Chapter Notes for DMC 216: Nonparametric Statistics

Luka Trikha

February 23, 2022

1 Chapter One: Intro to Nonparametric Statistics

Parametric statistics is based around the idea that the given data is of a normal distribution. In conjunction, nonparametric statistics is based around data that is collected from a non-normal distribution. This can mean data that is purly nominal (categorical), ranked, based on scales, or simply, do not follow a normal distribution (either visually, or through mathematical tests) Nonparametric statistics are immune to outliers–they do not matter.

Some parametric assumptions include samples that:

1. Are randomly drawn from a normally distributed population.
2. Consists of indendent observations, except for paired values.
3. Consists of values on an interval or ratio measurement scale.
4. Have respective populations of approximately equal variance.
5. Are adequatitely large.
   - $n > 30$
   - $n > 20$
   - $n > 10$
   - (Per group as an absolute minimum).
6. Approximately resembles a normal distribution.

Although it is not required to have all of these assumptions to be checked off when analyzing your data, it is sometimes safe to not have one of these assumptions in your data. For example, you may need to increase your sample size to normalize your data (which, in turns, shows that your data is *actually* normal, and not nonparametric).

There are different ways to measure scales with the given data:

- ***Dichotomous*** is a measure of two conditions. There are two types of dichotomous scales:
  - ***Discrete dichotomous*** has no particular order.
    - male vs. female, heads vs. tails.
  - ***Continuous dichotomous*** has a measurement.
    - pass/fail, young/old.
- ***Ordinal*** describes values that occur in some order of rank.
  - Distance of two ordinal values hold no value.
  - Likert-type is 'on a scale of 1-5'.
- ***Interval scale*** is a measure in which the distance between any two sequential values are the same.
  - -8° to -7° is the same as 55° to 56°.
- ***Ratio scale*** has an absolute zero value, and is determined as a ratio.
  - Screen brightness starts at 0%, which means it is off, and goes to 100%, which means it is fully brighten.
- ***Repeated values*** during ranking is called *ties*.
  - In case of tie, you give them the average of their rank values.

While there are some similarities to parametric testing, the nonparametric procedure follows as such:

1. *State the null (H_0) and research (alternative/H_a) hypothesis.*
   - H_0 indicates no difference exists between conditions, groups, or variables.
   - H_a indiciates there exists a difference between conditions, groups, or variables.

- Direction means a significant change in a particular direction (skewness).
- Nondirectional means there is a change, but there are two tails (symmetric) and you cannot say there is a change in any direction.

2. *Set the level of signifigance (usually, it is* 5%).
3. *Use appropriate test statistic.*
4. *Compute test statistic.*
5. *Determine value needed for rejection of the* H_0 *using appropriate table of critical values for the perticular statistic.*
6. *Compare obtained value with critical value.*
   - State whether or not to reject H_0.
7. *Interperate results.*

	Fail to reject H_0	**Reject** H_0
H_0 True	No error	Type I error; α
H_0 False	Type II error; β	No error

Table 1: Results of H_0 outcome.

8. *Report results.*

2 Chapter Two: Testing Data for Normality

There are many different ways to figure out if the data you are using is normal. You can sometimes visualize it with a box-plot, histogram, and scatter plots. However, there are times that interpreting visualizations of data can not clearly tell you whether or not the data is normal. One way to find out if data is normal is by conducting some statistical math to find the skewness and kurtosis.

2.1 Describing data and the normal distribution

Sample is a collection of several independent, random measurements of a particular variable associated with a population. When you have a large sample size, the more likely your data will be normalized. However, there

are times when you do not have a lot of data, and sometimes your sample sizes can be less than 30. As stated earlier, having a sample size less than 30 can be a flag of non-normality, but there are ways to go about proving normality even with a small sample size. This measurement is called *variance*, s^2. Variance describes the scale over which the samples vary about the mean value, and be computed using Formula (1):

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \tag{1}$$

Where x_i is an individual value in the distribution, $\bar{x}$ is the distribution's mean, and n is the number of values in the distribution. In order to compare sample variances, it is suggested to take the variance ratio by taking the largest sample variances and dividing it by the smallest sample variances.

A more common way of expressing sample variance is *standard deviation*, s; the standard deviation is the square root of variance where $s = \sqrt{s^2}$. Standard deviation can be calculated by using Formula (2):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \tag{2}$$

A small standard deviation indicated that a sample's values are fairly concentrated about its mean, whereas a large standard deviation indicates that a sample's values are fairly spread out.

If we want to compare two or more samples from different distributions, then we need to compute the *standard score*. We can use the *z-score* of the multiple distributions, and it can be calculated by Formula (3):

$$z = \frac{x_i - \bar{x}}{s} \tag{3}$$

where x_i is an individual value in the distribution, $\bar{x}$ is the distribution's mean, and s is the distribution's standard deviation. A z-score that is below the mean will always negative, while a z-score that is above the mean will always be positive.

2.2 Computing and Testing Kurtosis and Skewness for Sample Normality

Kurtosis is a measure of a sample or population that identifie how flat or peaked it is with respect to a normal distribution. There are two different distributions to identify kurtosis:

- ***Leptokurtic distribution*** has a positive (pointy) kurtosis.
- ***Platykurtic distribution*** has a negative (flat) kurtosis.

Skewness is the horizontal symmetry with respect to a normal distribution. If a model is said to be skewed to the left, that means a distribution has a high concentration of data on the right side, and if a model is said to be skewed to the right, there distribution has a high concentration of data to the left. Right skewness is represented as a positive value, while left skewness is represented as a negative value.

There are five steps to determine sample normality in terms of kurtosis and skewness:

1. *Determine the sample's mean and standard deviation.*
   - You need to find the mean ($\bar{x}$) and the standard deviation (s) first before computing kurtosis and skewness values.
   - Recall to find sample standard deviation, use Formula (2), and to find the sample mean, you can use Formula (4).
2. *Determine the sample's kurtosis and skewness.*
3. *Calculate the standard error of the kurtosis and the standard error of the skewness.*
4. *Calculate the z-score for the kurtosis and the z-score fo the skewness.*
5. *Calculate the z-score with the critical region obtained from the normal distribution.*

$$\bar{x} = \frac{\sum x_i}{n} \tag{4}$$

Where $\sum x_i$ is the sum of the values in the sample and n is the number of values in the sample. The kurtosis K and standard error of the kurtosis SE_K are found using Formula (5) and (6) respectively:

$$K = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{S} \right)^4 \right] - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{5}$$

and

$$SE_K = \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)n+3)}} \tag{6}$$

The skewness S_K and standard error of the skewness SE_{S_K}, are found using Formula (7) and (8) respectively.

$$S_K = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{7}$$

$$SE_{S_K} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{8}$$

Normality can be evaluated by using the z-score for the kurtosis, z_K, and the z-score for the skewness, z_{S_K}. You can use Formula (9) and (10) respectively to find those z-scores:

$$z_K = \frac{K - 0}{SE_K} \tag{9}$$

$$z_{S_K} = \frac{S_K - 0}{SE_{S_K}} \tag{10}$$

You then use these z-scores with the values of the normal distributionfor a desired level of confidence α. If you set $\alpha = 0.05$, then the calculated z-scores for an approximately normal distribution must fall between -1.96 and $+1.96$.

2.3 Computing the Kolmogorov-Smirnov One-Sample Test

–Compares two distributions to see if they can plausibly be from the same distribution.

–Comonly used to test for normality.

–Uses the point where the two diverge ithe most to form the test statistics.

2.4 KS-Test (One-Sample)

- Can be used to test a sample for normaility by comparing it to a normal distribution.
- a "middle" and SD are computed and used as the mean and DS of the normal distribution.
- Can also be used to test against any other distribution.

The Kolmogorov-Smirnov one-sample test is a procedure to examine the agreement between two sets of values; the test compares two cumulative frequency distributions. A cumulative frequency distribution is useful for finding the number of observations above or below a particular value in a data sample. It is created by taking a given frequency and adding all the preceding frequencies in the list. Once you create a cumulative distribution for the observed and empirical frequency distribution, the test will allow us to find the point at which the two distributions show the largest divergence; which is then used to identify a two-tailed probability estimate p to determine if the samples are statistically similar or different

To start, we need to find the empirical frequency distribution $\hat{f}_{x_i}$ based on the observed sample. We calculate the observed frequency distribution's midpoint M and standard deviation s. The midpoint and standard deviation are found using Formula (11) and (12) respectivily:

$$M = (x_{max} + x_{min}) \div 2 \tag{11}$$

where x_{max} is the largest value in the sample and x_{min} is the smallest value in the sample, and

$$s = \sqrt{\frac{\sum (f_i x_i^2) - \frac{(\sum f_i x_i)^2}{n}}{n-1}} \tag{12}$$

where x_i is a given value in the observed sample, f_i is the frequency of a given value in the observed sample, and n is the number of values in the observed sample.

We now use the midpoint and standard deviation to calculate the z-scores [Formula (13)] for the sample values of x_i.

$$z = \left| \frac{x_i - M}{s} \right| \tag{13}$$

We now use the z-scores just obtained to determine the probability associated with each sample value, $\hat{p}_{x_i}$. These p-values are the relative frequencies of the empirical frequency distribution $\hat{f}_r$.

Now, we find the relative values of the observed frequency distribution f_r using Formula (14):

$$f_r = \frac{f_i}{n} \tag{14}$$

where f_i is the frequency of a given value in the observed sample and n is the number of values in the observed sample.

3 Wilcoxon Signed Rank test

Able to analyze the difference of trends within a dataset; higher + and − signs means a greater diffrerence, while if they are equal, signifyes no difference. USE rank.avg ON GOOGLE SHEETS

-H_0: difference between the pairs follows a *symmetric distribution* around zero.

-H_1: differenece between the pairs does not follow a symmetric distribution around zero.

Wilcoxon Signed Ranked Test:

- "Paired" Test–two numbers per unit
- Oftern a "before and after" - CAUSALITY?!?!?!?
- The data itself need not be ranking data.
- Can be used with small samples sizes.

Steps for a Wilcoxon Signed Ranked Test:

- Forumulate Hypothesis
- to be continud after class

A note on causality

Even with statistical signifigance for a randomized control experiment, double blind study or paired test such as Wilcoxon, we still need a logical explanation.

Statistics can tell you if something is happening, but it cannot tell you why or ***how*** *is is happening.*

3.1 filler

4 Binomial Distribution

Notation: $B(n,p)$ -"B" - for binomical -"n" - number of trials -n - number of trials -p - probability of "success"

Example: $B(10,0.5)$ is the distribution of number of heads for a coin lip with ten flips.

Trials = n

Probability of Success = p

Desired number of heads: x

What is $P(X=x)$ for $B(n,p)$?

$\frac{6}{(2^4)}$

$P(x=3)$ for $B(5,0.5)$ $P(x=3)$ for $B(5,0.3)$

$\frac{1}{1}$

4.1 Binomial CLT

A Binomial distribution $B(n, p)$ is approximated by $N(np, \sqrt{np(1-p)})$.
Assuming: $np \geq 10$ & $n(1-p) \geq 10$.

4.2 Binomial and the Sign Test