

University of California Davis

STA104: Non-Parametric Statistics

Lecturer: Erin Melcon

Scribe: Ramneek Narayan

Last Revision: April 18, 2020

Table of Contents

0	About/Usage	3
0.1	About this Book	3
0.2	How to use this Book	3
1	Week 1: Introduction	4
1.1	Lecture 1: Overview	4
1.2	Lecture 2: HT/CIs for Median (Binomial Distribution)	8
1.3	Lecture 3: Median CIs and Percentiles/C.D.F.s	12
2	Week 2: 2-Sample Tests	18
2.1	Lecture 4: CIs for Percentiles & Permutation Tests	18
2.2	Lecture 5: Permutation Tests (cont.) & WRS Test	23
2.3	Lecture 6: WRS (cont.) & Approximations	28
2.4	Appendix (Week 2)	33
3	Week 3: More 2-Sample Tests & Comparisons	38
3.1	Lecture 7: Mann-Whitney Test	38
3.2	Lecture 8: Shift CI (cont.) & KS Test	43
3.3	Lecture 9: Comparison of Two-Sample Tests	47
3.4	Appendix (Week 3)	51
4	Week 4: More Non-Parametric ANOVA	52

4.1	Lecture 10: ANOVA Permutation & KW Test	52
4.2	Lecture 11: KW Example & Mult. Comparisons	57
5	Week 5: Group Comparisons (cont.) & Linear Tests	62
5.1	Lecture 12: Permutation Cutoffs/Rev. of Linear Tests	62
5.2	Lecture 13: Non-Parametric Linear Tests	66
5.3	Lecture 14: Ranked Correlation Tests	70
5.4	Appendix (Week 5)	74
6	Week 6: Correlation Tests (cont.) & Tests for Independence	76
6.1	Lecture 15: Hypothesis Tests for Kendall's Tau	76
6.2	Lecture 16: More on Correlation & Contingency Tables	81
6.3	Lecture 17: Permutation Test for Independence	84
7	Week 7: Prob. Comparisons & Bootstrapping	87
7.1	Lecture 18: Independence & Bootstrapping (Intro)	87
7.2	Lecture 19: Bootstrap Point/Interval Estimation	90
7.3	Lecture 20: BCA Bootstrap CI	94
8	Week 8: Bootstrap "t" Interval	98
8.1	Lecture 21: Bootstrap "t" Interval	98
9	Week 9: Interval Comparisons & KNN	102
9.1	Lecture 22: Interval Comparisons	102
9.2	Lecture 23: K-Nearest Neighbors (KNN)	104
10	Week 10: KNN (cont.)	107
10.1	Lecture 24: More KNN & CV	107
	References	111

Chapter 0

About/Usage

0.1 About this Book

These are some notes that were typeset by Ramneek Narayan after he completed STA104. There are new additions such as theorems, definitions, appendices, references, and proofs. These notes were created in an effort to make the material for the course more accessible and reader friendly. I'm sure there are typos and if you spot any, let the writers know.

0.2 How to use this Book

This book was written with the student in mind and comes with colored environments to make reading easier. In addition, at the end of each environment are symbols used conclude the environment (show that it is completed); they are there for organization and for your ease of reading. We list the environments below for clarity:

Example = Red Violet concludes with '♥'

Remark = Teal concludes with '♦'

Definition = Lime Green concludes with '♠'

Theorem = Royal Purple concludes with '■'

Proposition = Mulberry concludes with '■'

Note = Orange concludes with '▲'

Emphasis = Royal Blue concludes with '★'

Read at your own pace and if anything doesn't make sense, argue with the instructor! It makes sense at the end sometimes. Other than that, the book is pretty straight forward to read. We hope you enjoy reading it!

Chapter 1

Week 1: Introduction

1.1 Lecture 1: Overview

Non-parametric statistics uses techniques that do not require typical assumptions of traditional techniques (so-called *parametric statistics*). To see why we still use nonparametric methods in statistics today, consider a traditional text for a single mean. It has the following form:

$$H_0 : \mu = \mu_0 \text{ or } H_0 : \mu \leq \mu_0 \text{ or } H_0 : \mu \geq \mu_0 \text{ vs.}$$

$$H_A : \mu \neq \mu_0 \text{ or } H_A : \mu > \mu_0 \text{ or } H_A : \mu < \mu_0$$

The test statistic normally used is the t -statistic $t_s = (\bar{x} - \mu_0)/(s/\sqrt{n})$ with degrees of freedom $df = n - 1$. Each respective H_A has the following p-values:

$$2P(t > |t_s|), \quad P(t > t_s), \quad P(t < t_s)$$

As always we use the same decision rule for making an inference:

1. If p-value $< \alpha$, reject H_0
2. If p-value $\geq \alpha$, fail to reject H_0

When we conducted the t -test, **what assumptions did we make and why?**

Two assumptions that stand out are:

Assumptions of t -test

1. Random sample was taken, i.e. X_i 's are all independent of each other
2. \bar{X} (the random variable (r.v.) denoting all possible sample means) is assumed to be approximately normal, i.e. $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$. We know this because:
 - $n \geq 30$ (using the Central Limit Theorem) OR
 - Population is normal



We need these assumptions because:

Reasons for Assumptions

1. Random sampling allows us to simplify the formula for the variance of the sample mean $\sigma_{\bar{X}}^2$ as $\sigma_{\bar{X}}^2 = \sigma_X^2/n$ since $V(\bar{X}) = \sum V(X_i)/n^2 = V(X)/n$ assuming mutual independence.
2. Normality of \bar{X} lets us say $t = (\bar{X} - \mu_0)/(S/\sqrt{n})$ is t_{n-1} distributed, i.e. $t \sim t_{n-1}$. It also in turn allows us to make confidence intervals (CIs) and hypothesis tests (HTs).

Note: We use this statistic because $\sigma_{\bar{X}}$ is unknown. If it was known, then we replace S with $\sigma_{\bar{X}}$ in the test statistic t , giving $t \sim N(0, 1)$ here.



Notice, however, that the key part in making this test was that we assumed **some test-statistic had a known named distribution**. Then, we used this distribution to find percentiles for CIs and probabilities for p-values.

Sometimes this doesn't happen when actually collect data.

To remedy this, we have non-parametric statistics. It addresses the issue that arises when we do not have the assumptions we need to assume a named distribution. For this reason, non-parametric statistics is often called "*distribution free statistics*."

1.1.1 What happens if assumptions are violated?

Typically, the assumption that is most often violated is whatever distribution we assumed (typically a normal distribution). This makes the errors involved in HTs and CIs to grow really fast.

Recall:

- $\alpha = P(\text{Type I Error}) = P(\text{Reject } H_0 | H_0 \text{ true})$
- $\beta = P(\text{Type II Error}) = P(\text{Fail to reject } H_0 | H_0 \text{ false})$
- $\text{Power} = 1 - \beta = P(\text{Reject } H_0 | H_0 \text{ false})$



A good measure of the accuracy of a hypothesis test is **power**; we are especially interested in a test's ability to detect a false null hypothesis since we usually have an intuition for suspecting (creating) an alternative hypothesis (H_A). Now, if the assumptions for distributions hold, parametric tests will often have more power. However, when assumptions are violated, **non-parametric tests will have higher power**.

1.1.2 Common Parametric Tests

Some common tests you may be familiar with:

Well-known Parametric Tests

1. $H_0 : \mu = \mu_0$ (test for a single mean)
 - Requires $\bar{X} \sim N(\mu_X, \sigma_X/\sqrt{n})$
2. $H_0 : \mu_1 - \mu_2 = \Delta_0$ (comparison of two means)
 - Requires \bar{X}_1, \bar{X}_2 are independent and $\bar{X}_i \sim N(\mu_i, \sigma_i/\sqrt{n})$ for $i \in \{1, 2\}$
3. $H_0 : \mu_1 = \mu_2 = \dots = \mu_l$ (ANOVA)
 - Requires $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_{\epsilon_i}^2)$ and l groups are mutually independent
4. $H_0 : \beta_i = 0$, assuming $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \epsilon_i$
 - Requires $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$



Each test has its own caveats and they are especially weakened when their assumptions do not hold due to...

1. Outliers
2. Small sample sizes
3. Normal assumption violated

Fortunately, there are some solutions to these problems and we can

1. Remove outliers, where appropriate
2. Data transformations under non-normality

However, there are currently no known methods for small sample sizes and our "solutions" above don't always work. Hence prompting a need for non-parametric methods.

1.1.3 Broad Overview of Non-Parametrics

Here are some common non-parametric techniques:

Common Non-Parametric Methods

1. **Binomial Based Tests:** We assume something is distributed binomial, which is somewhat less strict than normal

2. **Permutation Tests:** We randomize data among groups, and use a permutation distribution to form HT and CIs
3. **Bootstrap Methods:** Resample the data from itself with replacement to form a bootstrap distribution to use for CIs and HTs.
4. **"Modern"/Machine Learning Techniques:** Primarily "K nearest neighbors", Regression/Classification Trees, and possibly Linear Discriminant Analysis (LDA)



1.2 Lecture 2: HT/CIs for Median (Binomial Distribution)

The reason we often use the mean is because of the CLT, which tells us the mean is approximately normal if a random sample was taken from a population with sample size greater than 30 ($n \geq 30$).

But, when we can't assume this, the median is often just as good since it is also a measure of central tendency. Before going into more detail, recall the definition of the median:

Definition 1.2.1 (Median). For continuous data, the **median** is the value $X_{(m)}$ such that $\approx 50\%$ of the data lies below it and $\approx 50\%$ lies above it. I.e.,

$$P(X < X_{(m)}) \approx 0.50, \quad P(X > X_{(m)}) \approx 0.50$$



When we sample from a population, we compute what are known as *sample medians*, marked as $x_{(m)}^j$ for the j -th sample. If we were to plot a histogram of all these sample medians, we would eventually arrive at the **sampling distribution of the population median**. The random variable that describes this distribution is known as θ_m and our best bet as to what the population median is known as the **hypothesized median**, denoted as θ_m° .

Now, stating H_0 and H_A in terms of θ_m and θ_m° yields the following hypotheses:

Hypotheses for Median HT

1. $H_0 : \theta_m \leq \theta_m^\circ$ vs. $H_A : \theta_m > \theta_m^\circ$
2. $H_0 : \theta_m \geq \theta_m^\circ$ vs. $H_A : \theta_m < \theta_m^\circ$
3. $H_0 : \theta_m = \theta_m^\circ$ vs. $H_A : \theta_m \neq \theta_m^\circ$



We could also phrase the above hypotheses in terms of probabilities if we set $p = P(X_i > \theta_m^\circ)$ where X_i is any data point we sampled. Then the hypotheses become

$$H_0 : p \leq 0.5 \text{ vs. } H_A : p > 0.5,$$

for example.

Consider marking those observations that are above our predicted value of the median, we can formalize this as:

$$B_i^+ = \begin{cases} 1 & \text{if } X_i > \theta_m^\circ \\ 0 & \text{if } X_i < \theta_m^\circ \end{cases}$$

Immediately, we can see that $B_i^+ \sim \text{Bernoulli}$ since it consists of only two outcomes. Notice that $\sum_i B_i^+ = \# \text{ of } X_i > \theta_m$. We will call this quantity B^+ .

If the null hypothesis H_0 is true (we will use the equality sign to make computations easier for the other two versions of the null hypothesis, so $\theta_m = \theta_m^\circ$), then

$$B^+ \sim \text{Bin}(n, 1/2)$$

since it is a sum of Bernoulli trials.

Thus, our p-value, which is the probability of observing our sample data **as or more extreme if the H_0 is true**, is:

P-values for Median HT

1. If $H_A : \theta_m > \theta_m^\circ \Rightarrow \text{p-value} = P(B^+ \leq b^+)$
2. If $H_A : \theta_m < \theta_m^\circ \Rightarrow \text{p-value} = P(B^+ \geq b^+)$
3. If $H_A : \theta_m \neq \theta_m^\circ \Rightarrow \text{p-value} = \min\{P(B^+ \geq b^+), P(B^+ \leq b^+)\}$

Where $B^+ \sim \text{Bin}(n, 1/2)$ as before and b^+ is the observed value of this r.v. from our sample. ★

As usual, reject H_0 if p-value $\leq \alpha$. Notice, B^+ is essentially the test statistic.

Notice that if $H_0 : \theta_m < \theta_m^\circ$ is true, then $P(B^+ \geq b^+)$ should be at least 0.5. Similarly, if H_0 is not true, we would expect $P(B^+ \geq b^+)$ to be less than 0.5. The other hypotheses reverse the direction of the inequality or make $P(B^+ \geq b^+)$ exactly 0.5.

Example 1.2.1 (Midterm Scores). Suppose a sample of 12 students had the following midterm I scores:

Scores : 55, 64, 65, 67, 67, 68, 69, 72, 73, 75, 80, 88

Suppose the hypothesis is the median is at least 70.

(a) State H_0 and H_A

- **Solution:** The claim is that $\theta_m \geq 70$, as such we will put it as a null hypothesis and use our data to seek evidence to the contrary. Thus, the null and alternative hypotheses are:

$$H_0 : \theta_m \geq 70 \quad H_A : \theta_m < 70$$

(b) Calculate the appropriate p-value

- **Solution:** Since $H_A : \theta_m < 70$, we define the p-value as $P(B^+ \leq b^+)$. In this problem we have $b^+ = \# \text{ of observations } > 70 = 5$. The p-value is then $P(B^+ \leq 5) = P(B^+ = 5) + P(B^+ = 4) + \dots + P(B^+ = 0) \approx 0.3872$ (using R or table of binomial probabilities).

(c) Interpret the p-value in terms of the problem

- **Solution:** If the null hypothesis was true ($\theta_m \geq 70$), then the chance we would observe our data or "more extreme" (less values greater than 70) is about 0.3872.

(d) State your conclusion in terms of the problem

- **Solution:** Since $p\text{-value} < \alpha$, we fail to reject the null and conclude the median is at least 70.



Remark 1.2.1. This test only requires

1. A random sample was taken from a continuous distribution

Also note that if any $X_i = \theta_m^\circ$, then we usually remove these values (thereby reducing n) and carry out the test.



1.2.1 Normal Approximation to Binomial Test

Now, with a reasonable sample size, we may assume (using the CLT) that

$$\sum B_i^+ \sim N(np, \sqrt{np(1-p)}), \text{ where } p = 1/2 \text{ under } H_0$$

So we can then make a test statistic using a Z (standard normal) distribution. Our hypotheses stay exactly the same and for clarity we rewrite below:

Hypotheses under Normal Approximation

1. $H_0 : \theta_m \leq \theta_m^\circ$ vs. $H_A : \theta_m > \theta_m^\circ$
2. $H_0 : \theta_m \geq \theta_m^\circ$ vs. $H_A : \theta_m < \theta_m^\circ$
3. $H_0 : \theta_m = \theta_m^\circ$ vs. $H_A : \theta_m \neq \theta_m^\circ$



Interestingly, the test statistic we calculate for all three types of tests remains the same: $S = B^+$. To have universal testing procedures, we normalize this accordingly to give final form:

$$Z = \frac{S - n(0.5)}{\sqrt{n(0.25)}} \quad \text{assuming } H_0 \text{ is true}$$

The p -values then follow upon geometric inspection of the standard normal curve:

P-values Under Normal Approximation

1. $H_A : \theta_m > \theta_m^o \implies \text{p-value} = P(Z > z)$
2. $H_A : \theta_m < \theta_m^o \implies \text{p-value} = P(Z < z)$
3. $H_A : \theta_m \neq \theta_m^o \implies \text{p-value} = P(|Z| > z) = 2P(Z > |z|)$



Example 1.2.2 (Midterm Scores Cont.). *Let's use the normal approximation on the same problem. The hypotheses are:*

$$H_0 : \theta_m \geq 70 \quad H_A : \theta_m < 70$$

and the observed value of the test statistic is $b^+ = 5$. In this problem then, $n = 12$ and $s = 5$. Computing the z-statistic is then:

$$z = \frac{5 - (12)(0.5)}{\sqrt{(12)(0.25)}} = \frac{-1}{\sqrt{3}} \approx -0.577$$

We then end up with a p-value of

$$P(Z < -0.577) = 0.2810$$

Again, we fail to reject H_0 since $\alpha = 0.05$ which matches the binomial test.



In order to conduct this test we assume:

1. Random sample was taken
2. At least 5 observations are above and below the hypothesized value of the median

1.3 Lecture 3: Median CIs and Percentiles/C.D.F.s

1.3.1 Confidence Intervals for the Median

The sample median is what's known as an *order statistic*, it is a certain spot on an ordered set of data. More formally, we define order statistics as follows:

Definition 1.3.1 (Order Statistics). *Given a random sample of data X_1, X_2, \dots, X_n , we may reorder the values from greatest to least yielding a permuted data set $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ where we read $X_{(i)}$ as the i -th highest data entry. The entries $X_{(i)}$ are then called **order statistics**. Notice that $X_{(1)} = \text{minimum } X_i$ and $X_{(n)} = \text{maximum } X_i$. ♠*

In order to construct CIs for a true population parameter, we somehow isolate the parameter of interest as we do with the sample mean \bar{X} yielding $\mu \in [\alpha, \beta]$ with some degree of confidence. For the population median, however, this method of isolation called *pivoting* is not necessary. We can instead make some progress by considering the nature of the population median θ_m . Clearly, by definition, if we order our data, then each point has a 50% chance of lying above this median. If we further describe our data using order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, then we arrive at the following theorem adapted from [9]

Theorem 1.3.1 (Median CI). *For any distribution free sample the population median is approximately bounded above and below the order statistics $X_{(i)}$ and $X_{(j-1)}$ by degree of confidence $\sum_{k=i}^{j-1} \binom{n}{k} (1/2)^n$. That is:*

$$P(X_{(i)} \leq \theta_m \leq X_{(j-1)}) \approx \sum_{k=i}^{j-1} \binom{n}{k} (1/2)^n$$

Proof. We begin with a simple case, that is, solving for the following probability:

$$P(X_{(k)} \leq \theta_m \leq X_{(k+1)}) \tag{1.3.1}$$

which is the chance of the true median landing within any two adjacent order statistics. We will now rewrite equation (1.3.1) using conjunctions and simplify using the law of multiplication for events:

$$P(X_{(k)} \leq \theta_m \& \theta_m \leq X_{(k+1)}) = P(X_{(k)} \leq \theta_m | \theta_m \leq X_{(k+1)}) P(\theta_m \leq X_{(k+1)}) = P(E_1 | E_2) P(E_2) \tag{1.3.2}$$

Now, if we know E_2 happened, the chance for the order statistic immediately before it to be below θ_m does not change as it is still possible for $X_{(k)}$ to fall on either side of the relation between it and θ_m (we don't know what value $X_{(k+1)}$ actually takes, just that it's greater than θ_m), so the simplification is

$$P(E_1)P(E_2) = P(X_{(k)} \leq \theta_m)P(\theta_m \leq X_{(k+1)}) \quad (1.3.3)$$

If $X_{(k)} \leq \theta_m$, then $X_{(1)}, X_{(2)}, \dots, X_{(k)} \leq \theta_m$ and similarly if $X_{(k+1)} \geq \theta_m$, then $X_{(k+1)}, X_{(k+2)}, \dots, X_{(n)} \geq \theta_m$. Again, each order statistic has a 50% chance of being exclusively greater or less than θ_m so

$$P(X_{(i)} \leq \theta_m) = 1/2 \quad P(X_{(j)} \geq \theta_m) = 1/2$$

The last bit of information we need to simplify (1.3.3) is the fact the actual sample values of any order statistic are unknown (in shorthand: $X_{(i)}$ could be X_α where $\alpha \in \{1, \dots, n\}$), so we must take into account all combinations of data points which is $\binom{n}{k}$. Thus,

$$P(X_{(k)} \leq \theta_m \leq X_{(k+1)}) = \binom{n}{k} (1/2)^k (1/2)^{n-k} = \binom{n}{k} (1/2)^n$$

To generalize this probability in order to attain a higher confidence level, we need to find

$$P(X_{(i)} \leq \theta_m \leq X_{(j)})$$

for any $i < j$. Notice that when $X_{(i)} \leq \theta_m \leq X_{(j)}$ a direct implication is that $[X_{(i)} \leq \theta_m \leq X_{(i+1)}] \oplus [X_{(i+1)} \leq \theta_m \leq X_{(i+2)}] \oplus \dots \oplus [X_{(j-1)} \leq \theta_m \leq X_{(j)}]$. These events are all disjoint since the bounds do not intersect.

The probability then becomes

$$\begin{aligned} P(X_{(i)} \leq \theta_m \leq X_{(j)}) &= \sum_{k=i}^{j-1} P(X_{(k)} \leq \theta_m \leq X_{(k+1)}) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} (1/2)^k (1/2)^{n-k} = \sum_{k=i}^{j-1} \binom{n}{k} (1/2)^n \end{aligned}$$

Naturally, it follows that since $P(\theta_m = X_{(j)}) = 0$ (we draw from a continuous population) we have

$$P(X_{(i)} \leq \theta_m \leq X_{(j-1)}) \approx \sum_{k=i}^{j-1} \binom{n}{k} (1/2)^n$$

as we sought to show. ■

This theorem allows us to make CIs based on the binomial distribution as well as those using a normal approximation to the binomial.

CIs for Median

(i) Based on the Binomial Distribution

- For a CI, we want a roughly symmetric interval where

$$P(a < \theta_m < b - 1) = 1 - \alpha$$

for some lower bound a , and upper bound $(b - 1)$.

Note: Because the binomial distribution is discrete, the lower bound is a and the upper bound is $(b - 1)$ to adjust for the discrete nature of the data. Consider, $X \sim \text{Bin}(n, 1/2)$, then $P(X > a) \neq P(X \geq a)$.

This CI computation to find a and $(b - 1)$ is typically done via computer, and the bounds are found as **locations** where

$$\sum_{k=a}^{b-1} \binom{n}{k} (0.5)^n \approx 1 - \alpha \quad (\text{see Median CI})$$

where the computer starts with

- a = first location below median
- b = first location above median

and works outwards from there. At the end, it uses the ordered data point $X_{(i)}$ in the (a) th and $(b - 1)$ th location. In a sentence, the CI states **there is a $(1 - \alpha)100\%$ chance of the true population median being contained by the a th and $b - 1$ th order statistic.**

(ii) Using the Normal Approximation to the Binomial

- Now, we still want $P(a < \theta_m < b - 1) = 1 - \alpha$, i.e. $P(\theta_m < a) = \alpha/2$ and $P(\theta_m > b - 1) = \alpha/2$. Notice the the statement for the binomial case is about collecting areas from a binomial variate $X \sim \text{Bin}(n, 1/2)$. We can approximate this binomial variate using a normal curve yielding $a \approx -z_{1-\alpha/2}\sqrt{V(X)} + E(X)$ and $b - 1 \approx z_{1-\alpha/2}\sqrt{V(X)} + E(X)$. The **locations** of the bounds then become

$$a = \text{lower bound location} = -z_{1-\alpha/2}(\sqrt{0.25n}) + 0.5n$$

$$b = \text{upper bound location} = z_{1-\alpha/2}(\sqrt{0.25n}) + 0.5n + 1$$

rounded to the nearest integer.

Note: $z_{1-\alpha/2}$ is the $(1 - \alpha/2)100\%$ percentile of the Z (standard normal) distribution.



Example 1.3.1 (Height Data). Suppose we had the following data on height:

Height: 48, 48, 50, 52, 54, 54, 55, 56, 56, 57, 57, 57, 58, 58, 59, 60, 62, 62, 63, 71

(a) Find the 95% CI for the median using the normal approximation to the binomial.

- **Solution:** We use the same form as found out for the locations only set $n = 20$:

$$\text{lower bound location} = a = -1.96(\sqrt{0.25(20)}) + 5(20) \approx 5.617 \approx 6$$

$$\text{upper bound location} = b = +1.96(\sqrt{0.25(20)}) + 5(20) + 1 \approx 15.3826 \approx 15$$

So the CI will be $(X_{(6)}, X_{(15)})$ or (54, 59) with degree of confidence 95%.

(b) Does your interval suggest that the median is above 56?

- **Solution:** While there are values that are above 56 in the CI (e.g. 58), there are also values less than 56 (e.g. 55). We can only have valid evidence for $\theta_m > 56$ if and only if all points in the CI are above 56. In short, no, both bounds are not over 56.

(c) Interpret the CI in terms of the problem.

- **Solution:** We are 95% confident that the true median is between 54 and 59 inches.

Note: Some computer programs will extrapolate to an approximate value for a decimal location. This isn't a problem, just round the bounds to make sure the answer makes sense (i.e. the bounds are part of the data set given). ♥

1.3.2 Estimating Percentiles and C.D.F.s

Recall: The Cumulative Distribution Function or C.D.F (F) for discrete data is defined to be

$$F(x) = \sum_{t=\min\{X\}}^x P(X = t) = P(X \leq x)$$



The empirical CDF \hat{F} is an estimate of $F(x)$ where

$$\hat{F}(x) \equiv (\# \text{ of } x_i \leq x)/n$$

Example 1.3.2 (Height Data cont.). Here's a reproduction of the data:

Height: 48, 48, 50, 52, 54, 54, 55, 56, 56, 57, 57, 57, 58, 58, 59, 60, 62, 62, 63, 71

Given the data we have, we can construct an empirical C.D.F. For a single data point, say 48, we know that there are exactly 2 numbers that are below or equal to 48. So, $\hat{F}(48) = 2/n$. The rest of the points are calculated below in a table:

x	48	50	52	54	55	56	57	58	59	60	62	63	71
$\hat{F}(x)$	2/20	3/20	4/20	6/20	7/20	8/20	12/20	14/20	15/20	16/20	18/20	19/20	20/20



As a random variable $\hat{F}(x)$ for a given value of x follows a binomial distribution divided by the sample size n . This is because we can write $\hat{F}(x)$ as

$$\hat{F}(x) = \frac{\mathbf{1}(X_i < x)}{n} = \frac{\sum_i^n Y_i}{n}$$

where each Y_i is a bernoulli trial indicating whether the data point is below x or not. We have $E(Y_i) = F(x)$ since $F(x)$ is the true value of proportion of data below x . Thus $\hat{F}(x)$ is an average of bernoulli trials or $\hat{F}(x) \sim \text{Bin}(n)/n$ and $E(\hat{F}(x)) = F(x)$, $V(x) = (F(x)(1 - F(x)))/n$.

If the sample size is large enough, then we can use the CLT to approximate the this distribution and give

$$\hat{F}(x) \sim N(F(x), \sqrt{(F(x)(1 - F(x)))/n})$$

Since $\hat{F}(x)$ and $F(x)$ are proportions, we can see them as \hat{p} and p respectively. The form of the above statement is then

$$\hat{p} \sim N(p, \sqrt{(p(1 - p))/n})$$

if one prefers.

With this approximation we also have $(1 - \alpha)100\%$ CI for $F(x)$:

$$\hat{F}(x) \pm z_{1-\alpha/2} \sqrt{\hat{F}(x)(1 - \hat{F}(x))/n}$$

Example 1.3.3 (Height Data cont.). Find a 95% CI for the C.D.F. at $x = 60$

- **Solution:** Looking at the CDF table, we see $\hat{F}(60) = 16/20 = 0.8$. The Z quantile for a 95% CI is also $z_{0.975} = 1.96$. This then makes the CI

$$95\% \text{ CI: } 0.8 \pm 1.96\sqrt{0.8(0.2)/20} \implies (0.624, 0.975)$$

Thus we estimate that $x = 60$ could be anything from the 62.4th and 97.5th percentile. ♥

Note: Ideally For a CDF $F(x)$ we would have at least 5 values above and below. But this may not happen for values of x that are close to the Min or Max. ▲

Chapter 2

Week 2: 2-Sample Tests

2.1 Lecture 4: CIs for Percentiles & Permutation Tests

2.1.1 Confidence Intervals For Percentiles

The CI for a median can be easily modified to be a CI for any percentile. Recall the CI for the median using the normal approximation for the binomial was

$$a = \text{lower bound location} = -z_{1-\alpha/2} \left(\sqrt{0.25n} \right) + 0.5n$$

$$b = \text{upper bound location} = z_{1-\alpha/2} \left(\sqrt{0.25n} \right) + 0.5n + 1$$

We note that $0.5n = E(X)$ and $\sqrt{0.25n} = \sqrt{V(X)}$. When we no longer speak of the median but rather any percentile $p_\beta^* \forall \beta \in [0, 1]$ we have $E(p_\beta^*) = n\beta$ and $\sqrt{V(p_\beta^*)} = \sqrt{p_\beta^*(1-p_\beta^*)n}$. A very similar result for p_β^* 's confidence interval can thus be obtained and a $(1-\alpha)$ 100% CI for the (p_β^*) 100th percentile is:

$$\text{lower bound location: } -z_{1-\alpha/2} \left(\sqrt{p_\beta^*(1-p_\beta^*)n} \right) + n\beta$$

$$\text{upper bound location: } z_{1-\alpha/2} \left(\sqrt{p_\beta^*(1-p_\beta^*)n} \right) + n\beta + 1$$

These confidence intervals are functions of **confidence level, percentile, and sample size**. For shorthand, $CI = CI(\alpha, \beta, n)$. ▲

Example 2.1.1. If $n = 30$ and find a 99% CI for the 25th percentile.

- **Solution:** We have $\alpha = 1, \beta = .25, n = 30$. Thus, the locations would be:

$$\text{lower bound location: } -2.575(\sqrt{.25(.75)30}) + 30(0.25) \approx 1.393 \implies 1\text{st}$$

upper bound location: $2.575(\sqrt{.25(.75)30}) + 30(0.25) + 1 \approx 14.607 \implies 14\text{th}$
 So the ordered values in the 1st and 14th spot are $X_{(1)}$ and $X_{(14)}$.



Note: It is possible for you to round a location to 0 or to $n + 1$. In this case, use 1 or n instead.



2.1.2 Comparing two Means

The goal is typically to determine if two means are statistically different (i.e. if $\mu_1 \neq \mu_2$ by data we collect). The typical assumptions for the **parametric test** are

1. Random sample from both groups
2. Groups are independent
3. \bar{X}_1 and \bar{X}_2 are normal, either through...
 - (i) $n_i \geq 30 \quad \forall i \in \{1, 2\}$ (CLT)
 - (ii) Both populations normal

2.1.3 Permutation Test for Two Groups

The idea behind most permutation tests is that we assume that the distribution of the two groups is identical. If that were true each observation should be equally likely to come from either group. Then, we create all possible two group samples possible (ideally). Sometimes data sets are very large so all possible permutations take up more computing resources and we cannot use non-parametric methods. Just a caveat.

Example 2.1.2 (Small Group Permutation). *Suppose our two samples are:*

Group 1: 2, 4, 6

Group 2: 5, 7, 9

Clearly with this sample there is an observed difference of the sample means. We reference this idea with D^{obs} such that $D^{obs} = \bar{x}_1 - \bar{x}_2$. If we then set $D_i =$ difference in means for i th permutation we can then compute and organize all of the differences when we permute. Now, let's assemble all possible groups we could have had:

Group 1	Group 2	$D_i = \bar{x}_1 - \bar{x}_2$
(2, 4, 6)	(5, 7, 9)	$4 - 7 = 3$
(2, 4, 5)	(6, 7, 9)	$3.67 - 7.33 = -3.66$
(2, 4, 7)	(5, 6, 9)	$4.33 - 6.67 = -2.34$
(2, 4, 9)	(5, 7, 6)	$5 - 6 = -1$
(5, 4, 6)	(2, 7, 9)	$5 - 6 = -1$
(7, 4, 6)	(5, 2, 9)	$5.67 - 5.33 = 0.34$
(9, 4, 6)	(5, 7, 2)	$6.33 - 4.67 = 1.66$
(2, 5, 6)	(4, 7, 9)	$4.33 - 6.67 = -2.34$
(2, 7, 6)	(5, 4, 9)	$5 - 6 = -1$
(2, 9, 6)	(5, 7, 4)	$5.67 - 5.33 = 0.34$
(2, 5, 7)	(4, 6, 9)	$4.67 - 6.33 = -1.66$
(2, 5, 9)	(4, 6, 7)	$5.33 - 5.67 = -0.34$
(2, 7, 9)	(4, 6, 5)	$6 - 5 = 1$
(4, 5, 7)	(2, 6, 9)	$5.33 - 5.67 = -0.34$
(4, 5, 9)	(2, 6, 7)	$6 - 5 = 1$
(4, 7, 9)	(2, 6, 5)	$4.33 - 6.67 = -2.34$
(6, 7, 9)	(2, 4, 5)	$7.33 - 3.67 = 3.66$
(6, 5, 9)	(2, 4, 7)	$6.67 - 4.33 = 2.34$
(6, 5, 7)	(2, 4, 9)	$6 - 5 = 1$
(5, 7, 9)	(2, 4, 6)	$7 - 4 = 3$

Notice that we only care about the difference of the means; we do not care about the order of the observations in the groups (2, 4, 6 has same effect as 6, 4, 2 as far as our tests are concerned).

Now, using the above we can create the **empirical discrete distribution** of possible differences in means:

D_i	-3.66	-3	-2.34	-1.66	-1	-0.34	0.34	1	1.66	2.34	3	3.66
Freq	1/20	1/20	2/20	1/20	3/20	2/20	2/20	3/20	1/20	2/20	1/20	1/20

Now, the observed mean difference is $D^{obs} = -3$. If we are considering "extreme" to be as or more negative, we then have:

p-value: If all observations are equally likely to be in either group (the distributions are equal), then we would observe our data or more extreme with probability 2/20. ♥

Note: Permutation tests are usually done in R (hard to do by hand; too many permutations). ▲

Procedure for Conducting Permutation Tests

Notice that we had to assume the distributions are the same. This would mean the means, standard deviations, minimums, maximums, etc... are all equal between groups. This is in

effect the null hypothesis. In order to conduct a permutation hypothesis test we follow the following steps:

Permutation HT Steps

Step 1: State H_0 and H_A

- Let $F_1(x)$ = CDF for group 1 and $F_2(x)$ = CDF for group 2. Assuming the distributions for both groups are the same, it must be the case that $F_1(x) = F_2(x)$ for all x in the data's domain. Thus, possible hypotheses are

H_0	H_A
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$

Where the inequality is valid for at least one x in every alternate hypothesis.

Notice that¹

- $F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \implies$ Distributions are different
- $F_1(x) \leq F_2(x) \implies$ Group 1 tends to be larger than Group 2, i.e. ($E(X_1) > E(X_2)$)
- $F_1(x) \geq F_2(x) \implies$ Group 2 tends to be larger than Group 1, i.e. ($E(X_1) < E(X_2)$)

Step 2: Calculate the observed statistic and all permutations.

- The observed statistic could be any number of things such as: $D^{\text{obs}} = \bar{x}_1 - \bar{x}_2$, $D^{\text{obs}} = \text{total}_1 - \text{total}_2$, or $D^{\text{obs}} = \text{median}_1 - \text{median}_2$. If we set m = sample size of group 1 and n = sample size of group 2, then there are

$$\binom{m+n}{m} = \binom{m+n}{n} = \frac{(m+n)!}{m!n!}$$

total possible permutations.

¹For a detailed proof of this, see the [Appendix](#)

Step 3: Calculate the permutation p-value. For each hypothesis, the p-value is:

Alternate Hyp.	p-value
$H_A : F_1(x) \leq F_2(x)$	$(\# \text{ of } D_i \geq D^{\text{obs}}) / \binom{m+n}{n}$
$H_A : F_1(x) \geq F_2(x)$	$(\# \text{ of } D_i \leq D^{\text{obs}}) / \binom{m+n}{n}$
$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$	$(\# \text{ of } D_i \geq D^{\text{obs}}) / \binom{m+n}{n}$

Step 4: Reject H_0 if p-value $\leq \alpha$



Note: Typically when we have asymmetric distributions, we use the median to compare outliers. When we have a roughly symmetric distribution, we use the total or mean. So, the choice of D^{obs} does not always have to involve the mean. ▲

2.2 Lecture 5: Permutation Tests (cont.) & WRS Test

2.2.1 Approximate Permutation Test

When the sample sizes get moderately large, all permutations can be difficult to calculate (or code). When this happens, instead of calculating literally all permutations we randomly generate permutations. For example, if $n = m = 10$, then $\binom{n+m}{n} = \binom{20}{10} = 184,756$ total permuted groups. This would take a computer some time to make all of the permutations and certainly for us even longer. Hence, the need to randomly generate these permutations.

Note: Randomly generating permutations will give an *approximate p-value* rather than the true permutation p-value. This thus makes it a **random variable** and subject to error. More permutations usually minimize this error. ▲

The steps for an approximate permutation test are given below:

Steps for an Approximate Permutation Test (for coding):

1. Record D^{obs}
2. Create one vector of all observations, \vec{q}
3. Randomly shuffle the $(m + n)$ observations, and assign first m to group 1 last n to group 2
4. Compute $D_i =$ observed difference in (means/medians/totals)
5. Repeat steps 3 & 4 **R > 2000 times**
6. Based on these R random values of D_i , we have an approximate permutation distribution. Thus, our approximate p-values are:

Alternate Hyp.	p-value
$H_A : F_1(x) \leq F_2(x)$	$(\# \text{ of } D_i \geq D^{\text{obs}})/R$
$H_A : F_1(x) \geq F_2(x)$	$(\# \text{ of } D_i \leq D^{\text{obs}})/R$
$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$	$(\# \text{ of } D_i \geq D^{\text{obs}})/R$

7. If p-value $< \alpha$, then reject H_0



Remark 2.2.1. You will get code that I have made and you can modify it as needed. ◆

2.2.2 Confidence Interval for P-value

Since we are finding an approximate p-value which is a random quantity, we may want to estimate the p-value further. Notice that the calculated p-value \tilde{p} in all cases has the form:

$$\tilde{p} = \frac{\sum_i^R \mathbf{1}(\varphi(D_i) \odot \varphi(D^{obs}))}{R} \quad \odot \in \{\geq, \leq\}$$

for some transformation φ . As such, we can view \tilde{p} as the sum of R Bernoulli trials where the chance of success is the true p-value (think of each indication as '1' being **this is as or more extreme**, the true chance of as or more extreme is set as p^* , so each summand $\sim \text{Bernoulli}(p^*)$). It then follows that

$$\tilde{p} \sim \frac{\text{Bin}(R, p^*)}{R}$$

Then, a $(1 - \alpha)100\%$ CI for a p-value p^* is

CI for permutation p-value

$$p^* \pm z_{1-\alpha/2} \sqrt{p^*(1-p^*)/R}$$



Example 2.2.1 (Theoretical CI). Suppose based on $R = 2000$ random permutations the approximate permutation p-value was 0.0432. Find the 90% approximate p-value.

- **Solution:** The 90% CI for our approximate p-value is

$$0.04321 \pm 1.645 \sqrt{0.0432(1 - 0.432)/2000} \implies (0.0357, 0.0507)$$



2.2.3 Normal Approximation To Permutations

When we use a normal approximation to permutations we again assume that both groups come from the same population. To perform hypothesis tests, we use the overall mean \bar{x}^* and overall standard deviation s^* in our test statistic. Since both groups are from the same population (under H_0), $\bar{X}^* \rightarrow \mu^*$ and $S^*/\sqrt{m} \rightarrow \sigma_{\bar{X}^*}^*$ as more equal size samples are taken.

Remark 2.2.2. For this test we need $m + n \geq 30$



The steps to conduct the hypothesis test using the permutation normal approximation are

Permutation Normal Approximation HT

Step 1: State H_0 and H_A

- A table for all cases is given below:

H_0	H_A
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$

Step 2: Our test statistic

$$Z = \frac{\bar{X}_1 - \bar{x}^*}{s^*/\sqrt{m}} \sim N(0, 1) \text{ (approx.)}$$

Step 3: Calculate the p-value

- A table for each alternate hypothesis p-value is given

H_A	p-value
$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$	$P(Z > z)$
$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$	$P(Z < z)$
$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$	$2P(Z > z)$

Step 4: If p-value $< \alpha$, reject H_0



2.2.4 Wilcoxon Rank Sum (WRS) Test

A different approach to the same problem (difference between two groups) is instead of using the actual data values, to use the ranks of the data values instead. The main way this works is if one group has larger observations than the other, in general it will also have larger ranks than the other. **Ranks are directly proportional to the values of the data, lower data points, lower ranks and vice versa.** Before we begin further, let's define specifically what a rank is:

Definition 2.2.1 (Rank). The **rank** of any data point x_i with $m + n$ entries is

$$R(x_i) = \sum_{j=1}^{m+n} \mathbf{1}(x_j \leq x_i)$$

If any points have the same point value (not rank), then we take the average of their ranks and assign these numbers this rank. For example, if $x_i = x_j$ but $R(x_i) \neq R(x_j)$ upon first assignment, then $R(x_i) = (R(x_i) + R(x_j))/2$ and $R(x_j) = (R(x_i) + R(x_j))/2$



Remark 2.2.3. Remark: (WRS) tends to outperform permutation tests in certain situations which we will discuss later.



In order to conduct the test we follow these steps:

WRS HT Steps

Step 1: State the hypotheses (they remain the same as with permutation tests since we are after the same inference)

- The hypotheses are reproduced below

H_0	H_A
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$

Step 2: Our test statistic requires the following steps:

- Combine the $(m + n)$ values into one group/vector: \vec{q}
- Calculate the **rank** for each data point
- Calculate the **total** rank in group 1 (arbitrary choice of groups). This is our test statistic,

$$W_{\text{obs}} = \sum_{\text{group 1}} R(x_i)$$

Step 3: To find The exact p-value you would calculate all $\binom{m+n}{n}$ permutations of the two groups and calculate the distribution of

$$W_i = \text{sum of ranks in group 1}$$

Then,

Alternate Hyp.

$$H_A : F_1(x) \leq F_2(x)$$

$$H_A : F_1(x) \geq F_2(x)$$

$$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$$

p-value

$$(\# \text{ of } W_i \geq W_{\text{obs}}) / \binom{n+m}{n}$$

$$(\# \text{ of } W_i \leq W^{\text{obs}}) / \binom{n+m}{n}$$

$$2 \min\{\text{both one-sided p-values}\}$$

Step 4: If p-value $< \alpha$, then reject H_0



Note: WRS tends to have higher power when the distribution is skewed and outliers are present since assigning ranks essentially removes all influence of both issues.

Permutation tests, however, tend to have higher power when the distribution is thought to be symmetric and when using the mean as a measure of central tendency. ▲

2.3 Lecture 6: WRS (cont.) & Approximations

Let's see the WRS test in an example:

Example 2.3.1 (Exam Scores). Suppose that exam scores for math majors and computer science majors in a statistics course were:

Math: 80, 85

Computer Science: 75, 80, 90

Assume we want to test if the groups means are different, i.e. $\mu_{\text{math}} \neq \mu_{\text{comp}}$.

(a) State H_0 and H_A

- **Solution:** Since this test is about mean inequality (no knowledge about which direction the inequality is), the null and alternative are

$$H_0 : F_1(x) = F_2(x)$$

$$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$$

(b) Calculate all possible values of the rank sums for Group 1.

- **Solution:** A table giving the data, ordinal ranks, and adjusted ranks (one we use for test) is given below:

Observations:	75	80	80	85	90
Ordinal Ranks (R') :	1	2	3	4	5
R(x_i) :	1	2.5	2.5	4	5

The number of permutations is the number of ways we can group data into group 1 (math group): $\binom{5}{2} = 10$. Looking at the ordinal ranks, we can list all possible ranks for group 1 as follows

R' Group 1:	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(2, 4)
R' (cont.):	(3, 4)	(2, 5)	(3, 5)	(4, 5)	(2, 3)

If we define a mapping $\psi : R' \rightarrow R$ such that (for example) $\psi(2) = 2.5$, we can get the WRS rank permutations and associated rank sums (W_i):

$R(x_i)$	W_i	$R(x_i)$ (cont.)	W_i
(1, 2.5)	3.5	(2.5, 4)	6.5
(1, 2.5)	3.5	(2.5, 5)	7.5
(1, 4)	5	(2.5, 5)	7.5
(1, 5)	6	(4, 5)	9
(2.5, 4)	6.5	(2.5, 2.5)	5

(c) Calculate the WRS test statistic and the appropriate p -value

- **Solution:**

$$W_{obs} = 6.5$$

$$p\text{-value: } 2(\#W_i \geq W_{obs} = 6.5)/10 = 2(5)/10 = 1 > \alpha = 0.05$$

(d) State your conclusion in terms of the problem.

- **Solution:** We fail to reject H_0 and conclude we may support that the distributions (or means, medians) are similar.



Remark 2.3.1. As with most hypothesis tests the smaller the sample the more difficult it will be to reject the null (the measure of central tendencies are equal). This is because of the high variability of small samples. Small samples also have ranks that reveal no actual information about the degree the values observed are from each other, so even if the values are extremely high, we still fail to reject the null. For example, suppose the data looked like:

Math: 20, 30

Computer Science: 10, 20, 100

we will have the same test result as before (ranks the same). Notice this test takes away the effect of outliers whilst keeping the sample size the same. **As always, it is easier to make an inference with more data. Then the distribution of the ranks is apparent.** ♦

2.3.1 Large Sample Approximation to WRS

The WRS test can be approximated with large enough sample sizes. Before discussing it, let's introduce some notation. Let

$N = m + n$, and $R(x_1), \dots, R(x_N)$ be the corresponding combined ranks of the two groups

Also, let $S_1 =$ sum of ranks for group 1.

Under the assumption that the distributions are equal every rank $R(x_i)$ should have been equally likely to come from both groups. If we haven't observed the data yet, then $R(x_i)$ are random variables with their own expectations and variances. It is as if we are drawing numbers $\{1, 2, \dots, N\}$ from a bag without replacement, then adjusting for the rank for our tests.

It can be shown that²

Facts about Rank Sums

$$E(S_1) = m\mu_R = \frac{m(N+1)}{2}$$

$$\sigma_S^2 = V(S_1) = \frac{mn\sigma_R^2}{N-1}$$

where $\mu_R = \frac{1}{N} \sum_i R(x_i)$ and $\sigma_R^2 = \frac{1}{N} \sum_i (R(x_i) - \mu_R)^2$. Then, if $N \geq 30$, we have

$$S_1 \sim N\left(m\mu_R, \frac{mn\sigma_R^2}{N-1}\right)$$

Note: if there are no ties, then

$$\sigma_R^2 = \frac{mn(N+1)}{2}$$



To conduct a hypothesis test using this approximation, we follow these steps:

WRS Normal Approximation HT Steps

Step 1: State the null and alternative. For each different experimental setup we have:

H_0	H_A
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \geq F_2(x)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$

Step 2: Our test statistic is

$$Z = \frac{(S_1)_{\text{obs}} - m\mu_R}{\sigma_R}$$

Step 3: Calculate the p-value, this follows the same form as with normal approximation to the permutation test.

- A table for each alternate hypothesis p-value is given

²see the [Appendix](#) for proofs

H_A	p-value
$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$	$P(Z > z)$
$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$	$P(Z < z)$
$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$	$2P(Z > z)$

Step 4: If p-value $< \alpha$, reject H_0



Example 2.3.2 (Grad Exam Scores). Two years of graduate students exam scores for the pre qualifying exam were compared and the data is as follows:

	Mean Rank	Std. Dev of Rank	Sample Size
Year 1 (Group 1)	14.86	8.480	15
Year 2 (Group 2)	16.13	8.771	15
Overall	15.5	8.630	30

The department **believes year 2 scored significantly higher than year 1**. Perform a hypothesis test to find out if this is plausible.

(a) State H_0 and H_A

- **Solution:** $H_0 : F_{yr\ 1}(x) = F_{yr\ 2}(x)$ and $H_A : F_{yr\ 1}(x) \geq F_{yr\ 2}(x) \quad (\mu_{yr\ 1} < \mu_{yr\ 2})$

(b) Calculate the appropriate test statistic and p-value using the large sample approximation

- **Solution:** The quantities we need in order to compute the Z statistic are the rank sum $(S_1)_{obs}$, mean of rank sum $E(S_1)$, and standard deviation of rank sum σ_S . We compute them below:

$$\mu_S = m\mu_R = (15)(15.5) = 232.5$$

$$\sigma_S^2 = mn\sigma_R^2/(N-1) = (15)(15)(8.630^2)/(30-1) \approx 577.838$$

$$\Rightarrow \sigma_S \approx \sqrt{577.838} \approx 24.0383$$

Hence,

$$W_{obs} = \left(\text{mean} = \frac{1}{m} \sum R(x_i) \right) (\text{sample size} = m) = (14.86)(15) \approx 222.9$$

$$\Rightarrow Z = (222.9 - 232.5)/24.0383 = -0.399$$

$$\Rightarrow \text{p-value} = P(Z < -0.399) \approx 0.3449$$

(c) State your decision in terms of the problem

- **Solution:** Since $p\text{-value} > \alpha$ fail to reject H_0 . We cannot conclude that there is a significant difference in the scores for the two years.

Note: when the $p\text{-value}$ is very large or very small we do not need to specify $\alpha = 0.10, 0.05, 0.01$ ♥

Now, there is no immediate CI that is associated with WRS. So, we can't tell how large (extreme) a difference between population means are. But, there is another technique that has equivalent results for the HT and also has a CI with it.

This is the Mann Whitney (MW) test. It has a very different structure but will yield the same results as the WRS, albeit with different assumptions.

As far as which to use either one, but MW has a natural extension to CIs so we can get more information about the groups should the null be false.

2.4 Appendix (Week 2)

We now prove stochastic dominance of one CDF over another implies the same relation with respect to the population means of the two distributions. In other words, we show $F_1(x) \leq F_2(x) \Rightarrow E(X_1) > E(X_2)$ and $F_1(x) \geq F_2(x) \Rightarrow E(X_1) < E(X_2)$ assuming the inequality holds for all x . We show a proof using continuous density functions as the underlying distributions are assumed to be continuous. All credit for this proof goes to Shitong Wei who introduced me to it during discussion in college.

Theorem 2.4.1 (Stochastic Dominance). *For continuous random variables X_1 and X_2*

1. $F_1(x) \leq F_2(x) \Rightarrow E(X_1) > E(X_2)$
2. $F_1(x) \geq F_2(x) \Rightarrow E(X_1) < E(X_2)$

Proof. We prove the case (1), but case (2) is shown by permutation of inequality symbols. Suppose there is a number γ such that $F_1(x) \leq F_2(x)$ for all $x \in [\gamma, \infty)$. Then, we have

$$\int_{-\infty}^c F_1(x) dx \leq \int_{-\infty}^c F_2(x) dx \quad (2.4.1)$$

for all $c \in [\gamma, \infty)$. Using the definition of the CDF for any random variable we simultaneously have

$$F_j(x) = \int_{-\infty}^x f_j(t) dt \quad \forall j \in \{1, 2\} \quad (2.4.2)$$

This lets us rewrite equation (2.4.2) as

$$\int_{-\infty}^c \int_{-\infty}^x f_1(t) dt dx \leq \int_{-\infty}^c \int_{-\infty}^x f_2(t) dt dx. \quad (2.4.3)$$

Now, the region $\mathcal{R} = (-\infty, c) \times (-\infty, x)$ looks like a **triangle** made from the upper left part of a square (we treat $(-\infty, \infty)$ like a point, it's the bottom left corner of the square $\mathcal{S} = (-\infty, c) \times (-\infty, c)$). This allows us to make the change of variables we would normally do for triangles in calculus. Thus we can write the region of integration as $\mathcal{R} = (-\infty, c) \times (t, c)$ and equation (2.4.3) becomes

$$\int_{-\infty}^c \int_t^c f_1(t) dx dt \leq \int_{-\infty}^c \int_t^c f_2(t) dx dt \quad (2.4.4)$$

Evaluating the integrals (the first is free of x so we can just take the difference of the upper and lower limits) we arrive at:

$$\int_{-\infty}^c (c-t)f_1(t)dt \leq \int_{-\infty}^c (c-t)f_2(t)dt \quad (2.4.5)$$

which simplifies to:

$$\int_{-\infty}^c cf_1(t)dt - \int_{-\infty}^c tf_1(t)dt \leq \int_{-\infty}^c cf_2(t)dt - \int_{-\infty}^c tf_2(t)dt \quad (2.4.6)$$

At this point (2.4.6) is valid for all real numbers c , and we will apply a limit to c to the infinite on both sides. Clearly,

$$\lim_{c \rightarrow \infty} \int_{-\infty}^c cf_j(t)dt = \infty \quad \forall j \in \{1, 2\} \quad (2.4.7)$$

But what's interesting is that the evaluation is **the same number** for both f_1 and f_2 since they both have area 1 under their curves over infinity (take the ratio between the equations described by (2.4.7) and see that they evaluate to 1; same number over infinity). So, we can (as $c \rightarrow \infty$) remove both from the equality when we take the limit. This leaves the limit as

$$-\int_{-\infty}^{\infty} tf_1(t)dt \leq -\int_{-\infty}^{\infty} tf_2(t)dt \quad (2.4.8)$$

which can be simplified into

$$\int_{-\infty}^{\infty} tf_1(t)dt \geq \int_{-\infty}^{\infty} tf_2(t)dt \quad (2.4.9)$$

and by the definition of expected value this is equivalent to saying

$$E(X_1) \geq E(X_2) \quad (2.4.10)$$

as we sought to show. This concludes the proof. ■

We will now prove some properties of ranks and S_1 using the adjusted ranks, i.e. $R(x_i) = \psi(R'(x_i)) = \psi(\text{ordinal rank})$. Keep in mind though the theorems are much easier to prove using ordinal ranks (set $\psi(R'(x_i)) = R(x_i)$).

Theorem 2.4.2 (Mean of Rank Sums). *We have*

$$E(S_1) = m\mu_R$$

where μ_R is the average of the ranks, i.e. $\mu_R = (N+1)/2$.

Proof. The expectation is linear, so we don't worry about rank dependence:

$$E(S_1) = E\left(\sum_{\text{group 1}} \psi(R'(x_i))\right) = \sum_{\text{group 1}} E(\psi(R'(x_i))) = m\mu_R$$

Now $\mu_R = \frac{1}{N} \sum_i \psi(R'(x_i)) = (1+2+\dots+N)/N = (N+1)/2$. Note that the sum of adjusted ranks always equals the sum of the ordinals used to make the adjusted rank. This makes the form of the mean sound. ■

Theorem 2.4.3 (Rank Covariance). *The covariance between any two adjusted ranks $\psi(R'(x_i))$, $\psi(R'(x_j))$ is:*

$$\text{Cov}(\psi(R'(x_i)), \psi(R'(x_j))) = -\frac{\sigma_R^2}{N-1}$$

Proof. Note: for shorthand, we use $\psi(R'(x_i))$ as $\psi(R'_i)$. We use the definition of covariance to yield

$$\text{Cov}(\psi(R'_i), \psi(R'_j)) = E(\psi(R'_i)\psi(R'_j)) - E(\psi(R'_i))E(\psi(R'_j)) \quad (2.4.11)$$

Now, the average rank is always the same regardless of the specific rank in question so the product of the expectations is:

$$E(\psi(R'_i))E(\psi(R'_j)) = \left(\frac{N+1}{2}\right)^2 = [E(\psi(R'_i))]^2 \quad (2.4.12)$$

Now we work on the other part of (2.4.11), by the law of total expectation we

have:

$$\begin{aligned}
E[\psi(R'_i)\psi(R'_j)] &= \sum_k E[\psi(R'_i)\psi(k)|\psi(R'_j) = \psi(k)]P[\psi(R'_j) = \psi(k)] \\
&= \frac{1}{N} \sum_k \psi(k)E(\psi(R'_i)|\psi(R'_j) = \psi(k)) \\
&= \frac{1}{N(N-1)} \sum_k \psi(k) \left[\sum_l \psi(l) - \psi(k) \right] \\
&= \frac{1}{N(N-1)} \sum_k \left[\psi(k) \frac{N(N+1)}{2} - \psi(k)^2 \right] \tag{2.4.13} \\
&= \frac{1}{N(N-1)} \left[\left(\frac{N(N+1)}{2} \right)^2 - \sum_k \psi(k)^2 \right] \\
&= \frac{N}{N-1} \left(\frac{N+1}{2} \right)^2 - \frac{E(\psi(R'_i)^2)}{N-1} \\
&= \frac{N}{N-1} [E(\psi(R'_i))]^2 - \frac{E(\psi(R'_i)^2)}{N-1}
\end{aligned}$$

Putting (2.4.12) and (2.4.13) together gives a neat simplification:

$$\begin{aligned}
&= \left(\frac{N}{N-1} - 1 \right) [E(\psi(R'_i))]^2 - \frac{E(\psi(R'_i)^2)}{N-1} \\
&= \frac{[E(\psi(R'_i))]^2 - E(\psi(R'_i)^2)}{N-1} \tag{2.4.14} \\
&= -\frac{\sigma_R^2}{N-1}
\end{aligned}$$

as we sought to show. ■

Theorem 2.4.4 (Variance of Rank Sums). *We have*

$$V(S_1) = \frac{mn\sigma_R^2}{N-1}$$

Where $\sigma_R^2 = E(\psi(R'(x_i))^2) - [E(\psi(R'(x_i)))]^2 = \frac{1}{N} \sum_i (\psi(R'(x_i)) - \mu_R)^2$

Proof. Since there is a covariance between the adjusted ranks (see [Rank Covariance](#)) we use

the definition of any sum of variances:

$$\begin{aligned}
 V(S_1) &= \sum_{\text{group1}} V(\psi(R'(x_i))) + \sum_{i \neq j} \text{Cov}(\psi(R'(x_i)), \psi(R'(x_j))) \\
 &= m\sigma_R^2 - m(m-1) \frac{\sigma_R^2}{N-1} \\
 &= \sigma_R^2 \left(m - \frac{m^2 - m}{N-1} \right) \\
 &= \sigma_R^2 \left(\frac{mN - m - m^2 + m}{N-1} \right) \\
 &= \sigma_R^2 \frac{mn}{N-1} = \frac{mn\sigma_R^2}{N-1} \quad (N - m = n)
 \end{aligned}$$

as we sought to show. ■

Proposition 2.4.1 (Rank Variance). *If there are no repeats in the data, i.e. we use ordinal ranks, then the rank sum variance σ_S^2 can be computed as*

$$\sigma_S^2 = \frac{mn(N+1)}{12}$$

Proof. Notice the new simplified form of the rank variance σ_R^2 since the data has ordinal ranks only:

$$\sigma_R^2 = E(R(x)^2) - [E(R(x))]^2 \quad (2.4.15)$$

$$= \frac{1}{N} \sum_{k=1}^N k^2 - \left[\frac{1}{N} \sum_{k=1}^N k \right]^2 \quad (2.4.16)$$

$$= \frac{1}{N} \left[\frac{N(N+1)(2N+1)}{6} \right] - \left[\frac{(N+1)}{2} \right]^2 \quad (2.4.17)$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} \quad (2.4.18)$$

$$= \frac{(N+1)[2(2N+1) - 3(N+1)]}{12} \quad (2.4.19)$$

$$= \frac{(N+1)(N-1)}{12} \quad (2.4.20)$$

Using **Variance of Rank Sums** we can readily compute the variance of the rank sum as

$$\sigma_S^2 = \frac{mn(N+1)(N-1)}{12(N-1)} = \frac{mn(N+1)}{12}$$

as we sought to show. ■

Chapter 3

Week 3: More 2-Sample Tests & Comparisons

3.1 Lecture 7: Mann-Whitney Test

3.1.1 Mann-Whitney Test

The Mann-Whitney (MW) test is another form of the WRS test under the strict assumption that the distributions of the two samples have the **same shape** [3]. This feature allows us to make CIs about the space between the two distributions. What makes the MW test work is **its ability to count how many observations from one group are below each observation from the other group**. We define the setup for test as follows:

Let X_1, \dots, X_m be our sample from group 1, **they are iid**

Let Y_1, \dots, Y_n be our sample from group 2, **they are iid with the same shape as group 1's distn's**

The steps for conducting a hypothesis test with this method are

MW Test Steps

Step 1: State H_0 and H_A . As usual...(same questions to answer)

H_0	H_A
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \quad (\mu_1 \geq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \geq F_2(x) \quad (\mu_1 \leq \mu_2)$
$H_0 : F_1(x) = F_2(x)$	$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x) \quad (\mu_1 \neq \mu_2)$

Step 2: Calculate test statistic

- Our test statistic is²

$$U_{MW} = \sum_i \sum_j \mathbf{1}(X_i < Y_j) + \frac{1}{2} \sum_i \sum_j \mathbf{1}(X_i = Y_j)$$

where

$$\mathbf{1}(X_i < Y_j) = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{o.w.} \end{cases} \quad \text{and} \quad \mathbf{1}(X_i = Y_j) = \begin{cases} 1 & \text{if } X_i = Y_j \\ 0 & \text{o.w.} \end{cases}$$

i.e., $U_{MW} = (\# \text{ of pairs } (X_i, Y_j) \text{ where } X_i < Y_j) + \frac{1}{2}(\# \text{ of pairs where } X_i = Y_j)$

Note: if group 1 is lower than group 2, U_{MW} will be close to the maximum number of pairs. If group 1 is larger than group 2, U_{MW} will be close to 0. Note the number of possible pairings is mn .

Step 3: Calculate the p-value

- Here the test statistic has its own known distribution the **U** or Mann Whitney distribution. "Tail" percentiles of this distribution are found in table A4 online or in R.

– Let $U_{1-\alpha/2} = (1 - \alpha/2)100\text{th percentile of } U$

– Let $U_{\alpha/2} = (\alpha/2)100\text{th percentile of } U$

Then the p-values for this test are ranges and are based off the alternative hypotheses (remember if group 1 is higher than group 2, then U_{MW} is low, but if group 1 is lower than group 2, then U_{MW} is high):

Hypothesis	Comp.	p-value
$H_A : F_1(x) \leq F_2(x)$	If $U_{MW} < U_{\alpha/2}$	$< \alpha/2$
$H_A : F_1(x) \geq F_2(x)$	If $U_{MW} > U_{1-\alpha/2}$	$< \alpha/2$
$H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$	If $U_{MW} < U_{\alpha/2} \text{ or } U_{MW} > U_{1-\alpha/2}$	$< \alpha$

Step 4: If p-value $< \alpha$, reject H_0



Example 3.1.1 (Gorilla Weights). *The weight of gorillas of the same age in two zoos are:*

Zoo 1: 145, 155, 170, 180

Zoo 2: 130, 160, 165, 170

²See the [Appendix](#) for how this relates to the WRS test.

We want to test if gorillas in zoo 1 weigh more in general.

(a) State H_0 and H_A

- **Solution:** $H_0 : F_1(x) = F_2(x)$ vs. $H_A : F_1(x) \leq F_2(x)$ ($\mu_1 \geq \mu_2$)

(b) List all possible pairs of observations (X_i, Y_j)

- **Solution:** There are $mn = (4)(4) = 16$ total possible pairs; we list them below:

(145, 130)	(155, 130)	(170, 130)	(180, 130)
•(145, 160)	•(155, 160)	(170, 160)	(180, 160)
•(145, 165)	•(155, 165)	(170, 165)	(180, 165)
•(145, 170)	•(155, 170)	*(170, 170)	(180, 170)

where (•) means $(X_i < Y_j)$ and (*) means $(X_i = Y_j)$

(c) Calculate the MW test statistic and the appropriate p -value

- **Solution:** The MW test statistic is

$$\begin{aligned}
 U_{MW} &= (\# \text{ of pairs}(X_i < Y_j)) + \frac{1}{2} (\# \text{ of pairs}(X_i = Y_j)) \\
 &= 6 + \frac{1}{2}(1) = 6.5
 \end{aligned}$$

In table A4 we have at $\alpha = 0.05$ or $\alpha = 0.10$

$$U_{\alpha/2} = \begin{cases} 0 & \text{if } \alpha = 0.05 \\ 1 & \text{if } \alpha = 0.10 \end{cases}$$

Thus, since $U_{MW} > 1$, p -value $> 0.10/2 = 0.05$

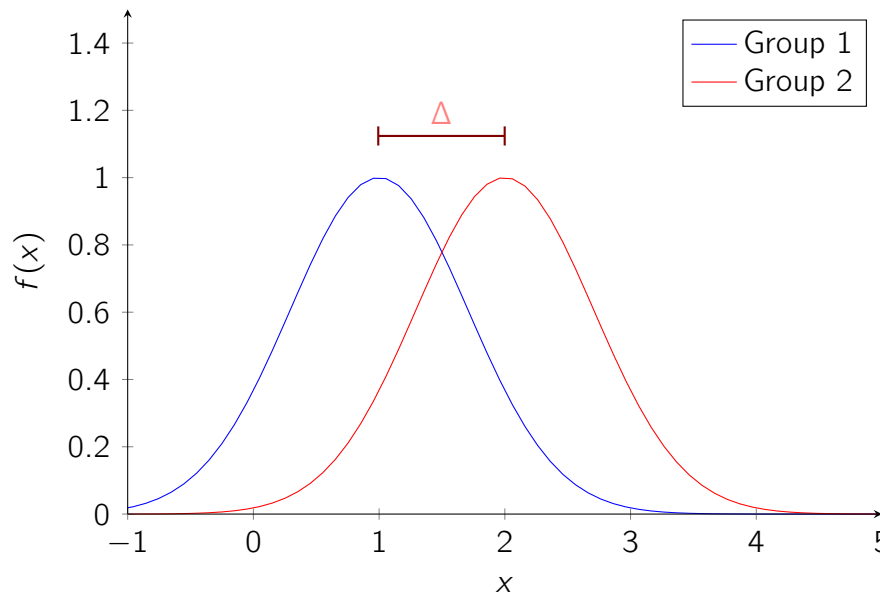
(d) Interpret the p -value in terms of the problem

- **Solution:** If in reality the distribution of weights for gorillas was the same between the two zoos, we would observe our data (MW statistic) or more extreme more than 5% of the time. Hence, **we fail to reject H_0** and do not have evidence to declare the two populations (zoos) to be different in gorilla weights.



3.1.2 CI for Shift Parameter

Instead of looking for a difference in means, medians, or totals between the two groups; we still assume the distributions have the same shape but one distribution is shifted some distance Δ from the other, this is a key assumption for the MW test. A graph of this idea is given below:



We call " Δ " the "shift parameter." This means we assume (for the picture above)

$$F_1(x) = F_2(x + \Delta)$$

which means that...

$$P(X_1 \leq x_1) = F_1(x) = F_2(x + \Delta) = P(X_2 \leq x_1 + \Delta) = P(X_2 - \Delta \leq x_1)$$

i.e., $P(X_1 \leq x_1) = P(X_2 - \Delta \leq x_1)$, so that X_1 and $X_2 - \Delta$ have the same distribution. X_2 is **higher than X_1 on average, so we subtract by the shift parameter to get back X_1 from X_2 .**

To find the CI for Δ the shift parameter, we follow these steps:

CI for Shift Parameter Δ

Step 1: Find all nm pair-wise differences, $X_i - Y_j$. **Some will be positive and others will be negative.**

Step 2: Order the pairwise differences and call them $\text{pwd}(1), \dots, \text{pwd}(nm)$. Notice that these are **order statistics**.

Step 3: We want the locations, call them k_a and k_b such that

$$P(\text{pwd}(k_a) \leq \Delta \leq \text{pwd}(k_b)) = 1 - \alpha$$

Notice if we set $\text{pwd}(k_a) = O_{(k_a)}$ and $\text{pwd}(k_b) = O_{(k_b)}$, then the CI becomes

$$P(O_{(k_a)} \leq \Delta \leq O_{(k_b)}) = 1 - \alpha$$

Which is the same form as we had for the population median's CI. We can also state this relationship as

$$P(k_a - 1 \leq U \leq k_b) = 1 - \alpha$$

Since in the extreme case where $O_{(k_b)} < 0$, we have

$$O_{(k_a)} \leq \Delta \leq O_{(k_b)} \implies k_a \leq U \leq k_b$$

Even if $O_{(k_b)} \not< 0$, the implication still holds as the bounds will be smaller and we can always get the same bounds as above as they are bigger. We finally adjust for the discrete nature of data: $k_a \mapsto k_a - 1$. **Remember, the discrete nature of the data is the reason we use $k_a - 1$.**

Thus we can use table A4 to find $U_{\alpha/2}$ and $U_{1-\alpha/2}$ and our locations for the CI are: $k_a = U_{\alpha/2} + 1$ and $k_b = U_{1-\alpha/2}$.

Essentially, we are using percentiles of U to find the locations for the bounds.



Example 3.1.2 (Theoretical Example). *If $m = 5$, $n = 5$, and we want the 95% CI, we go to 2.5% ($\alpha/2$) and find the percentiles. They are $U_{\alpha/2} = 4$ and $U_{1-\alpha/2} = 21$. This makes the CI for Δ as*

$$k_a = U_{\alpha/2} + 1 = 5\text{th location for ordered pairwise difference}$$

$$k_b = U_{1-\alpha/2} = 21\text{st location for ordered pairwise difference}$$

Thus, a 95% confidence interval is $(O_{(5)}, O_{(21)}) = (pwd(5), pwd(21))$.



3.2 Lecture 8: Shift CI (cont.) & KS Test

We examine another application of the shift CI using the Mann-Whitney Test:

Example 3.2.1 (Gorilla Shift CI). Recall the Gorilla example:

	pwd		pwd		pwd		pwd
(145, 130)	15	(155, 130)	25	(170, 130)	40	(180, 130)	50
(145, 160)	-15	(155, 160)	-5	(170, 160)	10	(180, 160)	20
(145, 165)	-20	(155, 165)	-10	(170, 165)	5	(180, 165)	15
(145, 170)	-25	(155, 170)	-15	(170, 170)	0	(180, 170)	10


Ordered Difference: -25, -20, -15, -15, -10, -5, 0, 5, 10, 10, 15, 15, 20, 25, 40, 50

Find the 90% CI for the shift parameter.


- **Solution:** To find the 90% CI we go to the section of the table for $\alpha/2 = 0.10/2 = 5\%$. Then, $m = n = 4$, and going to that combination finds "lower" = 1 (i.e. $U_{\alpha/2} = 1$) and "upper" = 15 (i.e. $U_{1-\alpha/2} = 15$).

Our CI is found using the locations

$$k_a = U_{\alpha/2} + 1 = 2 \quad k_b = U_{1-\alpha/2} = 15$$

i.e., our CI is (pwd(2), pwd(15)) or (20, 40). 

Note: As with most confidence interval For a difference we have three possible outcomes for our CI: **positive, negative, or neutral:**

- If the CI for Δ has both bounds > 0 , this suggests group 1's distribution is **larger** than group two's distribution.
- If the CI for Δ has both bounds < 0 , this suggests group 1's distribution is **smaller** than group two's distribution.
- In the CI for Δ contains zero, then it suggests that there is **no significant difference** between the two distributions. 

We will now cover a popular technique that can be used to compare two distributions or compare one group to a named distribution. This technique allows us to explicitly see if $F_1(x) \neq F_2(x)$.

3.2.1 Kolmogorov Smirnov (KS) Test

Unlike the previous tests which test sees if there is a difference based on a certain statistic (mean, median, total, etc...), the KS test looks for **any** type of difference (spread, center, etc...). To perform this test, we follow these steps:

KS Test Steps

Step 1: The KS test uses an **absolute difference**, so there is only one pair of hypotheses:

$$H_0 : F_1(x) = F_2(x) \quad H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$$

Step 2: The test statistic simply measures the **largest difference between the empirical CDFs**. For notation:

Let $\hat{F}_1(x)$ = empirical (observed) CDF or group 1

Let $\hat{F}_2(x)$ = empirical (observed) CDF or group 2

On order to obtain the test statistic, we

1. Combine the data from both groups into one set, S
2. Calculate $\hat{F}_1(x)$ for group 1 using both groups observations and $\hat{F}_2(x)$ for group 2 using both groups observations
3. Calculate the difference between $|\hat{F}_1(x) - \hat{F}_2(x)|$ for all $x \in S$
4. Our test statistic is then the maximum of these differences

$$K_s = \max_x |\hat{F}_1(x) - \hat{F}_2(x)|$$

Step 3: The p-value is a permutation p-value (same form)

$$\frac{(\# \text{ of } |\hat{F}_1(x) - \hat{F}_2(x)| \geq K_s)}{\binom{n+m}{m}}$$

Step 4: If p-value $< \alpha$, reject H_0



Example 3.2.2 (Machine Dispensing). *A machine is supposed to dispense 160 oz a liquid. Measurements were taken before and after maintenance:*

Before:	16.55	15.36	15.95	16.43	16.01
After:	16.05	15.98	16.10	15.88	15.91

We want to test if the distribution of the liquid was the same before and after maintenance.

We only are interested if they are the same or not, not so much the direction of the change, hence we use the KS test.

(a) State H_0 and H_A

• **Solution:**

$$H_0 : F_1(x) = F_2(x) \quad H_A : F_1(x) \leq F_2(x) \text{ or } F_1(x) \geq F_2(x)$$

(b) Calculate the KS test statistic

- **Solution:** a table summarizing the process of calculating the test statistic is as follows:

Group	1	2	2	1	2	1	2	2	1	1
Data	15.36	15.88	15.91	15.94	15.98	16.01	16.05	16.10	16.43	16.55
$\hat{F}_1(x)$	1/5	1/5	1/5	2/5	2/5	3/5	3/5	3/5	4/5	1
$\hat{F}_2(x)$	0	1/5	2/5	2/5	3/5	3/5	4/5	1	1	1
Diff	1/5	0	1/5	0	1/5	0	1/5	2/5	1/5	0

$$\Rightarrow K_s = \max |Diff| = 2/5$$

(c) Find the p -value associated with your test statistic:

- **Solution:** For the 252 possible permutations, the relative frequency of K_s follows:

K_s^*	0.2	0.4	0.6	0.8	1
$P(K = K_s^*)$	32/252	130/252	70/252	18/252	2/252

The p -value is thus:

$$P(K \geq K_s) = P(K \geq 2/5) = \frac{130 + 70 + 18 + 2}{252} \approx 0.873$$

(d) Interpret the p -value in terms of the problem

- **Solution:** If in reality the distribution of ounces dispensed before and after repair were the same; we would observe our test statistic or more extreme (greater) with probability 0.873.

(e) Give your conclusion

- **Solution:** Since p -value $>$ 0.05, we do not have enough evidence to reject H_0 . Conclude H_0 , the distributions are the same.



We give now a summary of all two sample tests that we have shown so far:

All two-sample tests

1. Permutation tests
2. Wilcoxon-Rank-Sum
3. Mann-Whitney
4. Kolmogorov Smirnov



Next we will discuss power simulations for certain comparisons of the above.

3.3 Lecture 9: Comparison of Two-Sample Tests

3.3.1 Comparison of Two-Sample Tests

First, we compare $D_i = (\text{power of } t\text{-test} - \text{power of permutation test})$ when the data is normal (best case scenario for a t -distribution. We give a table of the results; note that the groups simulated are balanced (equal numbers), i.e. $m = n$. We only give the size of one group.

α	sample size	R (# of simulations)					
		100	200	400	800	1600	3200
		D_i					
0.01	10	0.087	0.048	0.032	0.015	0.014	0.010
0.01	20	0.093	0.040	0.021	0.015	0.009	0.007
0.05	10	0.029	0.019	0.017	0.010	0.005	0.008
0.05	20	0.012	0.008	0.006	0.005	0.008	0.003

Notice that even in the best case scenario for a t -test, as long as R is large then there is very little difference in power for the permutation test vs t -test (values on the far right are close to 0 in all cases).

Next we will compare using the mean and median with the permutation test vs the WRS test. In the the table that follows

$$\text{Let } D_i = |\text{power for WRS} - \text{power for permutation}| \quad R = 1600$$

and the winning test (with the higher power) is given in parenthesis.

simulated dist.	statistic	D_i	
		$m = n = 10$	$m = n = 20$
Normal	mean	0.045 (P)	0.027 (P)
	median	0.031 (W)	0.059 (W)
Laplace	mean	0.035 (W)	0.080 (W)
	median	0.033 (P)	0.024 (P)
Cauchy	mean	0.276 (W)	0.502 (W)
	median	0.080 (P)	0.084 (P)

In summary,

Distribution	Statistic	Winner
Symmetric	mean	P
Symmetric	median	WRS
Asymmetric	mean	WRS
Asymmetric	median	P

This gives you an idea of when to use what test.

Note: KS is mainly used when you have no preference for a specific statistic. It is a more sensitive test and can find differences in center and spread of a distribution. But it does not give direction of the difference. ▲

3.3.2 Tests for three or more groups

We now discuss methods for comparing distributions of 3 or more groups. This is non-parametric ANOVA. In the text that follows we adopt the notations:

Notation: Assume we have k groups, then let...

- X_{ij} = j th observation from i th group
- n_i = sample size of i th group
- \bar{x}_i = sample mean of i th group
- s_i^2 = sample variance of i th group
- N = overall sample size = $\sum_i^k n_i$
- \bar{x} = overall sample mean = $\sum_i^k n_i \bar{x}_i / N$

The overall idea is the same as in ANOVA—we compare the difference in means to the overall mean and to the spread of each group.

Recall that

$$SST = \text{Sum of Squares Treatment} = \sum_{i=1}^k n_i (x_i - \bar{x})^2$$

$$MST = \text{Mean Sum of Squares Treatment} = SST / (k - 1)$$

These measure the difference between groups.

Also,

$$SSE = \text{Sum of Squares Error} = \sum_{i=1}^k (n_i - 1)s_i^2$$

$$MSE = \text{Mean Sum of Squares Error} = SSE/(N - k)$$

These measure the variances within each group. Our test statistic F_s compares how big the variation between groups is to that within each group. In other words,

$$F_s = \frac{MST}{MSE}$$

Notice...

- When F_s is large \Rightarrow variance between groups is larger than that within groups \Rightarrow means are significantly different
- When F_s is small \Rightarrow variance between groups is smaller than that within groups \Rightarrow means are **not** significantly different



Traditionally, ANOVA makes the following assumptions:

ANOVA Assumptions

1. Random samples are taken from all k groups
2. Measurements from all k groups are independent (observing one does not change what can be observed for the others)
3. $\sigma_1 = \sigma_2 = \sigma_3 = \dots = \sigma_k$ (Assessed by Levenes Test)
4. $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ (Assessed by QQ Plot or Shapiro-Wilks Test)



Then $F_s \sim F_{[k-1, N-k]}$ distribution.

But when the assumptions do not hold, we do not know what the distribution of F_s is. However, we can find a permutation distribution under the assumption that all of the groups have the same mean (center). In order to conduct this test, we follow these steps:

Steps for ANOVA Permutation Test

Step 1: State the hypotheses.

- The null and alternative are

$$H_0 : F_1(x) = F_2(x) = \dots = F_k(x)$$

$$H_A : F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x) \text{ for some } i \neq j$$

Step 2: Calculate the observed statistic:

$$F_{\text{obs}} = \frac{MST}{MSE}$$

Step 3: Find the permutation p-value

- There are $N!/n_1!n_2!\dots n_k!$ total permutations, and this value is typically unmanageable, so we use random permutations. The process is:
 - 1 Randomly assign the N observations into the k groups **R > 4000 times** (null lets us do this)
 - 2 Calculate the R values of F_s , denoted F_i
 - 3 Our p-value is ($\#$ of $F_i \geq F_{\text{obs}}/R$)

Step 4: If p-value $< \alpha$, reject H_0



3.4 Appendix (Week 3)

We give an explanation for as to why the MW test gives identical results to the WRS test.

Theorem 3.4.1 (MW and WRS Equivalence). *The test statistics used for the MW test and WRS test yield identical results, i.e.*

$$U_{MW} = \varphi(U_{WRS})$$

for some function φ .

Proof. Notice that U_{MW} gives the number of observations of group 2 that are greater than those of group 1, taking into account for ties between the data sets. Notice that when the two groups are put into the same data set, the rank of any entry in group 2 tells us the number of observations (either from group 1 or group 2) that are **less than** the current entry in group 2. For example, if $R(x) = 5$ where $x \in \text{Group 2}$, then there are 5 observations less than this observation from group 2; they could be in either group though. If we were to sum these ranks only for group 2, then we get the total number of observations that are less than the ones in group 2, including those in group 2 themselves. To get only those observations that are from group 1, we subtract the total amount of observations from group 2 that are counted as we sum. This would be $1 + 2 + 3 + \dots + n_2$ in sum this becomes

$$\frac{n_2(n_2 + 1)}{2}$$

Hence, the quantity that gives the number of observations from group 1 that are less than those from group 2 is

$$U_{MW} = R_2 - \frac{n_2(n_2 + 1)}{2}$$

But, notice that $R_2 = U_{WRS}$ so $U_{MW} = \varphi(U_{WRS})$ as we sought to show. This leads to the exact same inference since the p-value has form:

$$P(U_{MW} \geq u_{MW}) = P(R_2 \geq r_2)$$

so the results are the same. This concludes the proof. ■

Note: This proof was adapted from ideas in [7].

Chapter 4

Week 4: More Non-Parametric
ANOVA

4.1 Lecture 10: ANOVA Permutation & KW Test

We begin with an example of the ANOVA permutation test from the previous lecture.

Example 4.1.1 (Mice & Dye). *Mice were fed three amounts of red dye "Low," "Medium," and "High." To test effect of these dyes, one group was also given a "Control" dye. The time to death in weeks was measured with summary statistics below:*

	Control	Low	Medium	High
Mean	91.36	71.88	72.40	65.25
Std. Dev.	11.01	11.59	22.14	28.06
n_i	11	9	10	8

(a) What assumptions may be violated based on the above?

- **Solution:** While we cannot assess normality, it does seem that the standard deviations by group may not be equal (medium and high are twice as large as control and low).

(b) State the appropriate null and alternative.

- **Solution:**

$$H_0 : F_1(x) = F_2(x) = \dots = F_k(x)$$

$$H_A : F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x) \text{ for some } i \neq j$$

(c) Calculate the observed value of the test statistic.

- **Solution:** If we know $MST = 1266.68$, then

$$SSE = \sum_{i=1}^k (n_i - 1)s_i^2 = 10(11.01)^2 + 8(11.59)^2 + 9(22.14)^2 + 7(28.06)^2$$

$$= 12213$$

and

$$MSE = 12213 / (38 - 4) = 359.22$$

$$\implies F_{obs} = MST / MSE = 1266.68 / 359.22 = 3.5262$$

(d) Calculate the p -value or estimate it.

- **Solution:** Based on $R = 5000$ permutations, we found the following permutation distribution of F :

F^*	1	2	3	4	5	6	7
$P(F \geq F^*)$	0.31975	0.1495	0.047	0.018	0.008	0.005	0.002

Since $F_{obs} = 3.5262$ is between 3 and 4 our p -value is between 0.018 and 0.047.

(e) How many possible permutations were possible?

- **Solution:** This would be an extension of the binomial coefficient, called the **multi-nomial coefficient** where the groups are the treatment and control groups:

$$\frac{38!}{11!9!10!8!} = 2.467 \times 10^{20}$$

(f) State your conclusion in terms of the problem if $\alpha = 0.05$.

- **Solution:** The p -value is $< \alpha$, so we reject H_0 . We conclude that at least one of the distributions of time until death is different between the dosage groups, i.e. at least one mean comparison leads to inequality. **Based on the data, which groups do you think are most likely to differ?**



4.1.1 Kruskal-Wallis (KW) Test

Similar to the WRS test, KW test uses ranks rather than the actual data (X_{ij}) values. We can use this if we have outliers and need a large sample to work with. This test also allows us to make confidence intervals for differences in ranks. The process for the test is quite the same as with all permutation tests: (Data) \implies (Test Statistic) \implies (Permutation Distribution) \implies (p -value) \implies (Decision). To conduct this test, we follow the following steps:

KW Test Steps**Step 1:** State H_0 and H_A

- Same as with permutation ANOVA:

$$H_0 : F_1(x) = F_2(x) = \dots = F_k(x)$$

$$H_A : F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x) \text{ for some } i \neq j$$

Step 2: Calculate the test statistic

- By definition, we have

$$KW_{\text{obs}} = \frac{1}{S_R^2} \sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

$$= \frac{\text{variability of ranks **between groups**}}{\text{overall variability of ranks}}$$

Where S_R^2 = variance of ranks regardless of groups (overall variance).

Note: This form of the KW test works when ties are present or not.

Step 3: Calculate the approximate permutation p-value

- Permute the groups R times find KW_i for all R permutations. Then,

$$\text{p-value} = (\# \text{ of } KW_i \geq KW_{\text{obs}}) / R$$

Step 4: Reject H_0 if p-value $< \alpha$ 

Note: The (KW) test will have **higher power** than a permutation test when:

1. Outliers are present
2. The distribution of one or more groups is skewed
3. The distribution of one or more groups has "heavy tails"



4.1.2 Large Sample Approximation to KW

If the n_i 's are large but an assumption of ANOVA is violated we may use a large sample approximation to the KW test.

Motivation: In traditional ANOVA, we know that SST/σ_ϵ^2 is distributed χ^2 with d.f. = $k - 1$.

Now, for the KW test we replace the data X_{ij} with ranks R_{ij} (the corresponding ranks). We then can show that $SST_R = \sum_{i=1}^k n_i (\bar{R}_i - \frac{N+1}{2})^2$. But, the normalizing constant for the χ^2 distribution has changed (since we are using R_{ij}). It follows that since \bar{R}_i is an average, that over large samples, it will be approximately normally distributed with mean $\mu_{\bar{R}_i} = (N+1)/2$ and variance $S_{\bar{R}_i}^2$ or

$$\bar{R}_i \sim N\left(\frac{N+1}{2}, S_{\bar{R}_i}^2\right)$$

To get an idea as to how we could turn \bar{R}_i into a test statistic, we will assume no ties in the data and use ordinal ranks. If this is the case, then $E(\bar{R}_i) = (N+1)/2$ and we derive the variance as follows:

$$\begin{aligned} V(\bar{R}_i) &= \frac{1}{n_i^2} \left[\sum_{j=1}^{n_i} V(R_{ij}) + \sum_{j \neq k} \text{Cov}(R_{ij}, R_{ik}) \right] \\ &= \frac{1}{n_i^2} \left[n_i S_R^2 + (n_i)(n_i - 1) \frac{S_R^2}{N-1} \right] \\ &= \frac{1}{n_i^2} \left[\frac{n_i(N^2 - 1)}{12} + \frac{(n_i)(n_i - 1)(N+1)}{12} \right] \quad \left(S_R^2 = \frac{N^2 - 1}{12}; \text{ no ties} \right) \\ &= \frac{1}{n_i^2} \left[\frac{n_i(N+1)(N - n_i)}{12} \right] \\ &= \frac{(N+1)(N - n_i)}{n_i(12)} = \frac{(N+1)N}{n_i(12)} - \frac{1}{12} \approx \frac{(N+1)N}{n_i(12)} \end{aligned}$$

It naturally then follows that

$$\frac{\bar{R}_i - E(\bar{R}_i)}{\sqrt{V(\bar{R}_i)}} \sim N(0, 1) \quad (\text{approx.})$$

and thus

$$\left(\frac{\bar{R}_i - E(\bar{R}_i)}{\sqrt{V(\bar{R}_i)}} \right)^2 \sim \chi_{[1]}^2 \quad (\text{approx.})$$

making

$$\sum_{i=1}^k \left(\frac{\bar{R}_i - E(\bar{R}_i)}{\sqrt{V(\bar{R}_i)}} \right)^2 = \sum_{i=1}^k \left(\frac{n_i [\bar{R}_i - (N+1)/2]^2}{N(N+1)/12} \right) = \frac{N-1}{N(S_R^2)} \sum_{i=1}^k n_i [\bar{R}_i - N(N+1)/2]^2 \sim \chi_{[k-1]}^2$$

Notice if N is large enough, then $N - 1/N \approx 1$. This makes our new statistic as

$$\frac{1}{S_R^2} \sum_{i=1}^k n_i [\bar{R}_i - N(N+1)/2]^2 \sim \chi_{[k-1]}^2$$

Hence, it is suggested that the constant "c" s.t.

$$E[c(SST_R)] = k - 1$$

is $c = 1/S_R^2$ since it normalizes SST_R so each summand is the square of a standard normal variate [7] [1].

This gives the statistic as

$$KW = \frac{1}{S_R^2} \sum n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

and $KW \sim \chi_{[k-1]}^2$. Thus, the 4 steps to conduct a hypothesis test using this approximation are:

Normal Approximation to KW Test Steps

Step 1: State H_0 and H_A

$$H_0 : F_1(x) = F_2(x) = \dots = F_k(x)$$

$$H_A : F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x) \text{ for some } i \neq j$$

Step 2: Calculate the test statistic

$$KW = \frac{1}{S_R^2} \sum n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

Step 3: Calculate the p-value $p\text{-value} = P(\chi_{[k-1]}^2) \geq KW$

Step 4: If $p\text{-value} < \alpha$, reject H_0



4.2 Lecture 11: KW Example & Mult. Comparisons

We begin with an example using the large sample approximation for the KW test.

Example 4.2.1 (Food Saltiness). *The saltiness score from 0 to 5 for three food products was recorded with the following summary statistics:*

	I	II	III
\bar{x}_i	4	2.62	1.67
s_i	1.15	1.41	0.82
\bar{R}_i	15.86	10.31	6.25
n_i	7	8	6

where $S_R^2 = 36.9$.

(a) Name an assumption for parametric ANOVA and how we would assess it.

- **Solution:** *The data we are given isn't alone to determine if what we sampled came from a Normal Distribution, so we have to use a Q-Q Plot and/or Shapiro-Wilks Test to assess normality. Also, looking at the sample standard deviations s_i , we can see some variation, giving reason to assess equal population variances via Levene Test.*

(b) Find the test statistic and p-value for the large sample approximation to KW

- **Solution:** *Following the form given,*

$$\begin{aligned}
 KW_{obs} &= \frac{1}{S_R^2} \sum n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2 \\
 &= \frac{1}{36.9} [7(15.86 - 11)^2 + 8(10.31 - 11)^2 + 6(6.25 - 11)^2] \\
 &= 8.252
 \end{aligned}$$

Since $k = 3$, we know $df = k - 1 = 3 - 1 = 2$. This implies

$$P(\chi_{[2]}^2 > 8.252) \in [0.01, 0.025] \quad (\text{by chi-squared table})$$

(c) Suppose the permutation p-value based on 5000 random permutations is 0.0094. If $\alpha = 0.01$, do your conclusions from (b) and the permutation test agree? Explain.

- **Solution:** *They don't. We would fail to reject H_0 for the large sample approximation but reject for the permutation test. It appears that the permutation test is more strict in assessing group differences.*

(d) What should be taken into account when choosing which non-parametric test to use?

- **Solution:** *We should consider the **distribution of each group**, if there are any*

outliers, and if the assumptions of parametric ANOVA are violated. If they aren't, then a non-parametric test will have lower power than a parametric one.



The next question to consider when we reject the null is, "which groups are different?" The tests we have made only detect if any difference is present, to make a better inference it helps to know where the differences are. Now, if we have K groups there are $\binom{k}{2}$ possible pair wise combinations that can be made.

However, if we create $\binom{k}{2}$ hypothesis tests or (CIs) for comparing the groups (which would identify which groups were different), we would have the problem of multiple comparisons as we are making what are known as **simultaneous inferences**.

4.2.1 Corrections for Multiple Comparisons

Suppose we make " g " total CIs (or HTs). Then, we have the definitions:

Definition 4.2.1 (Error Rate/Confidence).

$$\begin{aligned} \text{"Overall" Error Rate} &= \text{Chance of at least one Type I error out of "g" CIs} \\ &= \text{Opposite of Chance no Type I Errors} \\ &= 1 - \text{Chance of no Type I Errors} \\ &= 1 - (1 - \alpha)^g \end{aligned}$$

and

$$\begin{aligned} \text{"Overall" Confidence} &= \text{Confidence in all CIs Simultaneously} \\ &= \text{All CIs "Correct" (contain parameter of interest)} \\ &= (1 - \alpha)^g \end{aligned}$$



Also, recall that α is the probability of Type I Error for a single HT/CI. Notice that when we make many CIs the overall confidence **decreases** (since $(1 - \alpha) \in (0, 1)$ implies for any $g \in \mathbb{N}$ that $(1 - \alpha)^g < (1 - \alpha)$).

A way to correct for this is to use a **Bonferroni correction** and make corrected simultaneous family-wise CIs. The correction makes use of **Boole's Inequality** which we state for reference:

Boole's Inequality: For any events E_i where $i \in \{1, \dots, n\}$, we have

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i)$$



We now state Bonferroni's Correction:

Theorem 4.2.1 (Bonferroni's Correction). *For g family-wise CIs with confidence $1 - \alpha$, we have*

$$(1 - \alpha)^g \leq (1 - \alpha/g)^g \leq 1 - \alpha$$

$$\iff (1 - \alpha) - (1 - \alpha)^g \geq (1 - \alpha) - (1 - \alpha/g)^g$$

Proof. We first prove $(1 - \alpha)^g \leq (1 - \alpha/g)^g$. Clearly, $1 - \alpha \leq 1 - \alpha/g$ since $\alpha > \alpha/g$ since g is by definition greater than 1. It then naturally follows that $(1 - \alpha)^g \leq (1 - \alpha/g)^g$ since x^g is an increasing function for all positive x .

Now we prove the next equality. Suppose we have g CIs with Type I Error rate α/g . If we denote events $E_i =$ Chance of no Type I Error, we have $P(E_i) = (1 - \alpha/g)$. We then proceed as follows:

$$\begin{aligned} P\left(\bigcap_{i=1}^g E_i\right) &= (1 - \alpha/g)^g \\ &= 1 - P\left(\bigcup_{i=1}^g E_i^c\right) \\ &\leq 1 - \sum_{i=1}^g P(E_i^c) && \text{(Boole's Ineq.)} \\ &= 1 - g(\alpha/g) \\ &= 1 - \alpha \end{aligned}$$

Hence, $(1 - \alpha/g)^g \leq 1 - \alpha$. From here, we have the first statement in the biconditional, the second follows by subtracting $1 - \alpha$ on both sides and multiplying by -1 . This concludes the proof. ■

Remark 4.2.1. *The correction is simple, make all the individual CIs have confidence level $(1 - \alpha/g)100\%$ instead of each having $(1 - \alpha)100\%$. This increases the width of all individual CIs but keeps the overall error rate controlled at $\leq \alpha$ using $\alpha = 0.10, 0.05, 0.01$. ◆*

We now give an example illustrating this technique.

Example 4.2.2 (Error Analysis). *Suppose $\alpha = 0.05$. Then we can construct a table showing the "overall" error for regular intervals and intervals using the Bonferroni correction.*

g	1	3	6	10	30
No Bonferroni: $1 - (1 - \alpha)^g$	0.05	0.1426	0.2649	0.401	0.78
Bonferroni: $1 - (1 - \alpha/g)^g$	0.05	0.0492	0.0490	0.0489	0.04881

Notice the error rate is stable for Bonferroni Corrected intervals (less variance about 0.05). ♥

In interpretation, all we add is "we are corrected/family-wise/Bonferroni/simultaneous/overall $(1 - \alpha)100\%$ confident..."

In short, for Bonferroni Correction: **anywhere you would use α replace it with α/g .**

With this method, we can make cutoffs for group comparisons, i.e. if the observed group differences are greater than some threshold value, the groups are most likely different. The **Bonferroni Cutoff** is as follows:

Bonferroni Cutoff (Non-Parametric Version): If $|\bar{R}_i - \bar{R}_j| \geq z_{1-\alpha/2g} \sqrt{S_R^2(1/n_i + 1/n_j)}$, then the average ranks for group i and j are significantly different. ★

There is also another choice when making pairwise comparisons. This is known as **Tukey's Honest Significant Difference (HSD)**. The parametric version of this test is

- If $|\bar{x}_i - \bar{x}_j| \geq q(\alpha, k, df = N - k) \sqrt{MSE(1/n_i + 1/n_j)}$ where $q(\cdot)$ denotes the Tukey table distribution, then the **averages** of group i and j are significantly different.

To make the non-parametric version, we replace everything that involved X_{ij} with R_{ij} . Thus,

HSD (Non-Parametric Version): If $|\bar{R}_i - \bar{R}_j| \geq q(\alpha, k, df = N - k) \sqrt{(S_R^2/2)(1/n_i + 1/n_j)}$, then the average ranks for group i and j are significantly different. ★

We now give some notation to make work easier when conducting hypotheses:

Cutoff Notation:

- Let BON = the Bonferroni cutoff = $z_{1-\alpha/2g} \sqrt{S_R^2(1/n_i + 1/n_j)}$
- Let HSD = the Tukey cutoff = $q(\alpha, k, df = N - k) \sqrt{(S_R^2/2)(1/n_i + 1/n_j)}$

Example 4.2.3 (Salt Example (cont.)). In the salt example we can calculate BON , HSD , and $|\bar{R}_i - \bar{R}_j|$. Let $\alpha = 0.05$. Find which groups are significantly different.

- **Solution:** There are $\binom{3}{2} = 3$ possible pairwise comparisons. We compute on possible

comparison, it is between I and III:

$$z_{1-0.05/2(3)} = 2.39 \text{ and}$$

$$\sqrt{S_R^2(1/7 + 1/8)} = 3.14$$

$$\Rightarrow \text{BON} = 8.07$$

If we keep computing, we arrive at this table:

	I vs. II	I vs. III	II vs. III
$ \bar{R}_i - \bar{R}_j $	$ 15.86 - 10.31 $	$ 15.86 - 6.25 $	$ 10.31 - 6.251 $
	= 5.55	= 9.61	= 4.06
BON	7.51	8.07	7.84
HSD	8.03	8.62	8.37

Thus, groups I and III have significantly different average ranks.



Chapter 5

Week 5: Group Comparisons (cont.) & Linear Tests

5.1 Lecture 12: Permutation Cutoffs/Rev. of Linear Tests

5.1.1 Permutation Cutoffs for HSD and Bonferroni

If we have low amount of data to work with, we can also find permutation based versions of Tukey's HSD criteria by approximating $q(\cdot)$. Or, if we prefer, we can find $\binom{k}{2}$ permutation HTs for two groups and compare the p-values to α/g (Bonferroni correction).

For Tukey's permutation HSD, the steps to find the cutoff are as follows:

Tukey's Permutation HSD:

Step 1: Randomly shuffle each observation into a group, **R > 4000 times** (as with non-parametric ANOVA)

Step 2: Pick a comparison (dispersion) measure, T_{ij} . Common values are $|\bar{x}_i - \bar{x}_j|$, $|\bar{R}_i - \bar{R}_j|$, $|\text{median}_i - \text{median}_j|$, and $(\bar{x}_i - \bar{x}_j) / \sqrt{MSE(1/n_i + 1/n_j)}$.

Step 3: For each R permutation, calculate $Q_R = \max_{i,j} |T_{ij}|$. We choose the maximum out of all the group comparisons since under the null, all groups are identical in distribution and center.

Step 4: Let $q^*(\alpha)$ be the $(1 - \alpha)100\%$ percentile of Q . Then, groups i and j are significantly different when

(a) $|T_{ij}^{\text{obs}}| > q^*(\alpha)$ OR (equivalently)

(b) $p\text{-value} = (\# \text{ of } Q_R \geq |T_{ij}^{\text{obs}}|) / R \leq \alpha$



Example 5.1.1 (Salt Example (cont.)). Assess any group difference with the same data.

• **Solution:**

- **Method 1 (Bonferroni):** In R , 3 WRS tests were performed with the following permutation p -values

	I vs. II	I vs. III	II vs. III
p -value	0.07506	0.00641	0.2254

Using Bonferroni correction, to be $\approx (1-\alpha)100\%$ confident in our joint inferences, we compare these to $\alpha/3$. If $\alpha = 0.05$ we compare them to $0.05/3 = 0.0167$. Thus, again group I vs. III are significantly different.

- **Method 2 (Permutation HSD):** Based on $R = 4000$ permutations, the Tukey HSD permutation cutoff is $q^*(\alpha) = 1.7505$. The value of the dispersion measure is $T_{ij} = |\bar{x}_i - \bar{x}_j|$. A table giving the comparisons is then:

Groups	I vs. II	I vs. III	II vs. III
T_{ij}	1.38	2.33	0.95
$q^*(\alpha)$	1.7505	1.7505	1.7505

This gives the same inference as with method 1.

So all methods with differing criteria agree that groups I and III are significantly different.



5.1.2 Kruskal-Wallis vs Permutation

Since KW uses ranks it tends to have **higher power** when...

1. There are outliers
2. The distributions are highly skewed
3. The distribution has "heavy tails" (ex: t -distribution)

5.1.3 Trends and Associations


First, we will consider associations between two numerical variables.

Recall: The simplest type of association is a **linear association** when one variable has a linear trend with another. ▲

One of the ways to measure the strength of a linear relationship is through **correlation**. For completeness, we give a definition here:

Definition 5.1.1 (Correlation). The **correlation** ρ between two random variates X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

Note that $\rho \in [0, 1]$ only. 

Parametric Test For Correlation

For review, we give the parametric version for a correlation test. Let ρ denote the population correlation between numeric random variables X and Y . Assume we measure n pairs of data, (x_i, y_i) . Then, to conduct a parametric correlation test, we follow these steps:

Parametric Correlation Test

Step 1: State the hypotheses

H_0	H_A
$H_0 : \rho = 0$	$H_A : \rho \neq 0$
$H_0 : \rho \geq 0$	$H_A : \rho < 0$
$H_0 : \rho \leq 0$	$H_A : \rho > 0$

Step 2: Calculate the test statistic


$$t_s = r \sqrt{\frac{n-2}{1-r^2}} \quad (\text{df} = n-1)$$

where

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Step 3: Calculate the p-value

H_A	p-value
$H_A : \rho \neq 0$	$2P(t > t_s)$
$H_A : \rho < 0$	$P(t < t_s)$
$H_A : \rho > 0$	$P(t > t_s)$

Step 4: Reject H_0 if p-value $< \alpha$ 

In this test, we assume...

1. Pairs are independent (random selection of pairs)
2. (x_i, y_i) are distributed bivariate normal

Alternatively, we could also create the linear regression line and create a test for the slope. In

this setup...

True Model:	$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
Least Squares Line:	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

where $\hat{\beta}_1 = r(s_x/s_y)$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. The steps, then, to conduct a parametric regression test are:

Least-Squares Parametric Test

Step 1: State the hypotheses

H_0	H_A
$H_0 : \beta_1 = 0$	$H_A : \beta_1 \neq 0$
$H_0 : \beta_1 \geq 0$	$H_A : \beta_1 < 0$
$H_0 : \beta_1 \leq 0$	$H_A : \beta_1 > 0$

Step 2: Calculate the test statistic

$$t_s = \hat{\beta}_1 \sqrt{\frac{\sum (x_i - \bar{x})^2}{MSE}} \quad (\text{df} = n - 2)$$

where

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

Step 3: Calculate the p-value

H_A	p-value
$H_A : \beta_1 \neq 0$	$2P(t > t_s)$
$H_A : \beta_1 < 0$	$P(t < t_s)$
$H_A : \beta_1 > 0$	$P(t > t_s)$

Step 4: Reject H_0 if p-value $< \alpha$



With linear regression we assume...

1. Pairs are randomly sampled and independent
2. $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$

Notice:

- If β_1 or $\rho = 0$, then no linear relationship between Y and X
- If β_1 or $\rho < 0$, then negative linear relationship
- If β_1 or $\rho > 0$, then positive linear relationship



5.2 Lecture 13: Non-Parametric Linear Tests

The most common reasons for using a non-parametric test are...

1. Many outliers present (violates normality)
2. Non-constant variance (violates normal distribution)
3. Small sample size (may not be able to conclude normality of data)

5.2.1 Permutation Test for Slope

If we assume that $H_0 : \beta_1 = 0$ is true, it means that Y does not tend to change with X . In other words, any value of Y should be equally likely to be paired with any X since **there is no association between the two variates**. In order to perform a permutation test for the slope, we follow these steps...

Permutation Least-Squares Test

Step 1: State H_0 and H_A

H_0	H_A
$H_0 : \beta_1 = 0$	$H_A : \beta_1 \neq 0$
$H_0 : \beta_1 \geq 0$	$H_A : \beta_1 < 0$
$H_0 : \beta_1 \leq 0$	$H_A : \beta_1 > 0$

Step 2: Calculate the observed test statistic:

$$\hat{\beta}_1^{\text{obs}} = \text{estimated least-squares slope} = r \frac{s_y}{s_x}$$

Step 3: Calculate the permutation p-value:

- To permute the groups, there are n ways to pair The first y_i with an x_i . Then $n - 1$ ways to pair The second y_i with an x_i &etc... This gives $n!$ total permutations. To obtain a permutation distribution for $\hat{\beta}_1$ we:
 - 1 Permute the data, and calculate $\hat{\beta}_i$
 - 2 Repeat for either...
 - All $n!$ permutations OR
 - **R > 3000** random permutations
 - 3 The actual or estimated permutation p-values are:

H_A	Actual	Estimated
$\beta_1 > 0$	(# of $\hat{\beta}_1 \geq \hat{\beta}_1^{\text{obs}}$)/ $n!$	(# of $\hat{\beta}_1 \geq \hat{\beta}_1^{\text{obs}}$)/ R
$\beta_1 < 0$	(# of $\hat{\beta}_1 \leq \hat{\beta}_1^{\text{obs}}$)/ $n!$	(# of $\hat{\beta}_1 \leq \hat{\beta}_1^{\text{obs}}$)/ R
$\beta_1 \neq 0$	(# of $ \hat{\beta}_1 \geq \hat{\beta}_1^{\text{obs}} $)/ $n!$	(# of $ \hat{\beta}_1 \geq \hat{\beta}_1^{\text{obs}} $)/ R

4 If $p\text{-value} < \alpha$, reject H_0 

Example 5.2.1 (Physical Demand vs. Salary). Ratings for salary and physical demand were recorded on a scale from 1 to 10. The **ranked** results were...

Salary (Y):	2	6	3	5	7	10	9	8	4	1
Demand (X):	5	2	3	8	10	9	1	7	6	4

With summary statistics

	Salary	Demand
Mean	5.5	5.5
Std. Dev.	3.03	3.03

and $r = 0.261$ as well as $n = 10$.

Note: The data has been ranked and there are no ties, so $\bar{y} = \bar{x}$ and $s_y = s_x$.

(a) Find the estimated slope

• **Solution:**

$$\hat{\beta}_1^{obs} = r \frac{s_y}{s_x} = 0.261 \left(\frac{3.03}{3.03} \right) = 0.261$$

(b) Based on $R = 4000$ random permutations, we find the following permutation distribution:

K	-1	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1
$P(\hat{\beta}_1^i) \geq K$	1	0.9920	0.9295	0.7550	0.5055	0.2475	0.0825	0.0075	0

Assuming $H_A : \beta_1 > 0$, estimate the p -value.

• **Solution:** Our p -value is $P(\hat{\beta}_1^i \geq 0.261)$ and by the table above we have

$$P(\hat{\beta}_1^i \geq 0.25) = 0.2475 \quad \text{and} \quad P(\hat{\beta}_1^i \geq 0.50) = 0.0825$$

$$\implies p\text{-value} \in [0.0825, 0.2475]$$

Note: From R , $(\# \text{ of } \hat{\beta}_1^i \geq 0.261)/4000 = 0.2355$

(c) State your conclusion in terms of the problem

• **Solution:** Since $p\text{-value} > \alpha$ for any $\alpha > 0.10$, we fail to reject H_0 . We can not conclude that there is a significant positive linear relationship for the scores of salary and scores of physical demand.

Note: There were $10! = 362,880$ possible permutations. We sampled (if each permutation was unique) $4000/10! \approx 1.1\%$ of them. **Do you think this is enough? Why?** ♥

5.2.2 Large Sample Approximation to Permutation

If an assumption of regression is violated and if $n \geq 30$, we can use a large sample approximation to the permutation slope test. But, before moving on, we state one proposition that is important for the test.

Proposition 5.2.1 (Slope with Null Sample Correlation). *For any sample where $r = 0$, assuming a (simple) least squares model is the best fit, we have $\beta_1 = 0$ too.*

Proof. With the use of a least-squares model and knowledge that $r = 0$, we have $\hat{\beta}_1 = r(s_y/s_x) = 0(s_y/s_x) = 0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \bar{y}$. This makes the sample line as

$$y_i = \bar{y}$$

Now, under least squares, $E(Y_i) = y_i$. This makes it so

$$\beta_0 + \beta_1 x_i = \bar{y} \quad \forall x_i \in \text{Domain}$$

It then follows that $\beta_1 = 0$ as β_0, β_1 are fixed constants but x_i 's are not. Thus, $r = 0 \Rightarrow \beta_1 = 0$ as we sought to show. ■

Notice that for any single sample where $r = 0$, any instance of this quantity for r **always** gives $\beta_1 = 0$. Any variation is solely attributed by the ϵ_i 's then.

We now state a method for conducting a large sample approximation for the permutation test for association:

Permutation Large Sample Approximation Test for Association

Step 1: State H_0 and H_A

H_0	H_A
$H_0 : \rho_1 = 0$	$H_A : \rho_1 \neq 0$
$H_0 : \rho_1 \geq 0$	$H_A : \rho_1 < 0$
$H_0 : \rho_1 \leq 0$	$H_A : \rho_1 > 0$

Step 2: Calculate the test-statistic

$$z_s = \frac{(r - 0)}{1/\sqrt{n-1}} = r\sqrt{n-1}$$

since under the null hypothesis (no association) $r \overset{approx}{\sim} N(0, 1/\sqrt{n-1})$ ¹

Step 3: Calculate the p-value

¹See the [Appendix](#) for a derivation

$H_A : \rho \neq 0$	$2P(Z > z_s)$
$H_A : \rho < 0$	$P(Z < z_s)$
$H_A : \rho > 0$	$P(Z > z_s)$



Example 5.2.2 (Salary & Physical Demand (cont.)). Continuing previous example, recall $n = 10$ and $r = 0.261$. Let's test for association using a large sample approximation to the permutation test. The test-statistic is

$$z_s = 0.261\sqrt{10 - 1} \approx 0.8253 \quad \text{and} \quad H_A : \rho > 0$$

So the p -value is $P(Z > 0.8253) \approx 0.2033$. Thus, again we fail to reject H_0 .

Notice that the p -values were quite different (0.2355 vs 0.2033). Thus, the technique we use has a lot of influence on the outcome. How would you explain this?

What likely occurred in this sample is our **sample size is too small**, so the p -value is not accurate in the normal approximation.



5.3 Lecture 14: Ranked Correlation Tests

Just like in the other non-parametric tests, there is a ranked version for correlation tests.

Notice that if we rank X and Y the general trend of linear relationships still hold. For example, if Y tends to increase with X the rank of Y should also tend to increase with the rank of X . This means that there is a rank correlation between X and Y .

5.3.1 Spearman's Rank Correlation

To calculate Spearman's Rank Correlation we use the traditional formula for correlation and replace (x_i, y_i) with the corresponding adjusted ranks.

Notation: Let...

- $R(x_i) = \text{rank for } x_i, \quad \forall i \in \{1, \dots, n\}$
- $R(y_i) = \text{rank for } y_i \quad \forall i \in \{1, \dots, n\}$
- $\bar{R}(x) = \text{average rank of } x_i \text{ and } \bar{R}(y) = \text{average rank of } y_i$
- $s_{R(x)} = \text{standard deviation of rank of } X$
- $s_{R(y)} = \text{standard deviation of rank of } Y$

Then, we have

$$r_s = \text{Spearman's Rank Correlation}$$

$$= \frac{1}{n-1} \sum_i \left(\frac{R(x_i) - \bar{R}(x)}{s_{R(x)}} \right) \left(\frac{R(y_i) - \bar{R}(y)}{s_{R(y)}} \right)$$



Then, the steps to perform a ranked correlation test are:

Spearman's Ranked Correlation Test

Step 1: State H_0 and H_A

H_0	H_A
$H_0 : \rho_s = 0$	$H_A : \rho_s \neq 0$
$H_0 : \rho_s \leq 0$	$H_A : \rho_s > 0$
$H_0 : \rho_s \geq 0$	$H_A : \rho_s < 0$

where $\rho_s = \text{population Spearman's correlation}$.

Step 2: Calculate the test-statistic, r_s as defined above

Step 3: Calculate the p-value. Note that for $n = 1, \dots, 10$, there are critical values of the Spearman's correlation in table A12. They give $P(r_s^* \geq c)$ for various c values. Use it to compute p-values for the following cases:

H_A	p-value
$\rho_s \neq 0$	$2P(r_s^* \geq r_s)$
$\rho_s > 0$	$P(r_s^* \geq r_s)$
$\rho_s < 0$	$P(r_s^* \leq r_s)$

Note: For the tables, you may have to use the fact that $P(r_s^* < -c) = P(r_s^* > c)$ if $H_0 : \rho_s < 0$. ★

Note: There is also a large sample approximation for Spearman's which is exactly the same as that for the large sample approximation to permutation except we replace r by r_s . Same process, different p-values though. ▲

Example 5.3.1 (Theoretical Spearman's). Suppose $r_s = -0.74$, $n = 8$, and $H_A \rho_s \neq 0$. Then, the table A12 gives $P(r_s^* \geq 0.74) \approx 0.023$, so that the p-value is $2P(r_s^* \geq |-0.74|) \approx 0.046 < 0.05$. If $\alpha = 0.05$, we reject H_0 and conclude there is a linear relationship between the ranks of X and Y . ♥

5.3.2 Kendall's Tau

An alternative that doesn't use ranks directly, but also does not use the original data is Kendall's tau (τ). Note that this test uses the same ideas as the Mann-Whitney Test.

Suppose we look at a pair of paired observations $\{(x_i, y_i), (x_j, y_j)\} \quad \forall i < j$ say (x_1, y_1) and (x_2, y_2) . Then...

1. If as X increases Y tends to also increase, then we should see $x_1 > x_2 \Rightarrow y_1 > y_2$
2. If as X increases Y tends to decrease, then we should see $x_1 > x_2 \Rightarrow y_1 < y_2$

We use this to describe "discordant" and "concordant" pairs:

Definition 5.3.1 (Concordant Pairs). A pair of data points $(x_i, y_i), (x_j, y_j)$ is said to be **concordant** (in agreement) if

$$x_i < x_j \Rightarrow y_i < y_j \quad \text{or} \quad x_i > x_j \Rightarrow y_i > y_j \\ \iff (\Delta x)(\Delta y) = (x_i - x_j)(y_i - y_j) > 0$$



Definition 5.3.2 (Discordant Pairs). A pair of data points $(x_i, y_i), (x_j, y_j)$ is said to be **dis-**

cordant (in disagreement) if

$$\begin{aligned} x_i < x_j &\Rightarrow y_i > y_j \quad \text{or} \quad x_i > x_j \Rightarrow y_i < y_j \\ \iff (\Delta x)(\Delta y) &= (x_i - x_j)(y_i - y_j) < 0 \end{aligned}$$



Additionally, if any pairs do not fit the definitions above, then they are "tied." This means either $x_i = x_j$ or $y_i = y_j$ which implies that $\Delta x \Delta y = (x_i - x_j)(y_i - y_j) = 0$.


This then implies...

- If most pairs are concordant \implies positive linear relationship
- If most pairs are discordant \implies negative linear relationship

We can use this to create a definition for "Kendall's Tau" as a measure similar to correlation:

Definition 5.3.3 (Kendall's Tau). *The "population" value of Kendall's Tau is*

$$\begin{aligned} \tau &= (\text{chance of concordant pairs}) - (\text{chance of discordant pairs}) \\ &= P((X_i - X_j)(Y_i - Y_j) > 0) - P((X_i - X_j)(Y_i - Y_j) < 0) \\ &= P((X_i - X_j)(Y_i - Y_j) > 0) - (1 - P((X_i - X_j)(Y_i - Y_j) > 0)) \\ &= 2P((X_i - X_j)(Y_i - Y_j) > 0) - 1 \end{aligned}$$

Note that since the population is assumed to be continuous, $P((X_i - X_j)(Y_i - Y_j) = 0) = 0$. This could also be thought of as a rescaled probability of concordant pairs. 

Notice if all pairs are concordant, then $\tau = 1$. If all are discordant, then $\tau = -1$. If exactly half are concordant and half are discordant, then $\tau = 0$. Thus, Kendall's Tau is mimicking the properties of traditional correlation ρ .

Now, if we have to estimate τ , there are $\binom{n}{2}$ total pairs $(x_i, x_j), (y_i, y_j)$. Then, let

$$U_{ij} = \begin{cases} 1 & \text{if } (x_i - x_j)(y_i - y_j) > 0 & \text{(concordant)} \\ \sqrt{-1} & \text{if } (x_i - x_j)(y_i - y_j) = 0 & \text{(tied)} \\ 0 & \text{if } (x_i - x_j)(y_i - y_j) < 0 & \text{(discordant)} \end{cases}$$

Then, we define

$$\begin{aligned} V_i &= \sum_{j=i+1}^n \text{Re}(U_{ij}) = \# \text{ of concordant pairs for } i\text{th value } (x_i, y_i) \\ &= \sum_{i < j} \mathbf{1}(U_{ij} = 1) \end{aligned}$$

Notice that we start at $j = i + 1$ so that there we are never comparing the same pair. For the same reason, $i \in \{1, \dots, n - 1\}$. The sample version of Kendall's Tau r_τ is then given by a very similar form as with the population; probabilities turn into sample proportions:

Definition 5.3.4 (Sample Kendall's Tau). *The sampled estimate of Kendall's Tau is given by*

$$\begin{aligned}
 r_\tau &= (\text{prop. concordant pairs, no ties}) - (\text{prop. discordant pairs, no ties}) \\
 &= \frac{\sum_{i < j} \mathbf{1}(U_{ij} = 1) - \sum_{i < j} \mathbf{1}(U_{ij} = 0)}{\binom{n}{2} - \sum_{i < j} \mathbf{1}(U_{ij} = \sqrt{-1})} \\
 &= \frac{2[\sum_{i < j} \mathbf{1}(U_{ij} = 1)] - [\binom{n}{2} - \sum_{i < j} \mathbf{1}(U_{ij} = \sqrt{-1})]}{\binom{n}{2} - \sum_{i < j} \mathbf{1}(U_{ij} = \sqrt{-1})} \\
 &= \frac{2 \left[\sum_{i=1}^{n-1} V_i \right]}{\binom{n}{2} - \sum_{i < j} \mathbf{1}(U_{ij} = \sqrt{-1})} - 1 \\
 &\approx \frac{2 \left[\sum_{i=1}^{n-1} V_i \right]}{\binom{n}{2}} - 1 \qquad \qquad \qquad (\text{ties are rare})
 \end{aligned}$$

Note: *The definition given above is useful for computation, a precise definition (given by [6]) of r_τ is*

$$\begin{aligned}
 r_\tau &= (\text{number of concordant pairs}) - (\text{number of discordant pairs}) \\
 &= \frac{1}{\binom{n}{2}} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)
 \end{aligned}$$

where

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

but it is not practical. ♠

5.4 Appendix (Week 5)

5.4.1 Sample Correlation Coefficient's Distribution

We now derive an approximation of the sampling distribution of r under H_0 or $\beta_1 = 0$. Note that in simple linear regression, it can be shown

$$\frac{\hat{\beta}_1 \sqrt{\sum (x_i - \bar{x})^2}}{s} = r \sqrt{\frac{n-2}{1-r^2}} \sim t_{[n-2]}$$

If we set $X = r \sqrt{\frac{n-2}{1-r^2}}$, then we can see that a distribution function for X can be obtained using a transformation of variables, i.e. we use the equation $f(x)dx = f(r)dr$ to find $f(r)$, the density function. If we do this, we find that (see [8])

$$f(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} (1-r^2)^{\frac{\nu-2}{2}}$$

where $\nu = n - 2$. This is the true sampling distribution for the sample correlation coefficient; note $r \in [-1, 1]$. Because r^2 is symmetric about 0, so is $f(r)$ and likewise, $E(r) = 0$. We now derive the variance of r :

Proposition 5.4.1 (Variance of r). *The variance of the sampling distribution of the correlation coefficient is*

$$V(r) = \frac{1}{n-1}$$

Proof. For brevity, we set

$$A = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}$$

Using the definition of variance, we see:

$$\begin{aligned} V(r) &= A \int_0^1 (r-0)^2 (1-r^2)^{\frac{\nu-2}{2}} dr \\ &= A \int_0^1 \frac{u(1-u)^{\frac{\nu-2}{2}}}{2\sqrt{u}} du && (u = r^2 \Rightarrow du = 2rdr) \\ &= A \int_0^1 u^{1/2} (1-u)^{\frac{\nu-2}{2}} du \\ &= A \int_0^1 u^{(3/2)-1} (1-u)^{(\nu/2)-1} du \end{aligned}$$

Notice the last equation is the area under the beta function where $\alpha = 3/2$ and $\beta = \nu/2$. We

can then compute the integral as

$$\begin{aligned}
 &= A \frac{\Gamma(3/2)\Gamma(\nu/2)}{\Gamma(\frac{\nu+3}{2})} \\
 &= \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)} \times \frac{\Gamma(3/2)\Gamma(\nu/2)}{\Gamma(\frac{\nu+3}{2})} && \text{(substitute value of } A\text{)} \\
 &= \frac{1}{\sqrt{\pi}} \frac{\Gamma(3/2)}{\frac{\nu+1}{2}} \\
 &= \frac{1}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} \frac{2}{\nu+1} && \left(\Gamma(3/2) = \frac{\sqrt{\pi}}{2}\right) \\
 &= \frac{1}{n-1} && (\nu = n-2)
 \end{aligned}$$

Hence, $V(r) = \frac{1}{n-1}$, as we sought to show. ■

Now that we know the mean and variance of r , we can, upon inspection of its curve, see that a normal distribution with the same mean and variance approximates r 's distribution (it's a little below the curve, though). Hence,

$$r \stackrel{appx}{\sim} N(0, 1/\sqrt{n-1})$$

as we sought to show. Notice, we assumed normality of errors in the simple linear regression to yield normality of the observed slope $\hat{\beta}_1$. In a non-parametric setting, with a large enough sample size, any linear combination of the ϵ_i 's will be approximately normally distributed, so this can still hold in that setting (as long as σ_{ϵ_i} 's are all the same).

Chapter 6

Week 6: Correlation Tests (cont.) & Tests for Independence

6.1 Lecture 15: Hypothesis Tests for Kendall's Tau

There are 3 types of hypothesis tests using Kendall's τ : **Exact Hypothesis Tests**, **Permutation Tests**, and **Asymptotic Approximation Tests**. We now give them in the order they are presented.

6.1.1 Exact Hypothesis Test for Tau)

The steps for conducting an exact hypothesis test for τ are as follows:

Exact Test for Kendall's τ

Step 1: State H_0 and H_A

H_0	H_A
$H_0 : \tau = 0$	$H_A : \tau \neq 0$
$H_0 : \tau \leq 0$	$H_A : \tau > 0$
$H_0 : \tau \geq 0$	$H_A : \tau < 0$

Step 2: Calculate test-statistic

$$r_\tau = \frac{2 \sum_{i=1}^{n-1} V_i}{\binom{n}{2}} - 1$$

Step 3: Calculate the p-value Similarly to Spearman's, Kendall has an exact distribution table for r_τ^* for $n \in \{1, \dots, 10\}$ (notice the low sample size). It gives $P(r_\tau^* > c)$ and again $P(r_\tau^* < -c) = P(r_\tau^* > c)$. The p-values for each H_A are then

H_A	p-value
$\tau > 0$	$P(r_\tau^* > r_\tau)$
$\tau < 0$	$P(r_\tau^* < r_\tau)$
$\tau \neq 0$	$2P(r_\tau^* > r_\tau)$

Step 4: Reject H_0 if p-value $< \alpha$



6.1.2 Permutation Test for Tau

We follow the same procedure as in the exact test, except the only difference is in the p-value step:

Step 3 (Permutation Test): For $R > 2000$ random permutations, calculate r_{τ_i} For each permutations

H_A	p-value
$\tau > 0$	$(\#r_{\tau_i} \geq r_{\tau_{\text{obs}}})/R$
$\tau < 0$	$(\#r_{\tau_i} \leq r_{\tau_{\text{obs}}})/R$
$\tau \neq 0$	$(\# r_{\tau_i} \geq r_{\tau_{\text{obs}}})/R$

where $r_{\tau_{\text{obs}}}$ = observed Kendall's Tau from original sample



6.1.3 Asymptotic Approximation for Tau

The following formula can be used with or without ties in the data (for either X or Y). First, we note the frequency of ties with the following notation:

- Let $s_i = \#$ of ties for the i th tied value of X
- Let $t_i = \#$ of ties for the i th tied value of Y

For example, if

$X :$	0	1	1	2	3	3	3	4	5	6	6
$Y :$	1	1	2	3	3	4	4	4	5	6	7

then,

- $s_1 = 2$ (two values of 1)
- $s_2 = 3$ (three values of 3)
- $s_3 = 2$ (two values of 6)
- $t_1 = 2$ (two 1's)
- $t_2 = 2$ (two 3's)
- $t_3 = 3$ (three 4's)

Now, if we let

$$A = \frac{\sum_i s_i(s_i - 1)(2s_i + 5) + \sum_j t_j(t_j - 1)(2t_j + 5)}{18}$$

$$B = \frac{[\sum_i s_i(s_i - 1)(2s_i - 2)] [\sum_j t_j(t_j - 1)(t_j - 2)]}{9n(n - 1)(n - 2)}$$

and

$$C = \frac{[\sum_i s_i(s_i - 1)] [\sum_j t_j(t_j - 1)]}{2n(n - 1)}$$

We now give some properties of these quantities:

Proposition 6.1.1 (A,B,C when no ties). *When there are no ties in the data, $A = B = C = 0$.*

Proof. Notice if there are no ties, s_i and t_j have values of 0 for every index. This makes each sum in the formulas for A, B, C 0. Hence, all three values are thus 0 if there are no ties. ■

Proposition 6.1.2 (2 Repeated Values for Y). *If Y has $t_i \leq 2$, then $B = 0$.*

Proof. If $t_i \leq 2$ for all i , then $t_i - 1 = 0$ or $t_i - 2 = 0$ depending on the value of t_i . In either case, the sum $\sum_j t_j(t_j - 1)(t_j - 2)$ evaluates to 0 which in turn leads to $B = 0$, as we sought to show. ■

Proposition 6.1.3 (Mutually Exclusive Ties). *If ties exist for only one of the sets X or Y , then $B = C = 0$.*

Proof. If there are only ties for one of X and Y , then only one of s_i or t_j has 0 for every value. This makes any sum with these quantities evaluate to 0. Since B and C have products using these quantities, they are guaranteed to be 0 in either case. ■

Now, we can prove that the variance of r_τ is

$$V(r_\tau) = \frac{4n + 10}{9(n^2 - n)} - \frac{4}{n^2(n - 1)}(A - B - C)$$

and that the mean of r_τ is $E(r_\tau) = 0$. We can now state the steps for this hypothesis testing method:

Asymptotic Approximation for τ Test

Step 1: State the hypotheses

\mathbf{H}_0	\mathbf{H}_A
$H_0 : \tau = 0$	$H_A : \tau \neq 0$
$H_0 : \tau \leq 0$	$H_A : \tau > 0$
$H_0 : \tau \geq 0$	$H_A : \tau < 0$

Step 2: Compute the test statistic

$$z_s = \frac{r_\tau}{\sqrt{V(r_\tau)}}$$

Step 3: Compute the p-value

H_A	p-value
$\tau > 0$	$P(Z > z_s)$
$\tau < 0$	$P(Z < z_s)$
$\tau \neq 0$	$2P(Z > z_s)$

Note: This is primarily used when $n \geq 30$.



Note: We now have three correlations:

- Parametric-Pearson's: r
- Ranks-Spearman's: r_s
- Kendall's τ : r_τ



Let's assess the validity of non-parametric tests for correlation by an example.

Example 6.1.1 (Age & Body Fat). Age and body fat percentage were measured for 9 subjects:

Age (X) :	23	23	27	27	38	41	45	49	50
BF(Y) :	9.5	27.9	7.8	17.18	31.4	25.9	27.4	25.2	31.1

With corresponding correlations:

$$r = 0.658 \quad r_s = 0.395 \quad r_\tau = 0.286$$

Assume the claim is that body fat percentage **increases** with age (positive correlation).

(a) Calculate the asymptotic z-scores for all correlations and the appropriate p-values.

- **Solution:** For each type of test, we have

$$\text{Pearson: } z_s = r\sqrt{(n-2)/(1-r^2)} = 0.658\sqrt{7/(1-0.658^2)} = 2.311$$

$$\text{Spearman: } z_s = r_s\sqrt{n-1} = 0.395\sqrt{9-1} = 1.117$$

$$\text{Kendall: } s_1 = s_2 = 2 \quad \text{no ties in } Y \implies B = C = 0$$

$$\begin{aligned} \text{so } A &= \sum_i s_i(s_i - 1)(2s_i + 5)/18 = [2(1)(4 + 5) + 2(1)(2 + 5)]/18 \\ &= 2 \end{aligned}$$

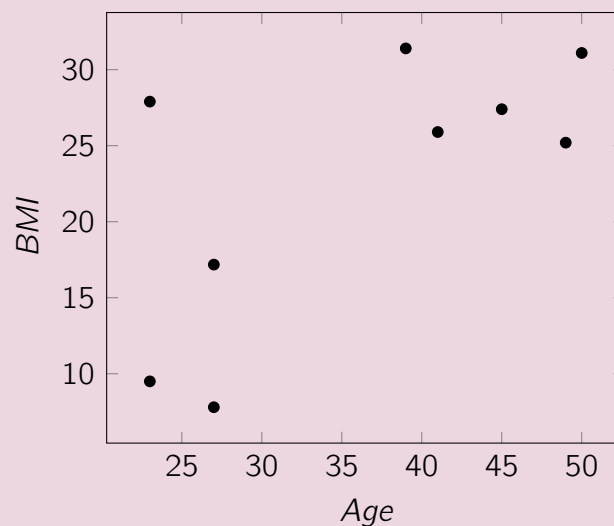
$$\begin{aligned}
 \text{and } V(r_\tau) &= \frac{4n + 10}{9(n^2 - n)} - \frac{4}{n^2(n - 1)}(A - B - C) \\
 &= (4(9) + 10)/(9(9^2 - 9)) - [4/(9^2(9 - 1))](2) \approx 0.0586 \\
 \implies z_s &= 0.286/\sqrt{0.0586} \approx 1.1815
 \end{aligned}$$

The p-values are thus:

statistic	p-value
r	$P(Z > 2.311) \approx 0.0104$
r_s	$P(Z > 1.117) \approx 0.1320$
r_τ	$P(Z > 1.1815) \approx 0.1189$

(b) Compare the results. Which p-value do you believe is more appropriate?

- **Solution:** Since the sample size is small, asymptotic distributions may not be accurate. But, if we have to use one, it is best not to use Pearson's. We can see why if we look at the plotted data:



Since it does not look bivariate normal (football/ellipse shaped), Pearson's parametric test is not applicable. Out of the above, use either Spearman's Rank test or Kendall's Tau.



6.2 Lecture 16: More on Correlation & Contingency Tables

6.2.1 When to use which Correlation

Some notes on which correlation test to use...

Correlation Notes

1. When there are no outliers and the distribution is approximately symmetric (but with low sample size), use a permutation test for the slope
2. When outliers are present in the data, use Spearman's or Kendall's since they remove effect of extreme values
3. Kendall and Spearman tend to have similar results, but Spearman tends to have higher power at low sample sizes and Kendall has higher power in large sample sizes.¹



6.2.2 Contingency Tables

A contingency table is used for two or more categorical variables and typically has the following form:

		Y				
		Cat 1	Cat 2	...	Cat c	
X	Cat 1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$ = row total for row 1
	Cat 2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$ = row total for row 2
	⋮	⋮	⋮	⋮	⋮	⋮
	Cat r	n_{r1}	n_{rc}	$n_{r.}$ = row total for row r
		$n_{.1}$	$n_{.2}$...	$n_{.c}$	$n_{..}$ = total sample size
		column total 1	column total 2	...	column total c	

This data typically comes from two methods of collection:

- (I) All n subjects are sampled randomly and independently so that **neither row nor column totals are known beforehand**
- (II) Row totals or column totals are known prior to the study and what is random is the **outcome of only one of the categorical variables**

¹See [4] for more information regarding power simulations

(I) is typically the result of an **observational study**. For example, 100 people are sampled and their major and gender are recorded.

(II) is used for **experiments**. For example, 50 subjects are randomly allocated into Drug group vs Placebo group so that there are 25 in each group. Then, the type of improvement is measured.

The goal with two categorical r.v.'s is to see if the outcome of one effects the outcome of another, i.e, if they are dependent or independent. Note, however, we cannot determine causality from any categorical analysis alone.

6.2.3 Notation

Let the population or true value of the probability of being in category i of X , j of Y is p_{ij} . Equivalently, $P(X = i, Y = j) = p_{ij}$. It then follows...

- $p_{i.} = P(X = i) =$ probability of being category i of X
- $p_{.j} = P(Y = j) =$ probability of being category j of Y
- $p_{i|j} = P(X = i|Y = j) =$ probability of being in category i of X , given in j of $Y = p_{ij}/p_{.j}$
- $p_{j|i} = P(Y = j|X = i) =$ probability of being in category j of Y , given in i of $X = p_{ij}/p_{i.}$

If two events A and B are independent, then

$$P(A, B) = P(A)P(B) \iff P(A|B) = P(A)$$

For categorical r.v.'s X and Y are independent, we have

$$p_{ij} = p_{i.}p_{.j} \iff p_{i|j} = p_{i.} \iff p_{j|i} = p_{.j}$$

6.2.4 Parametric χ^2 Test for Independence

Using the above notations, we can review the parametric test for independence.

Parametric χ^2 Test for Independence

Step 1: State H_0 and H_A

H_0 : Variables X, Y are independent

H_A : Variables X, Y are dependent

Step 2: Calculate test-statistic

- Here we compare the counts n_{ij} to what they should have been if the variables were actually independent (H_0 true). If they were independent we would see

$p_{ij} = p_i \cdot p_j \iff E(n_{ij}) = e_{ij} = np_i \cdot p_j$ where the left entry in the biconditional is the expected count based on average based on n subjects. In practice, we'll never know e_{ij} , but we can observe $\hat{e}_{ij} = n\hat{p}_i \cdot \hat{p}_j$ and simplification of this form gives

$$\begin{aligned}\hat{e}_{ij} &= n\hat{p}_i \cdot \hat{p}_j = n \left(\frac{n_{i.}}{n} \right) \left(\frac{n_{.j}}{n} \right) \\ &= \frac{n_{i.} \cdot n_{.j}}{n} = \frac{(\text{row total } i)(\text{row total } j)}{n}\end{aligned}$$

Notice that

$$\begin{aligned}\hat{e}_{ij} &= \frac{n_{i.}}{n} n_{.j} = \hat{p}_i \cdot n_{.j} \\ \iff \hat{p}_i &= \frac{\hat{e}_{ij}}{n_{.j}} \quad (\forall j)\end{aligned}$$

i.e. the prob of being in category i should be the same no matter what column j is. Finally, our test statistic is

$$\chi^2_{[s]} = \sum_{i,j} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad df = (r-1)(c-1)$$

Which is distributed $\chi^2_{[(r-1)(c-1)]}$ if H_0 is true

Step 3: Calculate the p-value

$$\text{p-value} = P(\chi^2 \geq \chi^2_{[s]})$$



We note some assumptions to carry out this test:

χ^2 Test for Independence Assumptions

1. Random sample was taken, i.e. X_i 's are mutually independent and Y_j 's are mutually independent
2. $\hat{e}_{ij} \geq 5$ for all i, j

Caveat: If n_{ij} 's have vastly different magnitudes (high variance), we may not have a χ^2 distribution for $\chi^2_{[s]}$, even if $\hat{e}_{ij} \geq 5$



6.3 Lecture 17: Permutation Test for Independence

Now, we give some non-parametric tests concerning independence.

6.3.1 Permutation Test for Independence

Permutation Test for Independence

Step 1: State H_0 and H_A

H_0 : Variables X, Y are independent

H_A : Variables X, Y are dependent

Step 2: Calculate the test statistic. Note, our test statistic is the same as it was for the parametric test:

$$\chi^2_{[s, \text{obs}]} = \sum_{i,j} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad df = (r - 1)(c - 1)$$

Step 3: Calculate the permutation p-value. The steps for the permutation p-value are as follows:

- (i) Fix either the row or column totals observed (this is an arbitrary constraint). Then randomly assign each subject in the row into a column. (Think of only one variate at a time Y_1, \dots, Y_n . (for example), we then assign the associated values of the other variate X_1, \dots, X_n . randomly assuming a **uniform distribution** without replacement).

Note: We may do this because if H_0 is true, then $p_{j|i} = p_{i.}$. I.e. if you are in row i the prob of being in column j should be the same no matter what column you are in. Thus we take the $n_{i.}$ values and shuffle them into all columns.

- (ii) Calculate $\chi^2_{s,i}$ for your permutation data
 (iii) Repeat (i, ii) **R > 2000 times**

The permutation based p-value is then

$$(\# \text{ of } \chi^2_{s,i} \geq \chi^2_{s, \text{obs}}) / R$$

Step 4: Reject H_0 if p-value $< \alpha$



Note:

- Whether you fix the rows or the columns does not matter, they will result in the same p-value, i.e. we yield the same permutation tables.

- The total number of ways to shuffle the columns and fix the rows is $n!/[n_1!n_2!\dots n_r!]$. Similarly, the total of ways to shuffle the rows and fix the columns is $n!/[n_1!n_2!\dots n_c!]$. Both are equal in value. ▲

Example 6.3.1 (Pain Relief). Seven patients were put into two groups. Group I took over the counter pain medications according to a doctor's recommendation. Group II self medicated with OTC medicine. The subject's pain relief was rated with values S = slightly reduced, R = reduced, and E = eliminated. The results are as follows:

	S	R	E	
I	2	2	0	$n_{1.} = 4$
II	0	1	2	$n_{2.} = 3$
	$n_{.1} = 2$	$n_{.2} = 3$	$n_{.3} = 2$	$n = 7$

(a) How many permutations are there if we fix the row totals?

- **Solution:** $7!/3!4! = 35$

(b) Find the observed test statistic, $\chi_{S,obs}^2$

- **Solution:** The table of the observed expected values \hat{e}_{ij} follows:

\hat{e}_{ij}	S	R	E	
I	$\frac{\binom{4}{7}\binom{2}{7} = 8/7$	$\frac{\binom{4}{7}\binom{3}{7} = 12/7$	$\frac{\binom{4}{7}\binom{2}{7} = 8/7$	4
II	$\frac{\binom{3}{7}\binom{2}{7} = 6/7$	$\frac{\binom{3}{7}\binom{3}{7} = 9/7$	$\frac{\binom{3}{7}\binom{2}{7} = 6/7$	3
	2	3	2	7

$$\begin{aligned} \Rightarrow \chi_{S,obs}^2 &= (2 - 8/7)^2/(8/7) + (2 - 12/7)^2/(12/7) + \dots + (2 - 6/7)^2/(6/7) \\ &= 4.278 \end{aligned}$$

(c) If the distribution of $\chi_{S,i}^2$ for all possible permutations is:

$\chi_{S,i}^2$	0.194	2.236	4.278	4.956	7.0	total
Freq	12	12	6	4	1	35

Find the exact permutation p-value.

- **Solution:** Since $(\# \text{ of } \chi_{S,i}^2 \geq 4.278) = 11$, the p-value is $11/35 = 0.31$.

(d) Interpret your p-value in terms of the problem

- **Solution:** If in reality pain relief and group were independent, we would observe our data or more extreme 31% of the time.

(e) State your conclusion in terms of the problem

- **Solution:** Since $p\text{-value} > \alpha$, we fail to reject H_0 and conclude there is evidence to support that group and pain relief are independent. ♥

6.3.2 Comparing Conditional Probabilities

Now, if we reject H_0 , we next want to identify the direction of the dependence. For example, how does the value of X exactly depend on the value of Y ? Which values of X depend on certain values of Y ? These are similar questions we ask when we conducted ANOVA permutation tests.

To do this comparison, we compare $p_{j|i} - p_{j|i'}$ (or $p_{i|j} - p_{i|j'}$). These are known as the conditional probabilities of i for different groups j and j' (or j for i and i').

The primary method used is similar to Tukey's HSD but modified for proportions.

Notation Let $Z_{j|i}$ be the test statistic comparing j (some column) conditional on i (some row). Then, we have

$$Z_{j|i} \equiv \frac{\hat{p}_{j|i} - \hat{p}_{j|i'}}{\sqrt{\bar{p}(1 - \bar{p})(1/n_i + 1/n_{i'})}}$$

where $\hat{p}_{j|i} = n_{ij}/n_i$, $\hat{p}_{j|i'} = n_{i'j}/n_{i'}$, and $\bar{p} = (n_{ij} + n_{i'j})/(n_i + n_{i'})$. Effectively, $Z_{j|i}$ measures the standardized difference between two cell counts for a given column.

Now we find the permutation values we will use for our cutoffs.

Step 1: Calculate all "g" of the observed $Z_{j|i}$'s (these are the observed values we'll need later)

Step 2: Find a random permutation and calculate all "g" values of the $Z_{j|i}$'s based on this permutation. Then, let $Q_k = \max_{i,j} |Z_{j|i}|$ and calculate Q_k

Step 3: Repeat (Step 2) **R > 2000** times. Then, calculate

$$q^*(\alpha) = (1 - \alpha)100\text{th percentile of all } Q'_k\text{'s}$$

With the cutoff value calculated, we can decide which differences are significant. To do this, we compare $Z_{j|i}^{\text{obs}}$ to $q^*(\alpha)$. If $Z_{j|i}^{\text{obs}} > q^*(\alpha)$, we conclude that the proportions used in $Z_{j|i}^{\text{obs}}$ are significantly different.

Note: To tell what direction the dependence is after we have determined which $|Z_{j|i}^{\text{obs}}| > q^*(\alpha)$, we can tell by the sign of the difference between $p_{j|i}$ and $p_{j|i'}$.

- (i) If $p_{j|i} < p_{j|i'} \implies$ probability of j in group i is less than probability of j in group i'
- (ii) If $p_{j|i} > p_{j|i'} \implies$ probability of j in group i is greater than probability of j in group i'



Chapter 7

Week 7: Prob. Comparisons & Bootstrapping

7.1 Lecture 18: Independence & Bootstrapping (Intro)

We begin with an example about finding dependence and its direction.

Example 7.1.1 (Car Color & Gender). Color of car was compared with gender of buyer with the following results:

	R = Red	S = Silver	B = Black	
F	2	16	3	21
M	3	2	4	9
	5	18	7	30

The permutation p-value was: 0.0025.

(a) Find the estimated difference in buying each color, comparing by gender.

- **Solution:** A table giving the probabilities and differences is as follows:

Red	Silver	Black
$\hat{p}_{R F} = 2/21$	$\hat{p}_{S F} = 16/21$	$\hat{p}_{B F} = 3/21$
$\hat{p}_{R M} = 3/9$	$\hat{p}_{S M} = 2/9$	$\hat{p}_{B M} = 4/9$
$\hat{p}_{R F} - \hat{p}_{R M} = -0.238$	$\hat{p}_{S F} - \hat{p}_{S M} = 0.540$	$\hat{p}_{B F} - \hat{p}_{B M} = -0.302$

(b) Calculate the relevant $Z_{j|i}$ values

- **Solution:** A table giving the values is

R	S	B
$\bar{p}_R = 5/30$	$\bar{p}_S = 18/30$	$\bar{p}_B = 7/30$
$Z_R = -1.604$	$Z_S = 2.769$	$Z_B = -1.790$

Where, for example,

$$\frac{\hat{p}_{R|F} - \hat{p}_{R|M}}{\sqrt{\bar{p}_R(1 - \bar{p}_R)(1/n_F + 1/n_M)}}$$

was used in calculating Z_R .

(c) Based off of $R = 5000$ permutations, the value of the cutoff is $q^*(\alpha = 0.05) = 2.114$. Which groups are significantly different and how?

- **Solution:** The only group that is significantly different is the silver group since $Z_S = 2.769 > 2.114$. Thus, the proportion of males and females who buy silver cars is different with females tending to buy silver cars more often since the difference is positive. ♥

7.1.1 Class so Far...

We have covered...

1. Single sample median, CDF, and percentiles
2. Independent two-sample tests
3. Independent k-sample tests
4. Linear Regression tests
5. Tests for Independence

Recall for tests 2-5 we used "permuting" the data in some way. When we permute data, we resample into each group, without replacement. In other words, each observation from each group is used exactly once.

Another method that can be used that also creates a distribution based on only one dataset is **bootstrapping**. This can be used in a huge variety of tests including all of what we have covered so far.

7.1.2 Bootstrapping

Let θ be the parameter we are interested in estimating. This could be one of the previous statistics, such as μ , θ_m , a percentile, etc. Generally, we know the distribution of our estimate and can use that to create HT and CIs.

In permutation tests, for example, we create a permutation distribution to find p-values but not confidence intervals. However, the permutation distribution assumes the null hypothesis is true in order to make these distributions.

Bootstrap distributions are also data driven distributions that we create from a single sample, but do not assume a particular H_0 is true since the method of distribution generation is general. Let's review how we would find a sampling distribution of $\hat{\theta}$ given infinite resources:

1. Take a random sample from the population
2. Calculate $\hat{\theta}$ our sample estimate
3. Repeat (1) and (2) many, many times (ideally infinite)

This gives a sampling distribution of $\hat{\theta}$ since we would have many realizations of $\hat{\theta}$.

Of course, we don't actually do this; it is too time consuming and intensive. But, bootstrapping mimics this process **so long as we are sure the sample we took is representative of the population it came from.**

7.1.3 Bootstrapping Sample

A bootstrap sample and a bootstrap distribution have the following steps. Assume you have a single sample of X_1, \dots, X_n . Then,

1. A bootstrap sample is resampling from X_1, \dots, X_n **with replacement**. We mark these resampled values with an asterisk: X_1^*, \dots, X_n^* is one possible bootstrap sample
2. A bootstrap estimate $\hat{\theta}_i^B = \varphi(X_1^*, \dots, X_n^*)$ is formed from your bootstrap sample, note all X_i^* 's are mutually independent
3. Repeat (1) and (2) B times. B is typically in the 1000s
4. The B values or $\hat{\theta}_i^B$ give a bootstrap distribution

Resampling with replacement adds variation to each bootstrap sample and mimics resampling from the population, though the values are repeated. Next, we will learn how to use this bootstrap distribution for statistical inference and point/interval estimation.

Fun Fact: If we observe n data points with m types of numbers yielding m empirical probabilities $\hat{p}_1, \dots, \hat{p}_m$. The particular bootstrap sample we generate will follow a **multinomial distribution**. So, if we observe a bootstrap frequency vector (n_1, \dots, n_m) , the chance of this occurring is:

$$P((N_1, \dots, N_k) = (n_1, \dots, n_k)) = P(\text{bootstrap sample}) = \binom{n}{n_1, \dots, n_k} \hat{p}_1^{n_1}, \dots, \hat{p}_m^{n_m}$$

In addition, each bootstrap variate X_i^* follows a **categorical distribution** $\mathcal{C}(\hat{p}_1, \dots, \hat{p}_m)$ where we have

$$X_i^* = \begin{cases} \gamma_1 & \text{with chance } \hat{p}_1 \\ \vdots & \vdots \\ \gamma_m & \text{with chance } \hat{p}_m \end{cases}$$

where γ_i is the i th unique value observed in the original sample. ▲

7.2 Lecture 19: Bootstrap Point/Interval Estimation

7.2.1 Estimating a Parameter

Some quantities that are often used to assess how good our estimate is are (assuming we are allowed to resample from the population N times):

Assessing Goodness of Estimate

1. **Expected Value (Average):** $E(\hat{\theta}) \approx \frac{1}{N} \sum \hat{\theta}_i$, for large N
2. **Bias:** $\text{bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$
3. **Variance:** $V(\hat{\theta}) \approx \frac{1}{N} \sum (\hat{\theta}_i - E(\hat{\theta}))^2$ for large N
4. **Mean Squared Error (MSE):** $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + \text{bias}(\hat{\theta})^2$

Proof. By definition: $\text{MSE}(\hat{\theta}) = \frac{1}{N} \sum (\hat{\theta}_i - \theta)^2$. Manipulating this definition with the addition and subtraction of $E(\hat{\theta})$ gives:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \frac{1}{N} \sum (\hat{\theta}_i - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= \frac{1}{N} \sum [(\hat{\theta}_i - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(\hat{\theta}_i - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] \\ &= \frac{1}{N} \left[\sum (\hat{\theta}_i - E(\hat{\theta}))^2 \right] + (E(\hat{\theta}) - \theta)^2 + 0 && \left(\sum (\hat{\theta}_i - E(\hat{\theta})) \approx 0 \right) \\ &= V(\hat{\theta}) + \text{bias}(\hat{\theta})^2 && \text{(by definition)} \end{aligned}$$

This concludes the proof.

5. **Chebyshev-Markov in Equality (distribution independent):**

$$P\left(|\hat{\theta} - \theta| \leq k\sqrt{\text{MSE}}\right) \geq 1 - \frac{1}{k^2} \quad \forall k \geq 1$$

Note the use of $\sqrt{\text{MSE}}$ instead of $\sigma_{\hat{\theta}}$ for the inequality. We can do this because $\text{MSE} \approx \sigma_{\hat{\theta}}^2 = V(\hat{\theta})$ assuming we use a reasonably unbiased estimator for θ . It is best to keep the MSE rather than the variance only because in practice there will always be bias (finite amount of samples). In the infinite case, we can disregard the bias as it is 0 for an unbiased estimator.

As a statement, the probability means that the chance that $\hat{\theta}$ is within $k\sqrt{\text{MSE}}$ from θ (in either direction) is at least $1 - 1/k^2$. ▲

We can estimate all these values with a bootstrap distribution. Assume you have B^* bootstrap estimates of θ (ex: the sample mean). Call them $\hat{\theta}_1^B, \dots, \hat{\theta}_{B^*}^B$ (the superscript tells us that these are bootstrapped estimates). The corresponding bootstrap estimates of the true sampling distribution of $\hat{\theta}$ are:

Bootstrap Estimates Qualities

1. **Bootstrap Expected Value:** $\hat{E}(\hat{\theta}) = \frac{1}{B^*} \sum_{i=1}^{B^*} \hat{\theta}_i^B$
2. **Bootstrap Bias:** $\hat{bias}(\hat{\theta}) = \hat{E}(\hat{\theta} - \hat{\theta}^{obs}) = \hat{E}(\hat{\theta}) - \hat{\theta}^{obs}$ where $\hat{\theta}^{obs}$ = estimate from original sample
3. **Bootstrap Variance:** $\hat{V}(\hat{\theta}) = \frac{1}{B^*} \sum (\hat{\theta}_i^B - \hat{E}(\hat{\theta}))^2$
4. **Bootstrap MSE:** $M\hat{S}E = \hat{V}(\hat{\theta}) + \hat{bias}(\hat{\theta})^2$
5. **Bootstrap Chebyshev-Markov Inequality:** For a given value k , we have

$$\hat{P} \left(|\hat{\theta} - \theta| \leq k \sqrt{M\hat{S}E} \right) \geq 1 - \frac{1}{k^2}$$



None of these calculations required any known knowledge of a distribution for $\hat{\theta}$. This technique can be used for **any** sample size as well. Notice that the standard error or estimated standard deviation can be calculated as

$$\hat{S}E(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}$$

which we will use in some confidence intervals.

Example 7.2.1 (Bootstrap vs. Parametric). *Let's compare a parametric test with known values of $E(\hat{\theta})$, $bias(\hat{\theta})$, MSE , $SE(\hat{\theta})$. Suppose we simulate data from a population where $\mu = 37.8243$, $\sigma = 6.507154$. Let the estimator be $\hat{\theta} = \bar{X}$.*

A random sample of size 60 was taken with sample mean $\bar{x} = 37.5833$ standard deviation $SE(X) = 6.282$.

By parametric theory \bar{X} should have no bias (sample was i.i.d., sometimes this isn't always true though) and $SE(\bar{X}) = \hat{\sigma}_{\bar{x}} = s/\sqrt{n} = 6.282/\sqrt{60} = 0.81100$. Based on 5000 bootstrap samples, we find...

$\hat{E}(\hat{\theta}) = 37.57745$	$\hat{bias}(\hat{\theta}) = -0.0058767$
$\hat{S}E(\hat{\theta}) = \hat{\sigma}_{\bar{x}} = 0.807116$	$M\hat{S}E(\hat{\theta}) = 0.65147$

Notice our bootstrap SE is lower and we estimated our bias as -0.0058767. If we assume that $\hat{\theta}^{obs}$ came from the bootstrap distribution since we would never have knowledge about any parametric assumptions about the data we sampled, then we adjust $\hat{\theta}^{obs}$ by the bootstrap bias:

$$\begin{aligned} \hat{\theta}^{obs} - (-0.0058767) &= 37.5833 + 0.0058767 \\ &= 37.58921 \end{aligned}$$

Thus, bootstrap is fairly competitive even when parametric assumptions hold (the estimate

is close to the parametric (true) estimate). Notice bootstrapping assumed nothing about the data we collected.



The example gives way to a handy definition:

Definition 7.2.1 (Bias-Corrected Estimate). *The bootstrap bias-corrected estimate of θ is*

$$\hat{\theta}_c = \hat{\theta}^{obs} - \hat{bias}(\hat{\theta})$$



7.2.2 Bootstrap Confidence Intervals

There are many different types of bootstrap CIs, some of which mimic a traditional CIs while others do not. We give 2 such intervals now:

Bootstrap CIs

1. Percentile Method

- This method is simple to implement, but only works well when the distribution is symmetric. What "works well" is typically defined to be the realized confidence level. $(1 - \alpha)100\%$ is the theoretical confidence level; we won't get this level of confidence in practice.

A $(1 - \alpha)100\%$ percentile bootstrap CI is:

- Create B bootstrap estimates. Then the CI is $(\hat{\theta}_{\alpha/2}^B, \hat{\theta}_{1-\alpha/2}^B)$ i.e. the $(\alpha/2)100$ th and $(1 - \alpha/2)100$ th percentiles of the bootstrap distribution. While this CI is easy to implement, **it often has much lower coverage than it should.**

2. Empirical Bootstrap CI

- Typically when we make a CI (such as for the population mean μ), we use

$$P(\delta_{\alpha/2} < \hat{\theta} - \theta < \delta_{1-\alpha/2}) = 1 - \alpha$$

and we know the distribution of δ . Then, the CI is: $(\hat{\theta} - \delta_{\alpha/2}, \hat{\theta} - \delta_{1-\alpha/2})$. For this CI, we estimate the percentiles of the differences $\hat{\theta} - \theta$ with bootstrapping. We use these steps:

- Create a bootstrap sample, find $\hat{\theta}_i^B$
- Find $\delta_i^B = \hat{\theta}_i^B - \hat{\theta}^{obs}$
- Repeat (a) and (b) B times to form the distribution of δ .

The empirical bootstrap CI is then:

$$\begin{aligned} & (\hat{\theta}^{\text{obs}} - \delta_{\alpha/2}, \hat{\theta}^{\text{obs}} - \delta_{1-\alpha/2}) \\ \iff & (2\hat{\theta}^{\text{obs}} - \hat{\theta}_{\alpha/2}^B, 2\hat{\theta}^{\text{obs}} - \hat{\theta}_{1-\alpha/2}^B) \quad (\delta_p^B = \hat{\theta}_p^B - \hat{\theta}^{\text{obs}} \text{ for any percentile } p) \end{aligned}$$

This CI often has a higher coverage probability than the percentile method.



7.3 Lecture 20: BCA Bootstrap CI

7.3.1 Bootstrap CIs continued

We give an additional method for creating a bootstrap CI.

Bias Corrected and Accelerated (BCA) Bootstrap CI

This method requires the most mathematical explanation.

First it assumes that there exists some transformation T of $\hat{\theta}$ such that $T(\hat{\theta})$ is **normally distributed**. The transformation that allows this to happen is (see [5])

$$T(\hat{\theta}) = T(\theta) + \sigma_{T(\theta)}(Z - z_0) \quad (Z \sim N(0, 1))$$

where $\sigma_{T(\theta)} = \sqrt{V(T(\hat{\theta}))} = 1 + aT(\theta)$. Naturally, then, this means that

$$E(T(\hat{\theta})) = T(\theta) - z_0[1 + aT(\theta)]$$

where z_0 is some standard normal percentile.

Now if $T(\hat{\theta})$ is actually normally distributed, we know

$$\begin{aligned} & P\left(-z_{1-\alpha/2} \leq \frac{T(\hat{\theta}) - E(T(\hat{\theta}))}{\sqrt{V(T(\hat{\theta}))}} \leq z_{1-\alpha/2}\right) = 1 - \alpha \\ \implies & P\left(-z_{1-\alpha/2} \leq \frac{T(\hat{\theta}) - (T(\theta) - z_0(1 + aT(\theta)))}{1 + aT(\theta)} \leq z_{1-\alpha/2}\right) = 1 - \alpha \\ \implies & P\left(-z_{1-\alpha/2} \leq \frac{T(\hat{\theta}) - T(\theta)}{1 + aT(\theta)} + z_0 \leq z_{1-\alpha/2}\right) = 1 - \alpha \\ \implies & P\left(\frac{T(\hat{\theta}) + z_0 - z_{1-\alpha/2}}{1 - a(z_0 - z_{1-\alpha/2})} \leq T(\theta) \leq \frac{T(\hat{\theta}) + z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right) = 1 - \alpha \end{aligned}$$

Now, we do not actually know the distribution at $\hat{\theta}$ but if we were to estimate it with a bootstrap distribution $\hat{\theta}^B$. An assumption we make is that the bootstrap distribution is a close approximation to the true (unknown) sampling distribution. It is then the case that the bootstrap distribution is **conditional on the point estimate we observe**. Hence, we can write the transformation for the bootstrap distribution as (see [2])

$$T(\hat{\theta}^B) = T(\hat{\theta}) + \sigma_{T(\hat{\theta})}(Z - z_0)$$

where

$$\sigma_{T(\hat{\theta}^B)} = \sqrt{V(T(\hat{\theta}^B))} = 1 + aT(\hat{\theta})$$

and

$$E(T(\hat{\theta}^B)) = T(\hat{\theta}) - z_0(1 + aT(\hat{\theta}))$$

If we focus on only the upper bound in the last probability statement given above, we have an analogous statement:

$$P\left(T(\hat{\theta}) \leq \frac{T(\hat{\theta}^B) + z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} \middle| T(\hat{\theta})\right) \quad (7.3.1)$$

$$= P\left(\frac{T(\hat{\theta}) - E(\hat{\theta})}{\sqrt{V(T(\hat{\theta}))}} \leq z_0 + \frac{T(\hat{\theta}^B) + z_0 + z_{1-\alpha/2} - T(\hat{\theta})[1 - a(z_0 + z_{1-\alpha/2})]}{(1 - a(z_0 + z_{1-\alpha/2}))(1 + aT(\hat{\theta}))} \middle| T(\hat{\theta})\right) \quad (7.3.2)$$

$$= P\left(Z \leq z_0 + \frac{T(\hat{\theta}^B) - T(\hat{\theta})}{(1 - a(z_0 + z_{1-\alpha/2}))(1 + aT(\hat{\theta}))} + \frac{z_0 + z_{1-\alpha/2}[1 + aT(\hat{\theta})]}{(1 - a(z_0 + z_{1-\alpha/2}))(1 + aT(\hat{\theta}))} \middle| T(\hat{\theta})\right) \quad (7.3.3)$$

$$\approx P\left(Z \leq z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right) \quad (T(\hat{\theta}^B) \approx T(\hat{\theta}))$$

Where (7.3.2) was possible since $T(\hat{\theta})$ is still a random variate (we haven't observed $\hat{\theta}$ yet). In short, we have just shown:

$$\begin{aligned} & P\left(T(\hat{\theta}) \leq \frac{T(\hat{\theta}^B) + z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} \middle| T(\hat{\theta})\right) \\ & \approx P\left(Z \leq z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right) \end{aligned}$$

Which implies that the upper estimate for the BCA CI is at the location u of the bootstrap distribution where

$$u = F_{\hat{\theta}^B}^{-1}\left(\Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right)\right)$$

and $F_{\hat{\theta}^B}^{-1}(\cdot)$ is the quantile function of the bootstrap distribution. We use this form because we know the transformed bootstrap is approximately normal, so we can compute a quantile with the same area as the one given for the standard normal curve in the approximation. Since $T(\hat{\theta}^B)$ is just transformed data, areas stay the same and we can use quantiles of $\hat{\theta}^B$'s distribution to get the upper bound for the CI (this is similar to what we do when we standardize variates and then compute quantiles for the actual data, so long as the transformation is monotonic).

Using the same process, we can show the lower bound for the CI is

$$l = F_{\hat{\theta}^B}^{-1}\left(\Phi\left(z_0 + \frac{z_0 - z_{1-\alpha/2}}{1 - a(z_0 - z_{1-\alpha/2})}\right)\right)$$

In practice, we can never achieve the upper and lower bounds exactly due to the discrete nature of the bootstrap distribution. Note also that l and u are dependent on a and z_0 so it is fair to write $l = l(a, z_0)$ and $u = u(z, z_0)$. Our next task is to find what these constants are. First, we state how we interpret them [5]:

a = acceleration constant (increases variance of trans. distn.)
 z_0 = bias correction constant (shifts trans. distn.)

Notice: This means we have to estimate z_0 and a but we **do not have to** use the exact transformation of $T(\hat{\theta})$. We assume it exists but never actually need to know it, although we did give one form earlier.

- **Estimating z_0**

- z_0 is a measure of bias in the data using the median. If we think of $\hat{\theta}$ as the best estimate of θ , then we can get a count of how well the bootstrap estimates approximate θ by counting the amount of values below $\hat{\theta}$. Ideally, if there was no bias, each *transformed* bootstrap estimate would have an equal chance (50%) of being below $T(\hat{\theta})$. Hence,

$$\text{Let } p_0 = (\# \text{ of } T(\hat{\theta}^B) \leq T(\hat{\theta})) / B$$

Since the transformation T is monotonic, we have

$$\begin{aligned} p_0 &= (\# \text{ of } \hat{\theta}^B \leq \hat{\theta}) / B \\ &= \text{proportion of bootstrap } \hat{\theta}'s \leq \text{sample } \hat{\theta} \end{aligned}$$

Notice under our transformation $p_0 \approx P(T(\hat{\theta}^B) \leq T(\hat{\theta}))$, we normalize this accordingly

$$\begin{aligned} p_0 &\approx P\left(T(\hat{\theta}^B) - E(T(\hat{\theta}^B)) \leq T(\hat{\theta}) - [T(\hat{\theta}) - z_0(\sigma_{T(\hat{\theta}^B)})]\right) \\ &= P\left(\frac{T(\hat{\theta}^B) - E(T(\hat{\theta}^B))}{\sigma_{T(\hat{\theta}^B)}} \leq z_0\right) \\ &= P(Z \leq z_0) \end{aligned}$$

Thus, z_0 is a point such that $p_0 = P(Z \leq z_0)$. Notice, if p_0 is large, then z_0 is large as well. Also, under no bias: $E(T(\hat{\theta}^B)) = T(\hat{\theta})$.

- **Estimating a**

- a is a measure of **skewness in the data** to estimate the effect of each X_i (data) on the distribution of we leave the i th X_i out of the data and calculate $\hat{\theta}_{-i}$ (the estimate of θ without the i th X_i). This is known as **jackknife resampling**. Do this for all i and calculate $\hat{\theta}_{(-1)} = \text{mean of all } \hat{\theta}_{-i}$.

Based on bootstrap theory,

$$a = \frac{\sum_i (\hat{\theta}_{(-1)} - \hat{\theta}_{-i})^3}{6 [\sum_i (\hat{\theta}_{(-1)} + \hat{\theta}_{-i})^2]^{3/2}}$$

Notice we **cube** the differences to keep the sign of the skew. If $a = 0$, then there is symmetry in the data.

To summarize,

$$\text{BCA CI} = \left[F_{\hat{\theta}^B}^{-1}(\Phi(\gamma)), F_{\hat{\theta}^B}^{-1}(\Phi(\delta)) \right]$$

where $\gamma = z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}$, $\delta = z_0 + \frac{z_0 - z_{1-\alpha/2}}{1 - a(z_0 - z_{1-\alpha/2})}$

This CI will be done in R in practice. On an exam, it would be given or values for z_0 , a would be provided. ★

Note: If a bootstrap distribution is symmetric, the CI bounds for the Percentile, Empirical, and BCA will be similar. ▲

Example 7.3.1 (Interval Comparisons). *Suppose data from an exponential distribution (skewed positively) is simulated, with a population mean of 10 and standard deviation 10. Suppose we want to estimate*

1. The mean (10)
2. The median (6.9315)
3. The standard deviation (10)

Using R, The 95% CIs for all three are (with $B = 10,000$):

	mean	median	standard deviation
Percentile	(8.18, 14.044)	(6.360, 10.411)	(6.050, 14.443)
Empirical	(7.766, 13.627)	(5.795, 9.846)	(7.329, 15.722)
BCA	(8.535, 14.700)	(6.413, 10.411)	(7.480, 15.572)

Ideally, since the CI level is the same, the best intervals have a smaller width (difference between upper bound and lower bound). **Out of the ones given, which are best?**

Note: We can use these boot-strapping methods to find estimates for **any** θ based off of a single sample so long as an estimator based off of the data we have is known for it. ♥

Chapter 8

Week 8: Bootstrap "t" Interval

8.1 Lecture 21: Bootstrap "t" Interval

8.1.1 General Method

Some bootstrap CIs use the parametric form as a starting point, and then modify what is needed in the parametric assumptions are violated. In a parametric setting, we can assume

$$\hat{\theta} \sim N(E(\hat{\theta}), \sigma_{\hat{\theta}}^2)$$

for a single sample. A parametric CI is typically then:

$$\hat{\theta} \pm t_{1-\alpha/2} SE(\hat{\theta})$$

But, if assumptions are violated, then the distribution used is not actually normal (or t). But we can bootstrap it (create a bootstrap distribution)!

A "t" bootstrap interval has the following steps:

t Bootstrap Interval

Step 1: Find $\hat{\theta}$, the observed estimate

Step 2: Generate bootstrap sample

Step 3: Calculate $\hat{\theta}_i^B$, the bootstrap estimate

Step 4: Calculate $t_i^B = (\hat{\theta}_i^B - \hat{\theta}) / \hat{SE}^B(\hat{\theta})$

Step 5: Repeat 2-4 B times

We now have a bootstrap distribution of t^B . The "t" bootstrap CI is then:

$$[\hat{\theta} - t_{1-\alpha/2}^B SE(\hat{\theta}), \hat{\theta} + t_{1-\alpha/2}^B SE(\hat{\theta})]$$

Where

$$t_{1-\alpha/2}^B = (1 - \alpha/2)100\text{th percentile of } t^B$$

$$t_{\alpha/2}^B = (\alpha/2)100\text{th percentile of } t^B$$

Note: t^B is a distribution with negative and positive values but not necessarily symmetric (it's bootstrapped, so it reflects what information the sample gives it). Also, we have only replaced $t_{1-\alpha/2}$ and $t_{\alpha/2}$ from the parametric CI with bootstrap estimates of them. ★

When to use what CI for single samples?

1. When the theoretical SE and distribution of $\hat{\theta}$ are known and the bootstrap distribution is symmetric, "t" intervals tend to outperform all others
2. If the distribution is symmetric (not necessarily normal or t), then empirical CIs or BCA CIs tend to outperform the others
3. If there is significant skew in the distribution, BCA will outperform the others

Here are some examples of when you could use a t bootstrap CI:

1. $\hat{\theta} = \bar{x}$ (sample mean) since $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
2. $\hat{\theta} = \hat{p}$ (sample proportion) since $\hat{p} \sim N(p, \sqrt{p(1-p)/n})$

8.1.2 Bootstrapping with Two Samples

When we have two samples, we need to adjust how we create a bootstrap sample. First, we will look at the empirical, percentile, and BCA methods. Before we discuss the interval method, we give some notation:

Notation: Let Y_{ij} = j th value in i th group. Let n_i = # of observations in group i . Note the bounds for i and j are $i \in \{1, 2\}$ and $j \in \{1, 2, \dots, n_i\}$.

To form a bootstrap sample we follow these steps:

2-sample t Bootstrap CI Method

Step 1 Calculate $\hat{\theta}$ from your sample. Examples of $\hat{\theta}$ could be

$$\hat{\theta} = \bar{Y}_1 - \bar{Y}_2 \quad \hat{\theta} = \text{median}_1 - \text{median}_2 \quad \hat{\theta} = s_1 - s_2 \text{ (diff in sd. dev.)}$$

etc...

Step 2 Resample n_1 from Y_{11}, \dots, Y_{1n} with replacement. This forms a bootstrap sample for group 1. Similarly, resample n_2 from Y_{21}, \dots, Y_{2n} with replacement.

Step 3 Calculate $\hat{\theta}_i^B$ based on the bootstrap samples

Step 4 Repeat 2-3 B times to obtain the bootstrap distribution of $\hat{\theta}_i^B$



From here, we can calculate empirical and percentile CIs as usual (same forms as before). But, when we calculate "a" in the BCA CI, we need to adjust how we "leave one out" (conduct jackknife sampling).

Jackknife Correction for 2-samples: To "leave one out" we treat all $n_1 + n_2$ observations as one group and leave out one at a time. This means one of the groups sample size changes, but not both. We pool the samples essentially.

8.1.3 "t" Method for Two Samples

When $\hat{\theta}$ has a known distribution and SE (for example with $\hat{\theta} = \bar{Y}_1 - \bar{Y}_2$) we can also use the t-bootstrap CIs as an approximation to the CI derived from the following test statistic

$$Z_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{\bar{\epsilon}_1 - \bar{\epsilon}_2}{\sqrt{\sigma_{\epsilon_1}^2/n_1 + \sigma_{\epsilon_2}^2/n_2}}$$

Where $\bar{\epsilon}_i = \bar{Y}_i - \mu_i$ and are referred to as the "average errors." The test statistic involving only the errors is known as the **centered test statistic** since we centered each group by subtracting the sample averages with the mean. Notice $\sigma_i^2 = \sigma_{\epsilon_i}^2$ since the transformation $\bar{Y}_i - \mu_i \mapsto \bar{\epsilon}_i$ only shifts the data uniformly, keeping variance the same. Now, in order to estimate a t_s from bootstrapping we need to estimate $\sigma_1^2, \sigma_2^2, \bar{\epsilon}_1$, and $\bar{\epsilon}_2$. If we set

$$\epsilon_{ij} = Y_{ij} - \mu_i$$

it then follows that we can estimate this quantity with $e_{ij} = Y_{ij} - \bar{Y}_i$ and we naturally get

$$\bar{\epsilon}_i = \frac{1}{n_i} \sum_j e_{ij} = \frac{1}{n_i} \sum_j [Y_{ij} - \bar{Y}_i]$$

For the variance, we note

$$\begin{aligned}
 \sigma_i^2 &\approx s_i^2 = \frac{1}{n_i - 1} \sum_j (Y_{ij} - \bar{Y}_i)^2 \\
 &= \frac{1}{n_i - 1} \sum_j ([Y_{ij} - \mu_i] - [\bar{Y}_i - \mu_i])^2 \\
 &= \frac{1}{n_i - 1} \sum_j (\epsilon_{ij} - \bar{\epsilon}_i)^2 \\
 &\approx \frac{1}{n_i - 1} \sum_j (e_{ij} - \bar{e}_i)^2 = s_{e_i}^2 \quad (\bar{\epsilon}_i \approx \bar{e}_i, \epsilon_{ij} \approx e_{ij})
 \end{aligned}$$

We then estimate our test statistic with

$$t_e = \frac{\bar{e}_1 - \bar{e}_2}{\sqrt{s_{e_1}^2/n_1 + s_{e_2}^2/n_2}}$$

Note: $\bar{e}_i = 0$ for our observed sample but for all bootstrapped samples $\bar{e}_i^* \neq 0$ since it is now:

$$\bar{e}_i^* = \frac{1}{n_i} \sum_j (Y_{ij} - \bar{Y}_i)^B$$

So our bootstrap procedure is:

t Method for 2-samples

Step 1 Calculate all residuals $e_{ij} = Y_{ij} - \bar{Y}_i$

Step 2 Create bootstrap samples of the residuals of the data, resampling within groups only

Step 3 Calculate

$$t_{e,i}^B = \frac{\bar{e}_1^* - \bar{e}_2^*}{\sqrt{s_{e_1^*}^2/n_1 + s_{e_2^*}^2/n_2}}$$

based on bootstrap sample

Step 4 Repeat 2-3 B times to obtain the bootstrap distribution

Our CI is then:

$$\left[(\bar{Y}_1 - \bar{Y}_2) - t_{1-\alpha/2}^B \sqrt{s_1^2/n_1 + s_2^2/n_2}, (\bar{Y}_1 - \bar{Y}_2) - t_{\alpha/2}^B \sqrt{s_1^2/n_1 + s_2^2/n_2} \right]$$



Chapter 9

Week 9: Interval Comparisons &
KNN

9.1 Lecture 22: Interval Comparisons

We begin with an example that demonstrates multiple CIs and gives when to use which one.

Example 9.1.1 (Cattle Weight Gain). *The average daily weight gain in pounds for cattle based on two diets "A" and "B" are as follows:*

A:	1.40	1.23	1.02	0.98	1.34	1.36	1.15	1.27
B:	1.16	0.99	1.04	1.02	1.09	1.12	0.76	0.88

With summary statistics

	A	B
mean	1.22	1.01
std. dev.	0.156	0.132
size	8	8

(a) *Why might we consider a non-parametric technique?*

- **Solution:** *The sample size is small and there is no reason to believe that the populations are normal (hard to assess normality).*

(b) *Name three non-parametric techniques we could use to determine if group A tends to be larger than group B.*

- **Solution:** *We could use...*

1. *Permutation test*
2. *WRS/MW*
3. *Bootstrapping*

(c) *Calculate the 95% bootstrap intervals for the difference in means ($\mu_A - \mu_B$)*

- **Solution:** The 95% bootstrap confidence intervals are as follows:

	CI
Percentile	(0.0787, 0.3438)
BCA	(0.0794, 0.3500)
"t"	(0.0583, 0.3698)

(d) Calculate the widths and center for each CI

- **Solution:** Note that we define the width and center for a CI as

$$\text{Width} = (\text{Length of Interval}) = (\text{upper bound} - \text{lower bound})$$

$$\text{Center} = (\text{Midpoint of Interval}) = (\text{upper bound} + \text{lower bound})/2$$

This gives...

	Percentile	BCA	t
Widths:	0.2651	0.2706	0.3115
Center:	0.2135	0.2147	0.2140

If all CIs were appropriate, we could potentially use the widths to pick the best one. Smaller width \implies better interval. Notice that unlike most CIs, these are **not** centered about the sample estimates $\bar{Y}_A - \bar{Y}_B = 1.22 - 1.01 = 0.21$



9.1.1 When to use what type of CI

Some guidelines for choosing which Bootstrap CI to use:

1. Unless your bootstrap distribution of looks perfectly symmetric, using the percentile method is not generally suggested. This is because even in the presence of slight skew, the actual confidence level is often $> (1 - \alpha)100\%$ (more area coverage)
2. When the bootstrap distribution is approximately symmetric, the empirical distribution performs well
3. If skew exists in the bootstrap distribution, the BCA method is preferred since it is designed for those cases
4. If the distribution of is known (close to t or normal), a "t" method may be preferred (bootstrapped t -percentiles are close to actual t -percentiles)

Note: The main problem with bootstrapping is that the type of sample will heavily effect the estimates, CIs, etc... While this is true for most statistical tests, it is especially true for both permutation and bootstrap CIs/Tests. ▲

9.2 Lecture 23: K-Nearest Neighbors (KNN)

9.2.1 K-Nearest Neighbors

This is a "machine learning" technique that is focused on prediction of Y (either continuous or categorical) based on one or more X variables (typically continuous). Prediction can certainly be done with parametric modeling such as logistic or linear regression, but they can be viewed as having some problems:

1. Traditional (parametric) models typically do not outperform "machine learning" techniques. This is because traditional models have a strict framework and can have goals other than prediction. For example, traditional models could want to explain how X affects Y , rather than focus on prediction only.
2. Traditional models do not work for all types of data. If your data trend does not match your model, then your model will not perform well. Some data can't be put in a parametric setting because we lack information, so any parametric model will naturally not work well as assumptions are likely violated.
3. When sample sizes are large, \bar{X} has a small SE and thus statistically all variables are significant. They may not be practically significant on their own, however.

K-Nearest Neighbors (KNN) can be used without a model assumption. There are very little assumptions in general and it is very good at prediction. We do assume a random sample was taken, though. We do not, however, get any information on how X affects Y .

KNN Set-up

Assume all X 's are numerical. For example,

$Y = \text{Netflix Ratings}$

$X = \text{Previous Ratings, Previous Customer's Ratings}$

$Y = \text{House Price}$

$X = \text{Square Footage, Acres, Bedrooms, Distance to major highway, etc...}$

$Y = \text{Cell phones sold per month}$

$X = \text{price of phone, screen size, memory, camera quality, etc...}$

We need (as usual) a dataset $\mathcal{D} = \{(x_i, y_i) : i = 1, \dots, p\}$. Our aim is to predict one new variate we recently sampled given only one of the X 's and Y 's. I.e. predict a new house price given a new square footage. We give some notation:

Notation: Let x_j^* be the new data, with y_j^* unknown. Then, we set

$$D_{ij} = \text{measure of the "distance" between } x_i, x_j^* \quad \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, n^*\}$$

where

n = sample size of "known" data

n^* = sample size of "unknown" y_j^* 's

The main idea behind KNN is "use the K closest" known data points (x_i, y_i) 's to predict our unknown y_j^* . When Y is continuous, we would use the average of the nearest K to predict a new y_j^* , call it y_j^{pred} . ★

Example 9.2.1 (Weight vs. Height). Say we want to predict the weight of a subject based on their height. We have the following data:

Height:	64.5	73.3	68.8	65	69	64.5	66	66.3	68.8	64.5
Weight:	118	143	172	147	146	138	175	134	172	118
D_{ij} :	2.5	6.3	1.8	2.0	2.0	2.5	1.0	0.7	1.8	2.5

Our $x_j^* = 67$ and we want to predict y_j^* . Define $D_{ij} = |x_i - x_j^*|$ (absolute difference).

Let's say we want to use just **one nearest neighbors**. This would be the (x_i, y_i) with the lowest D_{ij} or (x_i, y_i) such that $\min_{i,j}(D_{ij})$ is achieved, which is $(134, 66.3)$ which implies $y_j^{\text{pred}} = 134$.

If we use the **two nearest neighbors**, we have the two smallest pairs of points with the lowest distance as $(134, 66.3)$ and $(175, 66)$ which implies (since weight is continuous) $y_j^{\text{pred}} = (175 + 134)/2 = 154$.

In summary, for different K 's we have

K	1	2	3	4	5	6	7	8	9	10
y_j^{pred}	134	154	160.33	163.25	160	157.67	154.8	150.25	146.67	146.3

Notice that K has a very strong effect on what y_j^{pred} is. ♥

Some Questions

1. What if multiple D_{ij} are tied?

- **Remedy:** You may either randomly select one to be the neighbor or increase K and use all of them

2. What if Y is categorical?

- **Remedy:** Use the highest probability out of all neighbors

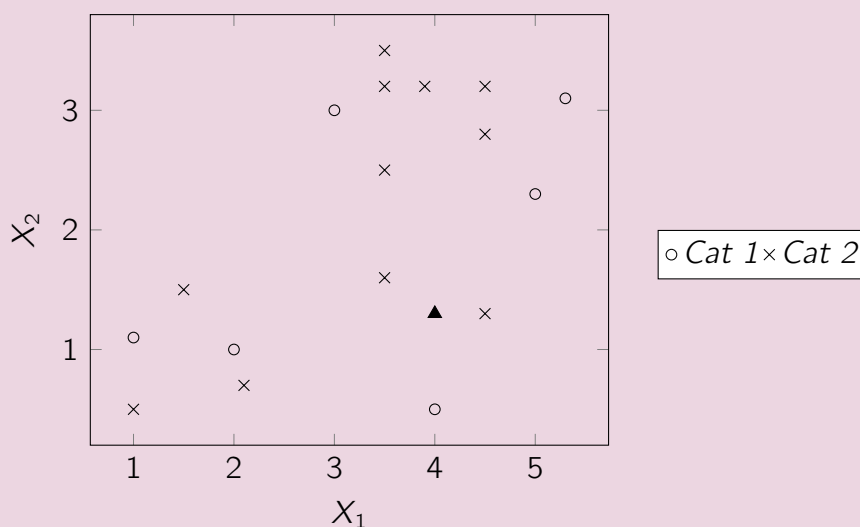
3. What distance should we use?

- **Remedy:** There are two very common measures, if we let there be c predictors, then each observation is really (\vec{x}_i, y_i) and a new observation is \vec{x}_j^* . Our aim is to minimize the distance between \vec{x}_i and \vec{x}_j^* and we can do this using **vector norms**:

- Euclidean: $D_{ij} = \|\vec{x}_i - \vec{x}_j^*\|_2$
- Manhattan: $D_{ij} = \|\vec{x}_i - \vec{x}_j^*\|_1$

We end with an example of using the nearest neighbor approach on categorical data:

Example 9.2.2 (Categorical KNN). Suppose a plot of our data looks like:



Where $Y = \circ, \times$ are the two categorical values Y can take. Then...

- If $k = 1$, the nearest to \blacktriangle is an $\times \implies y_j^{pred} = \times$
- If $k = 2$, the nearest to \blacktriangle is $\times, \times \implies y_j^{pred} = \times$
- If $k = 3$, the nearest to \blacktriangle is $\times, \times, \circ \implies y_j^{pred} = \times$
- If $k = 4$, the nearest to \blacktriangle is $\times, \times, \circ, \circ \implies y_j^{pred} = \times, \circ$ with 50% chance of either ♥

Chapter 10

Week 10: KNN (cont.)

10.1 Lecture 24: More KNN & CV

10.1.1 KNN (cont.)

Notice the K with the most volatility in prediction is $K = 1$ (one neighbor). This means that y_j^{pred} is subject to change the most when we predict with only one neighbor. On the other hand, the K with the least volatility is $K = n$, but will always predict a new observation as \bar{y} , (the sample mean for Y). Notice, if our prediction always changes for new measurements, then it does not effectively measure the relationship between the predictor and response. If the prediction is always the same, then we still haven't captured the true relationship between X and Y as we ignore any variability. So, we need a K that is in between the two. How do we pick that K ?

10.1.2 Cross Validation

In practice, we are highly interested in how well our model predicts new Y values. While we have some ways to this, we ideally could have some new data to test our model.

Cross validation can be used to see how well our data would do in predicting new data, and it is an extension of "leave one out" (jackknife) methods.

10.1.3 f -fold-CV

The idea is relatively simple. To pick the best K for many values of K we split our data (randomly) into f parts. The process is:

f -Fold-CV Process

Say $f = 10$. We "leave out" 1/10 of the data, and use KNN with the remaining (9/10)'s of the data to predict the (1/10)th we left out. Repeat this for every (1/10)th of the data we

left out. We then have

$$y_i = \text{original value of } y$$

$$\hat{y}_i^{CV} = \text{predicted } y_i \text{ using CV}$$

For all $i \in \{1, \dots, n\}$.



The overall CV error is then:

Overall CV Error

$$\sum_{i=1}^n (y_i - \hat{y}_i^{CV})^2$$



And a comparative measure to see if we did any better than \bar{y} is **PRE_{CV}**:

$$PRE_{CV} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i^{CV})^2}{\sum (y_i - \bar{y})^2}$$

= proportion of reduction in error when using
our current technique instead of \bar{y} based on CV

Note: PRE_{CV} is used for continuous Y



The benefit is that the model the (9/10)ths of the data was fit on has no association with the (1/10)th we left out. This gives a better measure of how our model may behave with "new" data.

Remark 10.1.1. CV can be useful in assessing competing "best" models in model selection in techniques or when trying to pick between various other techniques (such as using KNN) ◆

10.1.4 Error for Categorical Y

When we use KNN with categorical Y we don't have a numeric \hat{y} and y to calculate PRE_{CV} . But, we do have an alternative... If we let $y = a_1, a_2, \dots, a_c$ be the varying categories for y . Then, we have $\hat{y} = a_i$ where a_i is the category with the highest chance as the predicted category. The **error matrix** for our y predictions would then be:

Error Matrix (E)

	$\hat{y} = a_1$...	$\hat{y} = a_c$	
$y = a_1$	n_{11}	...	n_{1c}	$r_1 = \text{true total in category } a_1$
\vdots	\vdots	\ddots	\vdots	\vdots
$y = a_c$	n_{c1}	...	n_{cc}	$r_c = \text{true total in category } a_c$
	c_1	...	c_c	$n = \text{sample size}$
	pred. total in cat. a_1	...	pred. total in cat. a_c	



Ideally, in the matrix above we would like the trace (diagonal sum) to be the largest (close to n). So, one measure of the overall error rate would be

$$\text{Overall Error} = \frac{n - \text{trace}(\mathbf{E})}{n} = \frac{n - \sum n_{ii}}{n}$$

The overall rate (proportion) of correct predictions would then be: Correct = 1 - (Overall Error). When $c = 2$ (i.e. y has two categories or is **binary**), we often have more specific error rates.

10.1.5 Errors for Two Categories

For notation, we let $y = 1$ mean **subject has trait** and $y = 0$ mean **subject does not have trait**. The simplified error matrix is then:

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	n_{11}	n_{12}
$y = 0$	n_{21}	n_{22}

where the constraints are

$$n_{ij} \in \{0, \dots, n\} \quad \forall i, j \in \{1, 2\}$$

$$\sum_{i,j} n_{ij} = n$$

i.e. everything sums up to sample size and each entry is no more than the sample size. We then further calculate:

1. Sensitivity = $P(\text{pred. success} | \text{true success}) = P(\hat{y} = 1 | y = 1) = (n_{11}) / (n_{11} + n_{12})$
2. Specificity = $P(\text{pred. fail} | \text{true fail}) = P(\hat{y} = 0 | y = 0) = (n_{22}) / (n_{21} + n_{22})$
3. $P(\text{Correct}) = (n_{11} + n_{22}) / n$

4. $P(\text{Not Correct}) = 1 - P(\text{Correct})$

Of course, since \hat{y} depends on K we would use CV and pick the lowest misclassification rate.

References

- [1] Kornel (<https://stats.stackexchange.com/users/91913/kornel>). *Why is the Kruskal Wallis test statistic approximately chi-square distributed?* Cross Validated. (version: 2018-03-23). eprint: <https://stats.stackexchange.com/q/336241>. url: <https://stats.stackexchange.com/q/336241>.
- [2] James Carpenter and John Bithell. "Bootstrap confidence intervals: when, which, what?" In: *Statistics in Medicine* 19.9 (2000), pp. 1141–1164. url: <https://www.tau.ac.il/~saharon/Boot/10.1.1.133.8405.pdf>.
- [3] *Chapter 14: Nonparametric Statistics*. url: <http://users.stat.ufl.edu/~winner/sta3024/chapter14.pdf>.
- [4] Nian Shong Chok. "PEARSONS VERSUS SPEARMANS AND KENDALLS CORRELATION COEFFICIENTS FOR CONTINUOUS DATA". MA thesis. University of Pittsburgh, 2010. url: https://d-scholarship.pitt.edu/8056/1/Chokns_etd2010.pdf.
- [5] Gregory Imholte. *Better Bootstrap Confidence Intervals*. 2012. url: <https://faculty.washington.edu/heagerty/Courses/b572/public/GregImholte-1.pdf>.
- [6] *Kendall rank correlation coefficient*. url: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient.
- [7] Paul Kvam and Brani Vidakovic. "Nonparametric Statistics with Applications in Science and Engineering". In: chap. 7. url: <http://zoe.bme.gatech.edu/~bv20/isy6404/Bank/npmarginal.pdf>.
- [8] Eric W. Weisstein. *Correlation Coefficient–Gaussian Bivariate Distribution*. url: <https://sanweb.lib.msu.edu/crcmath/math/math/c/c703.htm>.
- [9] Steven J. Wilson. *Confidence Intervals for Percentiles and Medians*. url: <http://www.milefoot.com/math/stat/ci-medians.htm>.