**Introduction**

In statistics today we are quick to stomp out any ideas students or researchers have about stating, for instance, “I have proven that the Earth is round.” Instead we shuffle them towards the conventional caveated conclusion, “The Earth is round, (p < .05)” (Cohen, 1994). Many such issues plague statisticians today; the bulk of which come from a lack of understanding of the plethora of tools in the statistical toolbox. For my thesis, I propose to write a methods paper discussing this downfall, frequently referred to as Null Hypothesis Significance Testing (NHST), and propose alternative options by comparing and contrasting NHST with nonparametric statistics, and a new form of statistical analysis, Observation Oriented Modeling. I will do this by introducing the three methods and then applying them to a dataset collected on the QWERTY effect.

**Null Hypothesis Significance Testing (NHST)**

To fully understand the NHST procedure, the reader must first be able to recognize the procedure and then look at the procedure’s origins. NHST is made up of 3 main steps: 1.) Create only one hypothesis that states that there are no differences between populations. 2.) Use the conventional rejection region of 5% to reject the null hypothesis. 3.) Always do these steps (Gigerenzer, 2004). Readers may be wondering why this procedure is used. The best way to answer this question is to explain how the NHST procedure originated.

Many believe that Ronald A. Fisher is to blame for the patchwork procedure of NHST. In reality, Fisher’s ideas looked nothing like NHST as it is taught and applied in current classrooms. Fisher believed in creating one “null” hypothesis, which he described as a hypothesis to be “nullified”, not as zero change or a zero correlation. For example, in traditional Fisher NHST, a null hypothesis could be ρ = .2. Second, Fisher believed that the use of an omnibus 5% level of significance showed a “lack of statistical thinking,” and instead we should report the exact significance value, which is in line with current APA standards (APA, 1999). Third, Fisher promoted NHST as a last resort of the many statistical tests available to be used for those problems. Since Fisher is not wholly to blame for this procedure as used today, we turn to two men who criticized Fisher’s NHST and created their own decision theory.

Jerzy Neyman and Egon S. Pearson created Neyman-Pearson decision theory, which consists of two hypotheses (a null and an alternative hypothesis) and a binary decision criteria. For example, in our experiment our null hypothesis could be that individuals prefer RN words and DN words equally and our alternative hypothesis would be that participants like RN words more than DN words. First, the production of two hypotheses creates the need for an alpha and beta level to be defined, as power, Type II error rates, and effect sizes come into play. Second, Neyman and Pearson clearly state that a hypothesis should not be blindly supported based solely on the support of one statistical test, but that replication and reproduction of results are also imperative. The key to this decision theory is being able to look at the outcome and make the most cost-effective decision. Third, this analysis is clearly against setting omnibus alphas and betas and is geared toward being able to adjust the analysis to the needs of the particular task at hand.

A quick comparison of these three processes shows that neither Fisher nor Neyman-Pearson is responsible for the null ritual. The blame can only be placed on ambition. During the time these procedures were arising, the field of Psychology was known as a “soft” science. Many researchers took up these procedures in order to show proof of the phenomena they were studying and to gain acceptance for the field as a science. In the rush to demonstrate our capabilities, the steps of the processes were muddled, the rational for the procedures were misplaced, and the role of the researcher as a sentient part of the analysis process was let slip away. Yet, this odd procedure has been allowed to continue on and is being taught, by even this writer, to introductory statistics classes everywhere. None of the pleas from such highly regarded individuals as B. F. Skinner, John W. Tukey, Ralph L. Rosnow and Robert Rosenthal, Jacob Cohen, Etienne P. LeBel and Kurt R. Peters, Gerd Gigerenzer, and many more have not been heard, and NHST has continued to persist in Psychology classrooms, research, and journals. Herein, I propose that we take a look at what is really happening in statistics today, and take a real look at some of the tools in the statistician’s toolbox.

**QWERTY**

The QWERTY effect is the phenomenon whereby people rate words as more pleasant if the words are composed of more letters that lie on the right hand side of a QWERTY keyboard than on the left hand side. Most recently, Jasmin and Casasanto (2012) wrote about how the QWERTY effect influences our perceptions of old words (prior to the invention of the QWERTY keyboard), new (after the invention of the QWERTY keyboard), and made-up (pseudowords that imitate real words). To expand upon this research, the study presented here has examined more than simple left-right preference. Beilock and Holt (2007) demonstrated that individuals prefer tying easier combination of letters; those with combinations on opposite hands as opposed to a single hand. This finding leads us to believe that “typability” needs to be taken into account when considering the perceived pleasantness of these words. This experiment sought to mimic the Jasmin and Casasanto study and also to look into how “typability” affected pleasantness (i.e. are the letters right-left-right-left more pleasant than right-right-left-right?).

For the purpose of this study the words were separated into three categories: repeated keystrokes not paired together (RN), repeated keystrokes together (RY), and not repeated with different fingers (DN). We would expect that those that are more typable would be preferred over those that are less typable. For NHST, we would start by creating our null and alternative hypotheses, which would be for the null, that there would be no difference between the three categories of words, and for the alternative, that there would be some significant difference between the three groups at the p < .05 level. We would assess this by running a repeated measures Analysis of Variance (ANOVA), which would take into account the fact that the same participants rated all three types of words. Now, before the analysis can be run there are many assumptions that must be met for the analysis to be accurate. The assumptions are as follows: 1) data screening has been performed for accuracy, missing, multicollinearity, and outlier analysis, 2) data is normally distributed, 3) data combinations are linear, 4) variances are homogeneous, and 5) variables pass the test of sphericity. Once we are sure that these assumptions are met we can run our repeated measures ANOVA. If the F-test from this analysis is significant, we are then able to run post-hoc analyses between the three groups to find out where these three groups are significantly different. Then we will be able to calculate effect size for the analyses and can speak to the general results of the study (Cohen’s *d*).

As this process may sound tedious already it is important to remember that there are alternatives to the NHST approach. Here we will discuss two applicable alternatives: nonparametric statistics, and Observation Oriented Modeling.

**Nonparametric Analysis**

In contrast to NHST, nonparametric statistics place less emphasis on assumptions, use simpler models which require less calculation, are based on simpler theories which allows a researcher to assess their data and choose the correct analysis, and they are frequently more powerful than parametric statistics (in this case NHST).

For the purpose of this study a nonparametric analysis for several related samples needed to be chosen. The two most well know of these tests are the Friedman test and the Quade test. The Quade test was chosen over the better-known Friedman test because with only three experimental groups (as we have in this study), Friedman’s test suffers from a lack of power. The Quade test will be used to analyze the differences in perceived pleasantness between the three different typing combinations—repeated keystrokes not paired together (RN), repeated keystrokes together (RY), and not repeated with different fingers (DN). This test depends on three assumptions: 1. Variables are mutually independent, 2. Observations within each block (for each participant) may be ranked according to some criterion, and 3. A sample range is capable of being determined within each block so they may be ranked. The test focuses on the ranked observations within each block and the average participant rank for the different word types in this study (Conover, 1999).

The repeated measures ANOVA and Quade test are comparable by viewing the asymptotic relative efficiency (A.R.E.) of the tests. A.R.E. refers to how two tests compare when it comes to sample size. If both tests have comparable levels of power and significance, the test that requires the smallest sample size is preferred and has a higher A.R.E. The A.R.E. score is the ratio of the second test over the first test (Conover, 1999). When the first test is better, the value is close to one, though the value can approach infinity. For the purposes of this experiment, G\* Power (Faul, Erdfelder, Lang, & Buchner, 2007) will be used to compute the sample sizes necessary to complete the Quade test and the F test (repeated measures ANOVA).

**Observation Oriented Modeling**

Many researchers have claimed that changing the way we use null hypothesis significance testing could help psychology. Conversely, Grice (2012) argues that our problems are “much deeper and wider” than just the NHST. He argues that with a new philosophical idea (realism), the absence of estimating population parameters, the use of integrated models, a renewed appreciation of “eye-tests,” and the formation of deep structures of data, that the social sciences can grow and amass knowledge in a better way than before.

Grice argues that we first need to shift our philosophical position to be in line with St. Thomas Aquinas and Aristotle. By viewing psychology through the lens of realism instead of positivism we should be able to properly and effectively conduct research and analyze data. In contrast to positivism, realism is the belief that effects conform to their cause.By viewing science as knowing nature through its causes, we can use Aristotle’s four causes to think in terms of forming structures and processes for phenomena. Switching to this philosophy allows for techniques that match the daily activities of social scientists in their endeavors to unravel the story of how humans work.

When using OOM, researchers can stop focusing on samples as they relate to the population. As Grice stated in a presentation to MSU (2012), “Who is the population anyway?” If a researcher actually attempted to count the population, it would be impossible. Take for instance research concerning individuals with cancer. Are we considering everyone with any type of cancer? Do we count all individuals with cancer just in the United States? Even as we are counting them some may die off or recover and leave the group, while others may be diagnosed and join the group. Trying to estimate and make sense of the abstract idea of a population is difficult and probably not worthwhile. OOM does not use population parameters and the various underlying assumptions; instead the researcher looks at observations at the level of the individual. Skinner (1955) stated that methods allowing direct observation of the individual would be the most important in behavioral sciences, and OOM allows us to put this idea into practice.

As a replacement to the procedure in psychological research, Grice addressed the need for integrated models. As opposed to the simpler variable-based modeling where X causes Y, an integrated model accounts for the structures and processes, which go into the actual phenomena that we study. When a detailed framework is constructed for an experiment, it is easy to ask specific questions about the model and to add to or change said model as each experiment reveals something new. For example, we expect to find a QWERTY model where participants should rate non-repeating words with higher scores than repeating words because they are easier to type.

These integrated models can be run on any type of data the social scientists will collect. From ordinal rankings to frequency counts, all analyses are calculated in the same general fashion. This simplicity is due to the fact that observation oriented modeling works on the deep structure of the data. Through observation definition, the program then breaks these units into binary code. This technique creates a common language, which allows analysis through various techniques. Deep structures can be arranged to form a matrix, which can then be manipulated via matrix algebra, binary Procrustes rotation, and other operations to investigate the data.

In OOM p-values are no longer used. The most important value any OOM analysis is the PCC or percent complete match. This value represents how well your observations matched the stated or expected pattern or, in the case of causal modeling, how many of your observations conformed to a given cause. This proportion is found by taking the number of complete matches and dividing it by the total number of participants. This PCC value replaces all of the conventional values for effect size used in statistical analyses.

As a secondary form of reference value, a chance value or c-value, is obtained by bootstrapping observations anywhere from 100 to 5,000 times. These randomized data sets are then compared to the pattern, which has been designated. If the randomized data sets fit the pattern more often than the actual data does, the c-value will be high. Low c-values are indicative of distinct observations that are not likely due to chance. Although low c-values, like low p-values, are desirable, c-values do not adhere to a strict cut-off, and should be considered a secondary form of confirmation for the researcher that their results are distinctive.

In order to analyze the repeated measures data that the QWERTY study produced, I will be using OOM’s newest analysis: Ordinal Pattern Analysis. For this analysis data is entered and defined just as in any other analysis and then a hypothesized pattern is created. This analysis allows the researcher to designate the expected ordinal pattern in which the researcher can define each variable as being higher, lower, or equal to the other variables. For instance, with our data, we hypothesized that the ratings of RN would be higher than those of DN, which would both, in turn, be greater than ratings of RY. Once this pattern is defined, the program then analyzes the data to see if each individual’s set of observations match this expected ordinal pattern. A PCC and c-value is generated based on the number of individuals that completely match the expected ordinal pattern. This analysis does not form any type of linear or nonlinear equation or regression, but simply looks for those individuals who match the expected ordinal pattern.

**Proposed Data Plan**

In order to assess how these various procedures work, I propose to use all three methods of analysis on a dataset testing for the QWERTY effect. By analyzing the data from the QWERTY study with a repeated measures ANOVA, a Quade test, and an Ordinal Pattern Analysis we can see the differences between the three methods and examine how well each method fits and explains the data at hand.

**Method**

**Participants**

Participants consisted of 157 students who were recruited from Introductory Psychology classes through Missouri State University’s SONA system and were given class credit for their participation. Participants were tested in groups that ranged from one participant to 4 participants at a time in Pummel 207B.

**Materials**

The experiment consists of rating words on pleasantness. Originally a list of 240 words was compiled to offer a variety of real words and pseudowords that were typed with repeated keystrokes not paired together (RN), repeated keystrokes together (RY), and not repeated with different fingers (DN). Each participant was only asked to rate 120 of the 240 words to rule out the effect of fatigue on their judgments. Of the 120 words that participants rated, they were randomized and split between real words and pseudowords (60 real words and 60 pseudowords words). For clarification, a table of word examples is listed in the appendix in Table 1.

|  |  |  |
| --- | --- | --- |
| Table 1  *Examples of word types.* | | |
| Word Type | Real words | Pseudowords |
| Repeated keystrokes not together | cute | yame |
| Repeated keystrokes together | hide | neeb |
| Not repeated with different fingers | blue | pobe |

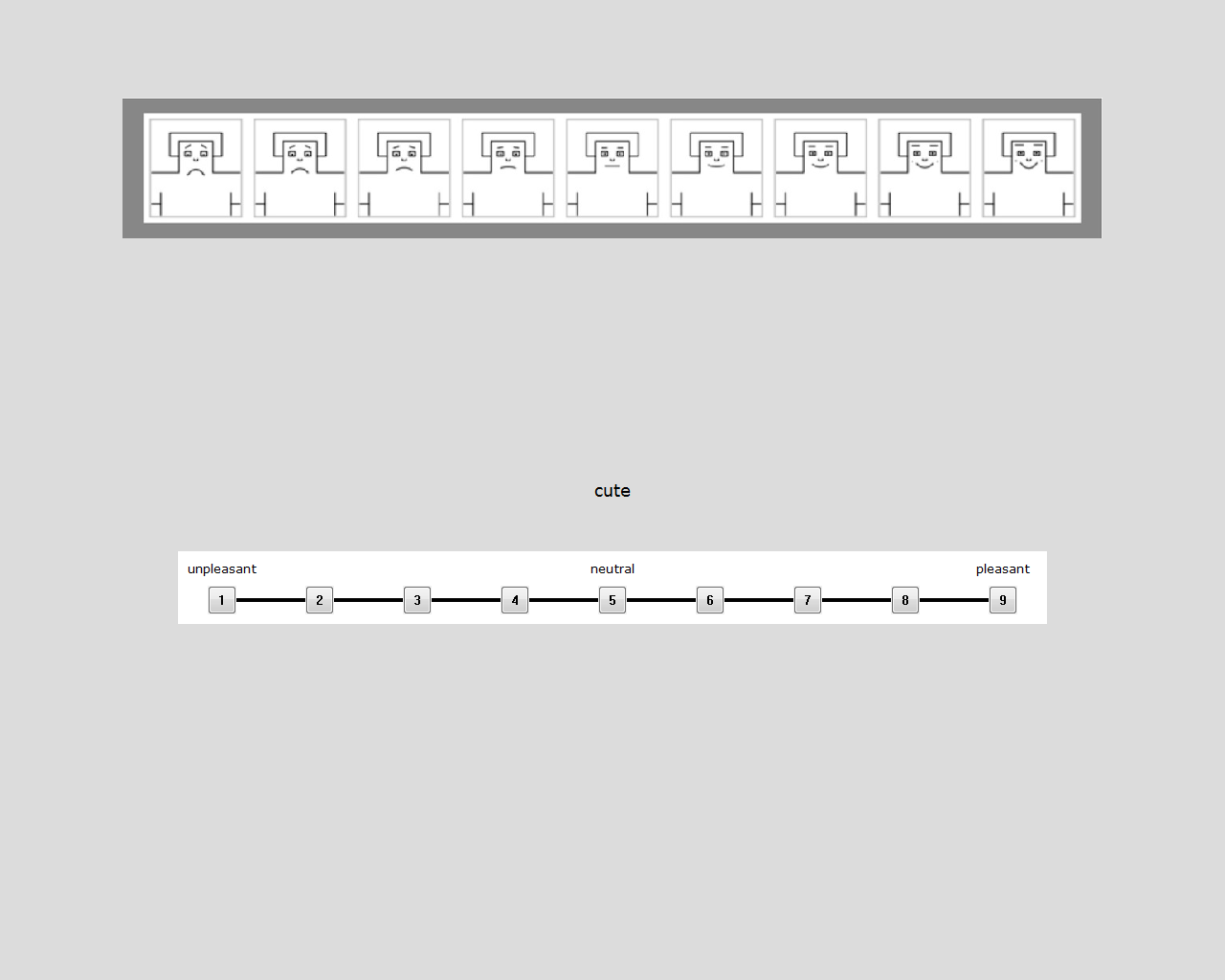
**Procedure**

Participants were asked to read and sign a consent form before entering the lab space and taking a seat at a computer. Participants were then asked to take a one minute long typing test that can be found at [www.typingtest.com/test.jsp](http://www.typingtest.com/test.jsp). The students were asked to complete as much of the Aesop’s fables test as they could in one minute. After the typing test, participants were asked to alert the researcher who recorded the typing speed and errors that appear on the screen after completion of the test.

The researcher then opens up the QWERTY experiment, enters their typing speed, and their accuracy rate, and then students are asked to complete the rest of the experiment. Participants were then asked to record which hand they write with, and then were shown a work and asked to rate how pleasant or unpleasant they find words. At no point in the experiment were participants asked to type the word they saw on the screen. Participants were asked to make this rating for 120 words. Participants were asked to use a self-assessment manikin (SAM; example attached) to rate their judgments. A SAM scale is a nine point visual emotional scale, which will allow the students to rate their perceived pleasantness of a word. After completing ratings of all 120 words, participants were allowed to exit the lab room. After completion of the experiment the researcher gave participants their class credit through the SONA system.

**Proposed Analysis**

I propose that the data be analyzed in three different analyses to compare and contrast the strengths and weaknesses of each of the methods. I first propose that a common repeated measures analysis of variance be run on the data with post-hoc paired-samples t-test to decipher any differences in how participants rated the three different types of words (RN, RY, DN). As participants rated all three types of words, a repeated measures ANOVA is appropriate for this data. Second, I propose that a nonparametric Quade test be run on the data with post hoc t-tests. The Quade test is the nonparametric version of a repeated measures test for multiple groups. Instead of running the data as is, the data is first ranked, and then given a Quade score, which can then be analyzed through a common ANOVA and t-tests. Third, I propose that an ordinal observations analysis be run on the data using Observation Oriented Modeling software. Observation Oriented Modeling, as described previously, is a new way to look at and analyzing data that focuses on participants at the individual level. The new ordinal analysis allows the researcher to designate a pattern that should theoretically be followed in the data, and then runs the data to see how well the actual observations match the theoretical pattern in an ordinal pattern. This procedure is also a repeated measure analysis as it looks at how an individual rates one word compared to another in all combinations of comparisons.



*Figure 1*. Screenshot of experimental rating procedure.

References

American Psychological Association (1999). Statistical methods in psychology journals:

Guidelines and explanations. *American Psychologist, 54* (8), 594-604.

Beilock, S. L. & Holt, L. E. (2007). Embodied preference judgments: Can likeability be

driven by the motor system? *Psychological Science, 18* (1), 51-57.

Conover, W. J. (1999). Practical nonparametric statistics. Elm Street Publishing Services,

Inc.

Cohen, J. (1994). The Earth is round (*p* < .05). *American Psychologist, 49* (12), 997-

1003.

Faul, F., Erdfelder, E., Lang, A, & Buchner, A. (2007). G\*Power 3: A flexible statistical

power analysis program for the social, behavioral, and biomedical sciences.

*Behavioral Research Methods, 39* (2), 175-191.

Grice, J. W. (2012, February 24). Observation oriented modeling: A common sense

approach to data conceptualization and analysis. Recorded at Missouri State

University. Podcast retrieved from iTunes RStats podcasts.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always

wanted to know about significance testing by were afraid to ask. In D. Kaplan

(Ed.) *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks, CA: Sage.

Jasmin, K. and Casasanto, D. (2012). The QWERTY Effect: How typing shapes the

meanings of words. *Psychonomic Bulletin Review, 19* (3), 499-504.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedure and the justification of

knowledge in psychological science. *American Psychologist, 44*(10), 1276-1284.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5),

221-233.