Review Questions 1

M.Z.

7/25/2020

Overview

These are the objectives of chapter one. I have decided to find the answer to each objective, so as to be able to get closer on the fundamental topics presented.

Question 1: coin flip process + interpretation

Refer to python code: https://github.com/dasvidanja/Statistics/blob/master/Statistics%20Fun/Simple%20Hypothesis%20Refined.py

Question 2: Define Probability

'The probability of something happening is the proportion of time that it is expected to happen when the same process is repeated over and over (paraphrasing from Freedman,et al.,Statistics, 2nd ed., Norton, 1991, 1st ed. 1978)'

Question 3: Define p-value

If we examine the results of the chance model simulations in this way, the probability of seeing a result as extreme as the observed value is called the p-value(or probability value). Even if our chance model had produced a very low probability, ruling out chance, this does not necessarily mean that the real cause is the alternative that we are considering. There are many other possible explanations. Just as we need to rule out chance, we need to rule out those as well.

This explanation gives a general overview of hypothesis testing and the p-value's relativity to it. 'A standard approach exists for answering the question "is chance responsible?" This approach is called a hypothesis test. To conduct one, we first build a plausible mathematical model of what we mean by chance in the situation at hand. Then, we use that model to estimate how likely it is, just by chance, to get a result as impressive as our actual result. If we find that an impressive improvement like the observed outcome would be very unlikely to happen by chance, we are inclined to reject chance as the explanation. If our observed result seems quite possible according to our chance model, we conclude that chanceis a reasonable explanation.'

Question 4: Key statistics in EDA

- Central Location:
  - Mean (a.k.a average): x bar is used for sample, mu used for population
  - Median: midlle number of sorted data
  - Mode: most common observation
  - Expected Value: theoretical mean of a RV
  - Percents: proportion (part/whole)* 100
- Variation:
  - Range: max-min , sensitive to outliers
  - Percentiles: 'The Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100−P) percent of the values take on this value or more.'
  - IQR: 75TH percentile - 25th percentile
  - Residuals: deviation from some typical value. Usually, difference between mean value and an observed one.
  - Variance: mean of the squared residuals, sigma for population (population vs sample variance difference formulas N vs N-1)
  - SD: square root of variance, same scale as the data & mathematically convenient
- Distance:
  - Euclidian Distance: distance between two or more points sqrt((a-b)^2+…)
- Data Check:
  - Outliers: value that does not fit the data
  - Frequency table: set of values(x axis) and their frequency (y axsi). Histogram is the graphical reppresentation
  - Boxplot: enclose central half of the data. Outliers are outside the whisker (1.5*IQR)
  - Errors: 'fat finger' phenomena which is simply refering to a typo in the data

Question 5: Data Format

Question 5: Data Format

- Database format: in which each row reppresents an observation and each column reppresent a feature regarding that particular observation
- Relational Database: data that is contained in a table format in which information is releated to each other through a series of ‘key’ connection (additional fields in the columns). These keys releate different tables (data) back to the same entity
- Flat file: tabular format made of rows and columns

Question 6: ‘Big Data’ vs traditional stats research data

- Big Data: for v’s
  - Volume: quantity/amount
  - Velocity: flow rate (high)
  - Variety: different types of data (dates, numbers,…)
  - Veracity: no control checks, different quality

-“The great scale of the flow of new data means that the challenge of extracting, manipulating, cleaning, and preparing data is now enormous, and the time spent doing that easily outweighs the time spent analyzing data. The level of programming expertise required for these steps is substantial. Having gone to great lengths to prepare the data, adding some statistical algorithms into the process to gain interesting knowledge seems like a modest step to the programmer.”

- In short, now days statistical analysis is part of a bigger process that requires the new statistician (data science) to be confortable to with data pre/processing & manipulation.

Question 7: Treatment & Control Groups in experiments

Experimental vs Observational

- observational: preexisting data an study
- experimental: collecting data to answer a prespecified question (experiment,prospective study)
- Experimental Design,
  - Assign observations to two groups: Control & Treatment group. Control will be the benchmark to which you will compare the results yielded by your new method used in the treatement group. Ideally, you want the two groups to be similar to each other so that any difference that might result in the end can be associated with the different treatemnet instead of other contributing factors that might compromise our findings/results.

    So, how do we ensure that we account for the other factors, or better said how do ensure that other factors are not interfering with our results? RANDOMIZATION is the best strategy, by randomly assigning individuals to the control/treatemnt group we are ensuring that the variations that come with each particualr individul are random, and not enough to influence our results.

    “True random assignment eliminates both conscious and unconscious bias in the assignmentto groups. It does not guarantee that the groups will be equal in all respects. However, it does guarantee that any departure from equality will be due simply to the chance allocation and that the larger the number of samples, the fewer differences the groups will have. With extremely large samples, differences due to chance virtually disappear and you are left with differences that are real- provided the assignment to groups is really random”

    Bottom line: randomization attempts to make different groups similar by allowing for randomness in both, and offsetting itself by its presence in both systems.
  - Plan ahead: what will the experiment be alike, what are the resources needes. Talking to exper statisticians is key.
  - Blinding: whenever working with humans there is a particularity that takes place: Hawthorne Effect- where individuals experience/exhibit positive effects merely by being subject of studies/attention (lol, people and their fargile impartiality). Hence, as researchers/ statisticians we need a shield against such reality. One, way is to provide a ‘dummy treatment’ called ‘placebo.’ The subject is not aware of this reality, and might indeed report positive results (placebo effect), with the difference this time that the people in charge (stats, reasearchers) are aware of the palcebo, and can indeed account for in in the developments. Not disclosing information with the subject is called single blinding, double blinding- not discolsing info with stufff, and triple- not discolsing info with people who evaluate results (not common).
  - Paired data, one strategy to compare the groups is to take two measurements for each group (say one year apart) record the differences for each group (year1-year2), and focus on the study of the differences. Check if the differences in the treatement group are the result of chance (having control group model as benchmark), or if there is indeed a difference that can’t be simply fruit of randomness (treatment works).