Appendix A: Statistical Methods

Guilherme Galhardo

1 Motivation

The following section explicates the statistical theory underlying our data analysis process and provides an overview of respective computational methods. Due to the ordinal and non-Gaussian nature of our data, typical parametric statistics —such as, say, mean, standard deviation, etc —and methods such as ANOVA are not appropriate. An in-depth discussion of this issue is outside the scope of this paper, but the key principle at play is that ordinal data doesn't admit the same kinds of distance metrics as numerical data.

For example, we can easily calculate that the distance, or difference, between an observation of \$1.00 and \$5.00 for the cost of an item is $|\$5.00 - \$1.00| = \$4.00$. And, critically, if we consider the distance between another pair of observations —such as $|\$9.00 - \$5.00| = \$4.00$ —we can say with certainty that the three observations are, so to speak, evenly spaced: i.e., that the difference between \$5.00 and \$1.00 is the same as the difference between \$9.00 and \$5.00. What, though, is the distance between a Likert-Scale observation of 'Agree' and 'Somewhat Agree?' Is it the same as both the difference between 'Agree' and 'Strongly Agree,' as well as the difference between 'Disagree' and 'Somewhat Disagree,' and so on? Could we say that these differences are exactly the same for any respondent in any situation, and that they are precisely quantifiable in a consistent, generalizable way? Can we even be certain that two observations of 'Agree' actually represent the exact same level of agreement?

Certainly not. All we can generally say about these observations is that, up to a given construct, 'Agree' represents "more" agreement than 'Somewhat Agree,' and that there is a well-ordered ranking from 'Strongly Disagree' to 'Strongly Agree.' There are, however, various methods to work with ordinal data in a numerical form, such as by transforming a seven-point Likert Scale into the ordered set 1, 2, 3, 4, 5, 6, 7 —or, equivalently, -3, -2, -1, 0, 1, 2, 3. This allows the use of non-parametric statistical methods, which make fewer and less restrictive assumptions about the nature and distribution of observations.

As such, non-parametric methods were chosen in Cliff's δ for effect size and the Brunner-Munzel test for significance. The Brunner-Munzel test, in particular, was instrumental to our analysis, since it is robust to as few as ten observations per factor level, and many of our predictor bins were quite small. Moreover, between seven and ten observations per factor level, the permuted Brunner-Munzel test sustains performance, which enabled analyses with predictor bins containing as few as eight observations.

2 Effect Size

Cliff's δ is a robust statistic that measures stochastic dominance between two distributions $\mathcal{X}, \mathcal{Y}$. More specifically, given samples $X = \{x_1 \dots x_i \dots x_n\}$ and $Y = \{y_1 \dots y_j \dots y_m\}$ with $X \sim \mathcal{X}$ and $Y \sim \mathcal{Y}$ and for an arbitrary pair of observations x_i, y_j, $\hat{\delta} = \mathbb{P}(x_i > y_j) - \mathbb{P}(x_i < y_j)$. Order does matter when comparing samples, and $\hat{\delta}$ will be positive specifically when X dominates Y, with the same test done in the reverse order yielding $-\hat{\delta}$. Cliff's δ is a non-parametric analogue of Cohen's d, and is therefore well-suited to distributions that are asymmetric and/or non-Gaussian.

2.1 Point Estimate

We can compute $\hat{\delta}$ as a point estimate of δ using the following formulae:

$$\delta(i, j) \coloneqq \begin{cases} +1, & \text{if } x_i > y_j \\ -1, & \text{if } x_i < y_j \\ 0, & \text{if } x_i = y_j \end{cases} \qquad \text{for } i \in [1, n], j \in [1, m] \tag{1}$$

$$\hat{\delta} = \frac{1}{nm} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(i, j) \tag{2}$$

δ ranges from -1 to 1, with 0 indicating stochastic equivalence between $\mathcal{X}$ and $\mathcal{Y}$. More specifically, $\delta = 1$ implies that $\mathcal{X} \succeq \mathcal{Y}$, while $\delta = -1$ implies that $\mathcal{Y} \succeq \mathcal{X}$. In terms of effect size, $|\hat{\delta}| \in [0, 0.15)$ is generally considered negligible, $|\hat{\delta}| \in [0.15, 0.33)$ small, $|\hat{\delta}| \in [0.33, 0.47)$ moderate, and $|\hat{\delta}| \in [0.47, 1]$ large.

2.2 Standard Error

The standard error of $\hat{\delta}$ is defined by the following:

$$\delta_r(i) \coloneqq \frac{1}{m} \cdot \sum_{j=1}^{m} \delta(i, j) \tag{3}$$

$$\delta_c(j) \coloneqq \frac{1}{n} \cdot \sum_{i=1}^{n} \delta(i, j) \tag{4}$$

$$d_r \coloneqq \sum_{i=1}^{n} (\delta_r(i) - \delta)^2 \tag{5}$$

$$d_c \coloneqq \sum_{j=1}^{m} (\delta_c(j) - \delta)^2 \tag{6}$$

$$s_\delta = \sqrt{\frac{n^2 \cdot d_r + m^2 \cdot d_c - \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\delta\left(i,j\right) - \delta\right)^2}{nm \cdot (n-1) \cdot (m-1)}} \tag{7}$$

2.3 Confidence Intervals

While a symmetric confidence interval with confidence level $(1-\alpha)\%$ can be constructed with $\hat{\delta} \pm s_\delta \cdot z_{\frac{\alpha}{2}}$, there exists a narrower, asymmetric interval given by the following expression:

$$\frac{\hat{\delta} - \hat{\delta}^3 \pm z_{\frac{\alpha}{2}} \cdot s_\delta \sqrt{1 - 2 \cdot \hat{\delta}^2 + \hat{\delta}^4 + \left(z_{\frac{\alpha}{2}} \cdot \hat{\delta}\right)^2}}{1 - \hat{\delta}^2 + \left(z_{\frac{\alpha}{2}} \cdot \hat{\delta}\right)^2} \tag{8}$$

Rather than symmetry of real parameter magnitude about a midpoint estimate, this interval privileges probabilistic symmetry. In other words, the real distance between the point estimate and either extreme of the interval may be different, but the probability captured by either segment about the point estimate will be the same. In fact, using this asymmetric confidence interval, one can perform a coarse test of statistical significance: if, for a $(1-\alpha)\%$ confidence interval I, $0 \in I$, then the result is not significant at level α. A finer-grained analysis is typically warranted, particularly in post-hoc testing, but this method serves as a reasonable frame of reference.

3 Significance Testing

The Brunner-Munzel test is a non-parametric test for significance that replaces the Mann-Whitney test in cases where the test samples of interest have different variances or even come from different families of distributions. In terms of parametric methods, Brunner-Munzel is a non-parametric analogue to the Satterthwaite-Smith-Welch t-Test. Given samples $X = \{x_1 \dots x_i \dots x_n\}$ and $Y = \{y_1 \dots y_j \dots y_m\}$ with $X \sim \mathcal{X}$ and $Y \sim \mathcal{Y}$, and for an arbitrary pair of observations $x_i,\ y_j$, Brunner-Munzel can be used to test null hypotheses of the following forms:

- **Two-Sided Hypothesis:**
  - $H_0 : \mathbb{P}\left(x_i < y_j\right) = \mathbb{P}\left(y_j < x_i\right)$
- **One-Sided Hypotheses:**
  - $H_0 : \mathbb{P}\left(x_i < y_j\right) > \mathbb{P}\left(y_j < x_i\right)$
  - $H_0 : \mathbb{P}\left(x_i < y_j\right) < \mathbb{P}\left(y_j < x_i\right)$

In the following subsections, we will explore the formulation of the test as described in (Brunner & Munzel, 2000).

3.1 Relative Treatment Effect

3.1.1 Notation

In order to evaluate the relative ranks of all observations $(x_i)_{i=1}^{n}$ and $(y_j)_{j=1}^{m}$, we first consider the union of the two test samples, $A \coloneqq X \cup Y$, with elements indexed as $a_{(k,\,l)}$ and cardinality $|A| = N = n + m$. For ease of computation, we rename the two test samples and their cardinalities such that $X = A_1$ with n_1 elements and $Y = A_2$ with n_2 elements, indexed by k such that $A = \bigcup_{k=1}^{2} A_k$. So, for any element $x_i \in X$ with $i \in [1,\, n]$, we write $a_{(1,\,l)}$, with $l \in [1,\, n_1]$. Likewise, for any element $y_j \in Y$ with $j \in [1,\, m]$, we write $a_{(2,\,l)}$, with $l \in [1,\, n_2]$. Moreover, we will denote the distributions of the test samples X and Y, $\mathcal{X}$ and $\mathcal{Y}$, by $F_k(a)$ such that $A_k \sim F_k$, with expectation μ_k, variance σ_k^2, and combined distribution function $H(a) = \frac{1}{N} \cdot \sum_{k=1}^{2} n_k \cdot F_k(a)$.

3.1.2 Derivation of $\hat{p}$

If we define the relative treatment effect as $p = \mathbb{P}\left(x_i < y_j\right) + \frac{1}{2} \cdot \mathbb{P}\left(x_i = y_j\right)$, $p = 0.5 \iff \mu_1 = \mu_2$ will correspond to the two-sided null hypothesis and $p < 0.5 \iff \mu_1 > \mu_2$ and $p > 0.5 \iff \mu_1 < \mu_2$ will respectively correspond to each of the one-sided nulls. In order to test these hypotheses, we will use normalizations of the distribution functions $F_k(a)$ such that $\mathcal{F}_k(a) = \frac{1}{2} \cdot \left[F_k^{-}(a) + F_k^{+}(a)\right]$. Here, for arbitrary $a_{(k,\,l)} \in A_k$, $F_k^{-}(a) = \mathbb{P}\left(a_{(k,\,l)} < a\right)$ is the left-continuous distribution function, and $F_k^{+}(a) = \mathbb{P}\left(a_{(k,\,l)} \leq a\right)$ is the right-continuous distribution function. We can then define the relative treatment effect as $p = \int F_1 \,\mathrm{d}F_2$

In order to estimate p by its sample parameter $\hat{p}$, we will use approximations of the distribution functions in $\hat{F}_k(a)$ such that our approximation of the normalized, combined distribution function is $\hat{\mathcal{H}}(a) = \frac{1}{N} \cdot \sum_{k=1}^{2} n_k \cdot \hat{\mathcal{F}}_k(a)$. Letting $R(k,\, l)$ be the rank of $a_{(k,\,l)}$ within A, $R_k(k,\, l)$ be the rank of $a_{(k,\,l)} \in A_k$ within A_k, and $\bar{R}_k$ be the mean of the within rankings R_k of sample A_k, we construct the following equations:

$$R(k,\, l) = \frac{1}{2} + N \cdot \hat{\mathcal{H}}(a_{(k,\,l)}) \tag{9}$$

$$\bar{R_k} = \frac{1}{n_k} \cdot \sum_{l=1}^{n_k} R(k,\, l) \tag{10}$$

In the case of tied rankings between Q such elements of an arbitrary set $\mathcal{A} \coloneqq \{a_q \,|\, a_{q_i} = a_{q_j}\, \forall\, i,\, j \in [1,\, Q]\}$, we assign the midrank between all Q elements to every $a_{(k,\,l)} \in \mathcal{A}$. That is, for the Q tied ranks in the ordinal set $\{a_{(1)},\, a_{(2)},\, \ldots,\, a_{(i=q_1)},\, \ldots,\, a_{(j=q_Q)},\, \ldots,\, a_{(N)}\}$, we would assign the mean rank of

$w_{\bar{q}} = \frac{1}{(j-i)+1} \cdot \sum_{w_q = i}^{j} (w_q)$ to all the $a_q \in \mathcal{A}$, where (w_q) is the rank of each element if we were to simply choose an arbitrary order for the elements of $\mathcal{A}$ and continue to increment the ranks as normal.

Let $\mathcal{C}(u)$ be the normalized count function, such that $\mathcal{C}(u) \mapsto \{0, \frac{1}{2}, 1\}$ as u evaluates to $<$, $=$, $>$, respectively. It then follows that we can calculate $\hat{p}$ using the formula below:

$$\hat{p} = E[\mathcal{C}(A_{(2,\,1)} - A_{(1,\,1)})] = \int \hat{F}_1 \, \mathrm{d}\hat{F}_2 = \frac{1}{n_1} \cdot \left(\bar{R}_2 - \frac{n_2 + 1}{2} \right) \tag{11}$$

3.1.3 Derivation of $\hat{\sigma}^2_N$

Define $G_{(k_i,\,l)} \coloneqq F_{k_j}\left(A_{(k_i,\,l)}\right)$, with $l \in [1,\, n_{k_i}]$, and $\bar{G}_{k_i} \coloneqq \frac{1}{n_{k_i}} \cdot \sum_{l=1}^{n_{k_i}} F_{k_j}\left(A_{(k_i,\,l)}\right)$. Then, $\sigma^2_{k_i}$ can be estimated using the approximation below:

$$\tilde{\sigma}^2_{k_i} = \frac{1}{n_{k_i} - 1} \cdot \sum_{l=1}^{n_{k_i}} \left(G_{(k_i,\,l)} - \bar{G}_{k_i}\right)^2 \tag{12}$$

Replacing the unknown distribution functions above with their previously derived, normalized approximations and letting the within rank $R_{k_j}(k_j,\, l) = \frac{1}{2} + n_{k_i} \cdot \hat{\mathcal{F}}_{k_i}\left(A_{(k_i,\,l)}\right)$, we proceed as follows:

$$n_{k_i} \cdot \hat{\mathcal{F}}_{k_i}\left(A_{(k_j,\,l)}\right) = N \cdot \hat{\mathcal{H}}\left(A_{(k_j,\,l)}\right) - n_{k_j} \cdot \hat{\mathcal{F}}_{k_j}\left(A_{(k_j,\,l)}\right) = R(k_j,\, l) - R_{k_j}(k_j,\, l) \tag{13}$$

Using the above, we can then calculate $\hat{\sigma}^2_{k_i}$ with the following formulae:

$$S^2_{k_i} = \frac{1}{n_{k_i} - 1} \cdot \sum_{l=1}^{n_{k_i}} \left(R(k_j,\, l) - R_{k_j}(k_j,\, l) - \bar{R}_k + \frac{n_{k_i} + 1}{2} \right)^2 \tag{14}$$

$$\hat{\sigma}^2_{k_i} = \frac{S^2_{k_i}}{(N - n_{k_i})^2} \tag{15}$$

Finally, with both sample variances accounted for, we can compute a combined variance of $\hat{\sigma}^2_N = N \cdot \left[\frac{\hat{\sigma}^2_1}{n_1} + \frac{\hat{\sigma}^2_2}{n_2} \right]$

3.1.4 The W^{BF}_N Test Statistic

In order to test the relative treatment effect, we will use the Behrens-Fisher W with a Satterthwaite-Smith-Welch distribution. Proving the normality of the test statistic

is beyond the scope of this paper, but suffice it to say that the following test statistic is asymptotically normal under the null hypothesis:

$$W_N^{BF} = \frac{\left(\hat{p} - \frac{1}{2}\right) \cdot \sqrt{N}}{\hat{\sigma}_N} = \frac{\bar{R}_2 - \bar{R}_1}{\hat{\sigma}_N \cdot \sqrt{N}} \tag{16}$$

Furthermore, we can estimate the degrees of freedom by using the Welch-Satterthwaite equation. If the degrees of freedom of each sample variance is given by $\nu_i = n_i - 1$, then:

$$\nu_{\chi'} \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2 \cdot \nu_1} + \frac{S_2^4}{n_2^2 \cdot \nu_2}} \tag{17}$$

And, in the case of small sample sizes–which might otherwise lead to degeneration of the variance estimators–we can instead apply the following estimate for degrees of freedom:

$$\hat{f} = \frac{\left(\sum_{i=1}^{2} \frac{\hat{\sigma}_i^2}{n_i}\right)^2}{\sum_{i=1}^{2} \frac{\left(\frac{\hat{\sigma}_i^2}{n_i}\right)^2}{n_i - 1}} = \frac{\left(\sum_{i=1}^{2} \frac{S_i^2}{N - n_i}\right)^2}{\sum_{i=1}^{2} \frac{\left[\frac{S_i^2}{N - n_i}\right]^2}{n_i - 1}} \tag{18}$$

4 References

Brunner, E., & Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. Biometrical Journal, 42(1), 17–25. doi:10.1002/(sici)1521-4036(200001)42:1¡17::aid-bimj17¿3.0.co;2-u