).

1. Be able to describe the simple criteria for each of the four levels of measurement:
   1. Nominal (2 points) – The lowest level of measurement is categorical data, that cannot be ranked. Usually text characters or numbers that represent a name attribute, but in no particular order. Statistical information that is usually obtained is the mode, counts and frequency.
   2. Ordinal (2 points) – The next level is still categorical but can be ranked or put in order, but the distance between the categories does not have to be equal. The classes are now distinct and now the mean, minimum, maximum and range can be measured (but the mean should not be).
   3. Interval (2 points)- The classes are distinct and have equal distances between categories ( or values). A difference of 10 values is the same any where on the scale.
   4. Ratio (2 points) – The highest level of measurement, in which anything could be measured from the data including standard deviations.
2. Provide an example of a variable for each of the four measurement levels below.
   1. Nominal (2 points) – Categorical data, or attributes. Location of UTSA: San Antonio, Height of Mike: 5’9,.
   2. Ordinal (2 points) – Military ranks: Lieutenant, Captain, Major. They can be put in order but there is no way to quantify the distinction between them
   3. Interval (be careful – be sure it is interval and not ratio!) (2 points) – Temperature is the best example (in Fahrenheit). A 10 degree difference is the same, such as 90-100 or 55-65, and does not change.
   4. Ratio (2 points) – Weight can be added, subtracted, the standard deviation and men (time series). I could run several statistical tests on a weight (or a sample)
3. Name at least two criteria from the IDRE chart that are used in determining which statistical technique can be used in a situation. (3 points)

Two criteria include: 1) How many dependent variables >> multivariate or univariate; 2) Which level of measurement is the dependent variable >> Nominal, Ordinal, Interval, Ratio.

1. Briefly explain the difference between descriptive and inferential statistics. (4 points)

Descriptive statistics >> describes basic information about the data, it doesn’t infer any new information, just conveys what is already present in the data. Such as 99% of all animals are dangerous in Australia. The highly inflated statistic does not predict anything, just describes data points based on what I deem dangerous.

Inferential Statistics >> We are able to predict and infer new information from the data we have. If all the information I have says that Jelly fish stings are painful, I can predict(or infer) getting stung by a Box Jellyfish == A Bad time.

1. Almost every statistical technique you will come across has some sort of assumptions – even non-parametric statistics.
   1. Name one benefit that assumptions of a test provides you (2 points)

It gives you confidence that a condition has been met such as homogeneity of variance >> Lavine’s test gives me confidence that I can use my ANOVA model.

* 1. Name one cost that assumptions of a test carry (2 points)

The increase of type 1 errors increases the more test that are run (when testing a hypothesis). Therefore a correction is needed like Bonferroni (that reduces the significance level) to offset the cost of extra tests.

1. What happens if you violate the assumptions of a statistical test? Do the statistical police come and arrest you? (4 points) - No, because you can’t get arrested for an ethical violation. You should try and find more data, possibly examine your data more to find an explanation and even report the violation; because it is the Ethical action to take as a data scientist.
2. Using the IDRE chart, suggest the appropriate statistical test for each of the following business cases
   1. As a maker of colored contact lenses, you think that there may be relationship between the color of the contact lenses purchased and the gender of the purchaser. (2 points)

Chi Squared goodness of fit >> gender: IV, Nominal ; color: DV, Nominal >> both are categorical.

* 1. As an auctioneer of fine art, you think that there may be a different between the price paid for a piece of art between men and women. (2 points)

Two independent T test >> gender: IV, Nominal; price: DV, Ratio >> Do T test for both and compare the means.

* 1. You want to better understand how different versions and price mixes of your product – the Vegematic – have on the number of product sold. You hypothesize that color of product, price, region of the country (North, South, East, West), gender of purchaser, household income of purchaser have an effect on the number of pieces sold. You may also want to make some predictions about how many products would be sold under various levels of these variables. (2 points)

Factorial ANOVA>> Num sold: DV, Ratio; color,price,region,etc: IV >> It is DV, with multiple IVs (and their interactions).

* 1. As publisher of the popular magazine Rabbit Times, you think that there may be a relationship between the number of pages in the magazine and the number of copies of that issue sold. How do you find out the direction and how strong this relationship might be? (2 points)

Correlation >> Pages in mag: IV, Interval; Num sold: DV, Interval >> In a correlation plot you are plotting the relationship between IV and DV; A upward trend is postivie, and downward is negative correlation.

* 1. You are the maker of FelineHair – a hair growing drug for hairless cats. You want to test your drug against three other drugs to see which one grows the most hairs on the cats in the experiment. You also want to see if there are other differences in the effectiveness depending upon the gender of the cat and what color coat the cat has. You end up with a drug (4) x cat gender (2) by cat coat color (black, white, brown) experimental design. What analysis technique would you use for this experiment? (2 points)

I think it is a Mixed ANOVA >> Drug: DV, Categorical; Cat Gender (with ingroups), IV and Color (with ingroups) >> In this one the IV’s would all be tested with in group, then against the other group.