Notes on Statistics

Shasha Liao
Georgia Tech

December 20, 2020

Notes from Youtube Channel *StatQuest with Josh Starmer*

1 Explaining Concepts

- Population Variance: the average of the squared differences between the data and the population mean μ.
- Model: an approximation of the real data. We use models to explore relationships and we use statistics to determine how useful and how reliable our model is.
- Sampling a Distribution: use computer to pick a random number based on the probability described by the histogram or the curve. We do this to explore statistics.
- Hypothesis Testing: We create a hypothesis. If data give us strong evidence that the hypothesis is wrong, then we can reject the hypothesis. But if we have data that is similar to the hypothesis, but not exactly the same, then the best thing we can do is fail to reject the hypothesis.
  - Null hypothesis: the hypothesis that there is no difference between things (requires no preliminary data to make the statement).
  - Alternative hypothesis: the opposite of the null hypothesis when there are only two groups. If there are more than two groups, we will have several options for alternative hypothesis and using different alternative hypothesis may end up with different decisions about rejecting the null hypothesis or not.
- p-values: the probability of observing the given data given the null hypothesis is true. It takes values between 0 and 1, quantifying how confident we should be that the null hypothesis is true.
  - If we have a small p-value (less than a significance level α), we can reject the null hypothesis. But a small p-value dose not mean that the difference between two things in the null hypothesis is large. It only tells us the probability of the observing the

current data given the null hypothesis is true and helps us to decide whether we want to reject the null hypothesis or not.

- Type I and Type II error
  - Type I error: when the null hypothesis is true, but we rejected it. It corresponds to obtaining a false positive.
  - Type II error: when the null hypothesis is false, but we failed to reject it. It corresponds to having a false negative.

- Significance level α
  - Typically we choose $\alpha = 5\%$.
  - It means we are wiling to get a false positive conclusion 5 times out of 100.
  - Lowering the α value (say to 1%) will decrease the probability of making a false positive conclusion(type I error).

2 Notes from the course Descriptive Statistics in Udacity

2.1 Lesson 1: Intro to Research Methods

- construct: (e.g. effort, health, happiness...) there are many ways to define a construct
- operational definition: a way of turning constructs into variables we can measure
- lurking variables: can influence the relationships we measure between variables and should be controlled in an experiment
- population, population mean μ
- sample, sample mean $\bar{x}$, sampling error
- we estimate population parameters using sample statistics
- Show relationships/correlation $\Rightarrow$ Observational studies & Surveys
- Show causation $\Rightarrow$ Controlled experiment
- Survey: Cheap and easy, but might get untruthful responses, biased responses, respondents not understanding the questions (response bias), or respondent refusing to answer (nonresponse bias). Usually used to analyze constructs.
- Controlled experiment: blinding(participants), double blinding experiment(both participant and researchers), random assignment(equal chance and independent, represent the population better)

2.2 Lesson 4: Visualizing Data

- Frequency table, Relative Frequency(proportion, percentage), Interval/bin
- Histogram(bins/frequencies): x-axis is qualitative, adjust bin size(length of intervals) based on how much details you want
- Bar graph: x-axis is categorical, no bin size
- Biased graphs: need to check the values of the y-axis
- Histogram of Normal distribution: roughly symmetric, most data fall in the middle of the distribution
- Skewed Distribution: asymmetrical, skewed with the most data falling toward the left or right of the distribution

2.3 Lesson 8: Central Tendency

- Mode: the most common value in the distribution; for histogram, the mode is the range of the highest frequency; uniform distribution has no mode; multi-mode distribution;
- Median: sort the data; the median is the value in the middle; median is not influenced by outliers; median is more robust;
- Mean/Average: the mean can be described as a formula; all scores in the distribution affect the mean; many samples from the same population will have similar means; the mean of a sample can be used to make inferences about the population it came from; the mean will change if we add an extreme value to the dataset; outliers lead to a misleading average;
- Mode, Median, Mean are the three measures of center of the data; Median is better for skewed data;
- Mean
  - Best used as a measure of center if our data is approximately symmetric and dose not contain outliers
  - There is a simple formula to compute mean
  - The mean is very sensitive to outliers. It will always get pulled in the direction of the largest outliers.
- Median
  - Best used as a measure of center when outliers are presented in the data since median will not be affected by extremely small or large observations

  - The median is the data point where 50% of the observations are above and below that datapoint. To find the median, we first sort the dataset and consider cases when the dataset has an odd or even number of observations; if odd, then the median is the number in the middle; if even, then the median is the average of the two numbers in the middle.

- Mode
  - Best used as a measure of center when analyzing categorical datasets. The mode is the number, range of numbers, or category that occurs the most frequently
  - The mode is also very resistant to outliers since it relies on which observation occurs the most and not the actual value of the observation

2.4 Lesson 11: Variability

- Two distributions can have the same mean, median, and mode, but have different variances which measure how spread out a distribution is.
- We don't use range as a measure of the variability of our data because it is not robust enough. If we add an outlier, the range can be changed a lot.
-
  - Q_1: the first quartile
  - Q_2: median
  - Q_3: the third quartile
  - Interquartile Range(IQR): Q_3 - Q_1
- Def: a data x is considered an outlier if $x < Q_1 - 1.5(IQR)$ or $x > Q_3 + 1.5(IQR)$.
- Boxplots: Min, Q_1, Q_2, Q_3, max, and outliers. The length of the box in the middle reflects how spread the data set is.
- IQR dose not depend on every data.
- Variance: a number that takes account of all the data to measure the variability of the dataset. It is calculated as the average/mean squared deviation.
- Standard Deviation:
  - The most common measure of spread.
  - Calculated as the square root of the variance to make the unit 1D.
  - For normal distributions, 68% of data falls within 1 standard deviation of the mean; 95% of data falls within 2 standard deviations of the mean.

- Bessel's correction: In general, samples underestimate the amount of variability in a population because samples tend to be values in the middle of the population. Instead of dividing the sum of squared deviations by n, we divide it by $n-1$ to calculate the sample standard deviation to approximate the bigger population standard deviation.

2.5 Lesson 13: Standardizing

- The proportion of data values that are less than or greater than a certain value in the data set tells us how good or how bad something is.

- Absolute frequency and relative frequency (absolute frequency divided by n)

- The distribution curve allows us to calculate the proportion of data points between any two values on the x-axis.

- In a normal distribution, given a value in the x axis, we can calculate $z =$ the number of standard deviation that value is away from the mean. Then we can estimate the percent of data less than or greater than that value.

- Describe a value: 3.5 deviations below/above the mean.

- z score $= \frac{x-\mu}{\sigma}$ is the number of standard deviation any value is away from the mean.

- Converting every value to a z score in a normal distribution is the process of standardizing. Every number in the data set is written in terms of the number of standard deviations it is from the mean.

- A negative z score means that the original value is less than the mean.

- The standard normal distribution has mean 0 and standard deviation 1.

- Given the z score of a value and the standard deviation of the original distribution, we can use the formula of z score to convert it into the real world value x. In this way, we can convert a value in one normal distribution to another normal distribution and make it easy two compare two values in two different normal distributions.

2.6 Lesson 16: Normal Distribution

- PDF: probability density function. The area under the curve represents the probability. We can use calculus or the $Z-$table (for normal distribution) to find the area under the curve. The $Z-$table provides a quick way to find the probability of getting anything less than a given z-score.

2.7 Lesson 18: Sampling Distributions

- Sampling distribution is the distribution of the sample means.
- Standard Error: the standard deviation of the sampling distribution.
- Central Limit Theorem: given any distribution, if we keep drawing samples from it and record the sample means, then the distribution of sample means is approximately normal, the standard deviation of the sample means is approximately $\frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation and n is the sample size, and the mean of sample means is approximately the population mean.
- If the sample size is 1, the sample means will be the data values, and the distribution will not be normal but follows the population distribution.
- The larger the sample size n, the smaller the standard error, and the skinnier the sample means distribution. The smaller the sample size n, the larger the standard error, the wider the sample means distribution. In general, we want n to be large.

2.8 Conclusion

In this course, we learned how to summarize data by measuring its center and its variance.

3 Notes from the course Intro to Inferential Statistics in Udacity

Materials on how to test your hypothesis and begin to make predictions based on statistical results drawn from data!

3.1 Lesson 1: Introduction and review

- A review of the materials in Lesson 18: Sampling Distribution.

3.2 Lesson 2: Estimation

- Point Estimate. Just use one sample mean to estimate the population mean.
- Margin of error: approximately 95% of sample means fall within $\frac{2\sigma}{\sqrt{n}}$ of the population mean in a normal distribution. Here $\frac{2\sigma}{\sqrt{n}}$ is called the margin of error.
- Interval Estimate. Find a confidence interval which contains the population mean using the assumption that the sample mean is within two standard deviations of the population mean.

3.3 Lesson 4: Hypothesis Testing

-

3.4 Lesson 6: t-Test

-

3.5 Lesson 12: One-Way ANOVA

-

3.6 Lesson 14: ANOVA

-

3.7 Lesson 16: Correlation

-

3.8 Lesson 18: Regression

-

3.9 Lesson 20: X^2 Tests

-

3.10 Lesson 22: Final Project

-