**STA 2183W**

**Chapter 4 – Two-Sample Inference on Quantitative Variables**

In Chapter 2, we explored what it meant to perform inference on one sample (means and medians).   
However, we often wish to compare two populations or two treatments – or more!

The situations which wish to compare two populations/treatments are called *two-sample problems.* Here, our goal is either to compare the responses to two treatments or to compare the characteristics of two populations, with a separate sample from each treatment or each population.

**4.1 Identifying the Target Parameter**

When identifying two-sample problems, we must be careful as to what kind of samples we have. The inference we use will depend on whether we have two independent samples from two populations or whether we have paired samples from one population. The table below summarizes the notation used when we have *two* populations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population** | **Variable** | **Population Parameters** | | | | **Sample Statistics** | | | |
| **Size** | **Mean** | **Median** | **Variance** | **Size** | **Mean** | **Proportion** | **Variance** |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

As Chapter 2, the methodology we will use will depend on the *sample size* as well as the *target parameter*. Some key words and types of data associated with each target parameter are outlined below:

|  |  |  |
| --- | --- | --- |
| Parameter ( | Key Words | Type of Data |
|  | mean difference; difference in averages | Quantitative |
|  | difference between medians, compare medians/distributions | Quantitative |
|  | Compare shifts in distributions | Quantitative |
|  | ratio of variances; difference in variability or spread; compare variation | Quantitative |

**4.2 Comparing Two Variances**

We will only briefly discuss *inference* for two population variances, as its use is more fueled toward performing inference on means. If both populations are normal, then a difference in two variances will lead us to determining which type of t-test we will perform in Section 4.4.

Since variances are never negative, we end up taking the *ratio* of two variances, and estimating this ratio by . Under the assumption (our null hypothesis) that they are equal, this quotient is distributed in what we call an ***F-Distribution***, which contains two different degrees of freedom: a *numerator degrees of freedom* (num df) and a *denominator degrees of freedom* (den df). We will not go too in-depth with F-Tests for the time being and come back to it when we cover ANOVA.

**Hypothesis Test About Difference of Two Variances – Folded-F Test**

|  |
| --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** |
| 1. A simple random sample. 2. Both populations are normally distributed. |
| ***Step 2: State the null and alternative hypotheses.*** |
|  |
| ***Step 3: Compute the value of the test statistic.*** |
| with num df and den df |
| ***Step 4: Determine Rejection Regions / Determine the p-value.*** |
| *p-value:*   |  |  | | --- | --- | | **Two-Sided Test** | **P-Value** | | If |  | | If |  |   *Calculate the p-value from the following:* |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

Using SAS to Test Whether Two Population Variances Are Equal

* You can use **PROC** **TTEST** and add the command CLASS to determine the category for the variable.

**PROC** **TTEST** DATA=**DATASET**;

CLASS **CATEGORY\_OR\_POPULATION**;

VAR **VARIABLE\_DESIRED**;

**RUN**;

Using R to Test Whether Two Population Variances Are Equal

* After testing both variables for normality, we will use **var.test** to see if the variances differ.
* If the variables are split by the variable of interest (Y) and Group Type (X), we can use **Y~X**.

**var.test(Pop1, Pop2)**

**var.test(Y~X)**

**4.3 Comparing Two Means: Paired Difference Experiments**

FINALLY! We will begin our two-sample inference by comparing two means, and.

When we are comparing the two means, one property that has proven to be quite useful in performing inference is: are the two samples ***independent*** or ***dependent***? In Section 4.4, we explore the typical case – we assume the samples are independent. However, we will first explore what happens when the two samples are ***dependent***. These inferences will appear similar to the one-sample inferences you performed in Chapter 2.

When there is a large amount of variability among the experimental units, we often use **blocking**: comparing the different procedures within groups of relatively similar experimental units called ***blocks***. This is done by what is called a **matched pairs design**, where subjects are matched in pairs and given one of two treatments. Taste tests and before and after observations are also called a **matched pairs design** ora **paired difference experiment**.

There are many practical situations in which “pairing” is natural (before/after, similar individuals, etc.). If pairing is possible, it may yield better results than two independent samples (will demonstrate later).

To compare the responses to the two treatments in a matched pairs design, we simply find the **difference**(denoted as )between the responses within each pair. Then we apply the one-sample procedures to the differences to determine if there is a *statistically significant distinction* between the two dependent samples.

**Notation for Differences, :**

|  |  |  |
| --- | --- | --- |
|  | population mean of the   differences, . | population standard   deviation of the differences, |
| number of differences   (pairs) | sample mean of the   differences, | sample standard deviation   of the differences, |

**Note:**

While we can simplify the formula for the sample mean as , the same convention **cannot** be used for the sample standard deviation or variance (). You ***must*** find like you do with a general variance.

To construct a confidence interval for two dependent samples, we first must compute all of the paired differences () between the two samples, and then find the sample mean () and standard deviation () for these paired differences ( of them). Our confidence interval procedures are now similar to the one-sample procedures, with only slightly adjusted assumptions.

|  |
| --- |
| **Confidence Interval for a Mean of Paired Differences, ,  2 Dependent Samples** |
| **Assumptions:**   1. A simple random sample of differences selected from the target population of differences. 2. The population of differences is approximately normally distributed. |
|  |

**Note:**

More often than not, when we construct a confidence interval for the difference of two populations, , we are looking to see if differs in some way from (, , or ).

* If zero is *not* in the confidence interval, then evidence suggests that there so difference between the two means (and so either or ).
* If zero is in the confidence interval, then we believe that is a possible value for the true value of , and so it appears there is not sufficient evidence to conclude any difference exists.

With hypothesis tests, we will test the null hypothesis , where is the hypothesize average difference between the matched pairs.

* More often than not, we are only looking to see if differs in some way from (, , or ). In such cases, we take , and so:  
   is equivalent to writing it as:   
  (and so we will have either (1) , (2) , or (3) )
* Again, assumptions are the same as they were for the one-sample procedures for estimating a mean, only here we treat the differences as our observations. We also find rejection regions and p-values in the usual manner.

**Hypothesis Test About a Mean of Paired Differences , 2 Dependent Samples**

|  |  |  |
| --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | |
| 1. A simple random sample of differences selected from the target population of differences. 2. The population of differences is approximately normally distributed | | |
| ***Step 2: State the null and alternative hypotheses.*** | | |
| **Right-sided Test** | **Left-sided Test** | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | |
|  | | |
| ***Step 4: Determine the p-value.*** | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *p-value:*   |  |  | | --- | --- | | **Test** | **t-Tests** | | **Right-Sided** |  | | **Left-Sided** |  | | **Two-Sided** |  |   *Calculate the p-value from the following below:* |

|  |
| --- |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

|  |
| --- |
| ***Note:*** *The test is exact for normal populations and is approximately correct  for large samples from non-normal populations.* |

Using SAS to Perform Hypothesis Test About Paired Difference Means, 2 Dependent Samples

* You can used the **PAIRED** command in the **PROC TTEST** procedure to get a paired test.

**PROC** **TTEST** DATA=**DATASET**;

PAIRED **VAR1**\***VAR2**;

**RUN**;

* Alternatively, you can create a new variable for the differences, and then perform   
  **PROC UNIVARIATE** or **PROC TTEST** on the differences variable. Note you will most likely need to create the differences variable anyways to check normality on the differences.

**DATA** **NEW\_DATASET**;

SET **DATASET**;

diff = **VAR1** – **VAR2**;

**RUN**;

**PROC** **UNIVARIATE** DATA=**NEW\_DATASET** NORMAL;

VAR diff;

**RUN**;

Using R to Perform Hypothesis Test About Paired Difference Means, 2 Dependent Samples

* You can used the **paired** option in the **t.test** procedure to get a paired test.

**t.test(X, Y, alternative= "two.sided"/"less"/"greater", paired=TRUE, conf.int=CONF\_LEVEL)**

* Alternatively, you can create a new variable for the differences, and then perform   
  **t.test** on the differences variable. Note you will most likely need to create the differences variable anyways to check normality on the differences.

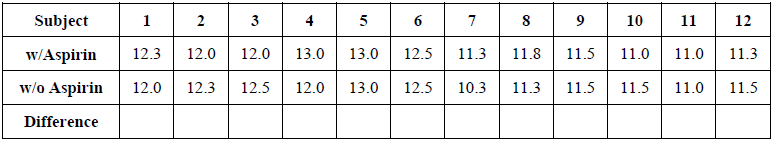
**diff <- X - Y**

**shapiro.test(diff)**

**boxplot(diff)**

**t.test(diff, mu=0, alternative= "two.sided"/"less"/"greater", conf.int=CONF\_LEVEL)**

**Example 4.1:**

****Blood clotting occurs due to a sequence of chemical reactions. The protein thrombin initiates blood clotting by working with another protein, prothrombin. It is common to measure an individual’s blood clotting time through prothrombin time--the time between the start of the thrombin-prothrombin reaction and the formation of the clot. Researchers wanted to study the effect of aspirin on prothrombin time. They randomly selected 12 subjects and measured the prothrombin time (in seconds), first without taking aspirin and again three hours later after taking two aspirin tablets. The data they obtained is included in the table below. Test the claim that aspirin increases the average time it takes for a blood clot to form.   
Use .

|  |
| --- |
| SAS Code: (Create Data Set BloodClot)  **DATA** BC2;  SET BloodClot;  Diff = WAspirin - WOAspirin;  **RUN**;  **PROC** **UNIVARIATE** DATA=BC2 NORMAL;  VAR Diff;  ODS SELECT TestsForNormality;  **RUN**;  **PROC** **SGPLOT** DATA=BC2;  VBOX Diff;  **RUN**;  **PROC** **TTEST** DATA=BloodClot;  PAIRED WAspirin\*WOAspirin ;  **RUN**;  **PROC** **TTEST** DATA=BC2;  VAR Diff;  **RUN**; |
| R Code: (Create Data Set BloodClot)  **BloodClot$diff <- BloodClot$WAspirin - BloodClot$WOAspirin**  **shapiro.test(BloodClot$diff)**  **boxplot(BloodClot$diff)**  **t.test(BloodClot$WAspirin, BloodClot$WOAspirin, alternative= "greater", paired=TRUE)**  **t.test(BloodClot$diff, mu=0, alternative = "greater")** |

**4.4 Comparing Two Population Means: Independent Sampling**

In the previous section, we examined two populations that are *dependent* in a concept called ***paired difference experiments***. We will now approach the topic of inference on ***two independent samples***.

When we compare *two population means*, Section 4.1 described that we wish to examine the *difference in population means*, say: . Thus, we get the ***point estimate*** for this parameter

|  |
| --- |
| ***Point Estimate of*** : |

Just like performing inference on *one-sample* means and proportions, we cannot perform any inferential methods until we have the sampling distribution of our point estimate.  
Below describes the *sampling distribution* of :

**The Sampling Distribution of :**

|  |
| --- |
| 1. The **mean** of the sampling distribution is: . 2. If the two samples are ***independent***, the **standard** **deviation** of the sampling distribution is: 3. The sampling distribution of is approximately normal for *large samples*, brought to us by the ***Central Limit Theorem***. |

**Notes:**

1. To obtain , note that for two random variables and :
2. When the samples are independent, then for two random variables and :  
     
   * Take the square root of both sides to get the standard deviation we have above.

When we have small samples, the Central Limit Theorem cannot be used, and so we must resort to using -procedures like we performed in Chapter 2. However, with two-sample procedures, there are various routes to construct confidence intervals and to perform hypothesis tests, each route based on certain conditions and assumptions about the structure of the problem. In particular, there are two possible -procedures:

1. The variances are the equal. ()
2. The variances aren’t equal ().

It is in the first case (the special case where the variances are equal) that the properties of the   
-distribution are actually adequately upheld. When we have independent samples with the same variances, then the inference on is a measure of distance (upheld by the normal distribution).

Because the variances are equal, we can construct what is called a **pooled sample estimator for ()** for use in confidence intervals and hypothesis tests. We can think of as the “weighted average” of the two sample variances, with the following formula:

Since we have independent samples, the degrees of freedom can be added :

The two-sample *t-*statistic is quite powerful and ***robust –*** even more than one-sample -procedures. While the above conditions are required to perform a true -procedure, we can relax the criteria to approximate normality with little disruption.

However, when the variances are not equal, as in situation, we cannot simply add the degrees of freedom together – such degrees of freedom are typically too large. However, in 1946, a formula was constructed to determine the approximate the degrees of freedom that fit such two-sample sampling distributions.

This construction is known as the **Satterthwaite Approximation** for the degrees of freedom (**)**:

We will resort back to the approximate standard error:

**Notes:**

* Do not fear – such computations are rarely seen to be something to do by hand.   
  We would rely on computer print-outs if we were to require such inference.

So, to perform confidence intervals, we have two options:

|  |  |
| --- | --- |
| **Independent Samples Confidence Interval (CI) for** | |
|  |  |
| **Assumptions:**   1. The samples are independently chosen random samples from the two populations. 2. Both sampled populations are approximately normal. 3. The two population variances are equal: | **Assumptions:**   1. The samples are independently chosen random samples from the two populations. 2. Both sampled populations are approximately normal. |
|  |  |
|  |  |

For hypothesis tests, we have the following procedures, with as the hypothesized difference from :

**Independent Samples Test for : Student’s t-Statistic**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | | |
|  | |  | |
| 1. The samples are independently chosen random samples from the two populations. 2. Both sampled populations are approximately normal. 3. The two population variances are equal: | | 1. The samples are independently chosen random samples from the two populations. 2. Both sampled populations are approximately normal. | |
| ***Step 2: State the null and alternative hypotheses.*** | | | |
| **Right-sided Test** | **Left-sided Test** | | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | | |
|  | |  | |
|  | |  | |
| ***Step 4: Determine the p-value.*** | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Test** | **t-Tests** | | **Right-Sided** |  | | **Left-Sided** |  | | **Two-Sided** |  |   *p-value:*  *Calculate the p-value from the following below:* |

|  |
| --- |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

Using SAS to Perform Inference About Difference in Means, 2 Independent Samples

* We will use PROC UNIVARIATE and PROC SGPLOT to determine normality, and then use   
  PROC TTEST to perform inference (CI and Hypothesis Tests). We will include the command CLASS to separate our observations and compare differences.

**PROC** **UNIVARIATE** DATA=**DATASET** NORMAL;

CLASS **VARIABLE\_FOR\_GROUP\_TYPE**;

VAR **VARIABLE\_DESIRED**;

ODS SELECT testsfornormality;

**RUN**;

**PROC** **SGPLOT** DATA=**DATASET**;

VBOX **VARIABLE\_DESIRED** / group=**VARIABLE\_FOR\_GROUP\_TYPE**;

TITLE 'TITLE';

**run**;

**PROC** **TTEST** DATA=**DATASET**;

CLASS **VARIABLE\_FOR\_GROUP\_TYPE**;

VAR **VARIABLE\_DESIRED**;

**RUN**;

Using R to Perform Inference About Difference in Means, 2 Independent Samples

* First, be sure to check for normality and to check for variances.
* If the data are separate into their own columns, you can use the following **t.test** procedure:

**t.test(X, Y, var.equal = TRUE/FALSE, alternative= "two.sided"/"less"/"greater",   
conf.int =CONFIDENCE\_LEVEL)**

* Alternatively, if the data is not separated into columns (such as when the data is melted), you can use this **t.test** procedure:

**t.test( RESPONSE\_VAR ~ GROUP\_VAR, var.equal = TRUE/FALSE, alternative= "two.sided"/"less"/"greater", conf.int=CONF\_LEVEL)**

[Note: You can choose to use TRUE or FALSE for the **var.equal** option to determine whether to use the Pooled or Satterthwaite variance.]

**Example 4.2:**

In an attempt to determine if two competing brands of cold medicine contain, on the average, the same amount of acetaminophen, twelve different tablets from each of the two competing brands were randomly selected and tested for the amount of acetaminophen each contains. The results (in milligrams) follow. Use a significance level of 0.01.

Brand A Brand B

517, 495, 503, 491 493, 508, 513, 521

503, 493, 505, 495 541, 533, 500, 515

498, 481, 499, 494 536, 498, 515, 515

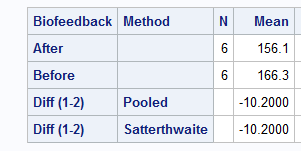
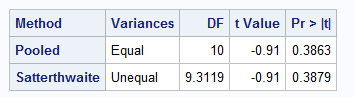
|  |
| --- |
| SAS Code: (Create Data Set cold\_meds)  **PROC** **UNIVARIATE** DATA=cold\_meds NORMAL;  CLASS brand;  VAR aceta;  ODS SELECT testsfornormality;  **RUN**;  **PROC** **SGPLOT** DATA=cold\_meds;  VBOX aceta / group=brand;  TITLE 'Amount of Acetaminophen in Cold Medicine';  **run**;  /\* Perform Inference for Means for Two Independent Samples \*/  **PROC** **TTEST** DATA=cold\_meds ALPHA=**0.01**;  CLASS brand;  VAR aceta;  **RUN**; |
| R Code: (Create Data Set cold.meds)  **shapiro.test(cold.meds$aceta[cold.meds$brand=="A"])**  **shapiro.test(cold.meds$aceta[cold.meds$brand=="B"])**  **boxplot(cold.meds$aceta[cold.meds$brand=="A"],cold.meds$aceta[cold.meds$brand=="B"], main="Amount of Acetaminophen in Cold Medicine",**  **xlab="Brand", ylab="Amount of Acetaminophen")**  **var.test(cold.meds$aceta[cold.meds$brand=="A"], cold.meds$aceta[cold.meds$brand=="B"])**  **t.test(cold.meds$aceta[cold.meds$brand=="A"], cold.meds$aceta[cold.meds$brand=="B"],**  **var.equal = T, conf.int =0.99 )**   * Since Populations are Not Separated by Columns, we can instead use Y~X and the **tapply** procedure to get all *factors* at once. (More when we get to ANOVA).   **tapply(cold.meds$aceta, cold.meds$brand, shapiro.test)**  **boxplot(cold.meds$aceta~cold.meds$brand, main="Amount of Acetaminophen in Cold Medicine", xlab="Brand", ylab="Amount of Acetaminophen")**  **var.test(cold.meds$aceta~cold.meds$brand)**  **t.test(cold.meds$aceta~cold.meds$brand,var.equal = T, conf.int =0.99 )** |

**Example 4.3 – Why *Independent* vs *Dependent* Matters**

In the past, many bodily functions were thought to be beyond conscious control. However, recent experimentation suggests that it may be possible for a person to control certain body functions if that person is trained in a program of *biofeedback* exercises. An experiment is conducted to show that blood pressure levels can be consciously reduced in people trained in this program. The blood pressure measurements (in millimeters of mercury) listed in the table represents the readings before and after the biofeedback training of six subjects.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Before* | 136.9 | 201.4 | 166.8 | 150.0 | 173.2 | 169.3 |
| *After* | 130.2 | 180.7 | 149.6 | 153.2 | 162.6 | 160.1 |

Is there evidence that the biofeedback exercises help reduce blood pressure? Test using .

* ****If we treat these as independent samples, we get the following results with SAS:

**Conclusion:** “No differences in before-and-after means.” Does this seem reasonable?

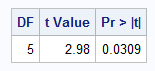
* The problem is that we treated these as independent samples, and they aren’t!!

***The independent samples t-test is not a valid procedure to use with this set of data.***

The measurements were taken on the **same** person, and we chose each person at random.

We do not have two independent random samples; we have dependent samples which are paired together.

We also have a large variation within samples, as evidenced by the large value of standard deviations incomparison to the relatively small differences between the sample means. Because the variability is so large, the independent-sample t-test can’t determine a difference between and. The difference in blood pressures of the two groups could be due to the biofeedback training, or it could just be that the two groups had different average blood pressures to start with.

* That’s why it’s better to take the *before-and-after* measurements.
* Now, perform the **matched-pairs** t-test like we did in Section 4.3.  
    
  We get the following SAS read-out:

What is the conclusion?

|  |
| --- |
| **What Should You Do If No Set of Assumptions Can Be Satisfied?**  If you are concerned that none of the possible assumptions are satisfied, you can either increase your sample sizes, or use the ***Wilcoxon Rank Sum Test*** for independent samples to test for a shift in population distribution. |

**4.5 Wilcoxon Signed-Rank Test**

We will now explore some ***Nonparametric*** procedures using two-samples with the data.

The first test is when we have ***dependent samples***. Note that when the two samples were dependent, we could take the differences and treat the dataset like a one-sample case and detect a shift in the distribution when the mean difference change. However, even if we have a one-sample problem, we still would need the distribution to be *approximately normal* for us to use a t-test procedure.

We could resort back to the sign test. However, even if normality is not available, there is one thing we should be able to expect from the differences: its distribution is *symmetric*.

The general reasoning draws from the assumption that there is no difference between the pairs, and so their distributions are the same. In general, if we have two distributions that are the same, the distribution of their differences will be symmetric around 0.

When a distribution is symmetric, we can up our inference levels with a test that works with symmetric distributions: ***Wilcoxon Signed-Rank Test***.

The procedure for the Wilcoxon Signed-Rank Test follows as such.

1. Calculate the differences in the pairs of observations
2. Subtract the null hypothesis value from all the differences (usually, ).
3. Delete all the zero values. Let be the number of nonzero values.
4. List the *absolute values* of the differences in increasing order, and assign them ranks  
    (or the average of ranks for ties).

Thus, if we had the following differences from Step 2:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Obs.** | **1** | **2** | **3** | **4** | **5** | **6** |
| **Values,** |  |  |  |  |  |  |

Then, we can take their absolute values and rank them accordingly.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Obs.** | **1** | **2** | **3** | **4** | **5** | **6** |
| **Values,** |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  | 3.2 |  |
| **Ranks** |  |  |  |  |  |  |

Let a *positive rank* be rank above that comes from a difference that is positive, and a *negative rank* be one that is negative. We will then use the ***test statistic*** {sum of positive ranks}.

So, in the above table, observations 2, 4, and 6 give us positive differences, and so the sum of their *positive ranks* is .

Now, if all the ranks were *negative*, then . If all the ranks were *positive*, then we’d get

Now, under the null hypothesis, there is no difference between the paired values, and so we would expect for every large value on one side, there would be another value equally large on the other side.   
Thus, we believe that the average sum for would be

The ***Wilcoxon Signed-Rank Test*** in general is a test that measures a *shift in location*, which would imply that there is a difference with the distribution.

With the ***Paired t-Test***, we are doing the same thing, but since we have normality, using the mean differences instead provides us with the same thing, as the distribution is symmetric around the mean.

When using R, we Again, SAS is a jerk and doesn’t give you . Instead, it gives you something analogous:

where is what we expect to be if there is no shift in location. We can now summarize our test:

**Wilcoxon Signed-Rank Test About Paired Differences and Shift in Location, 2 Dependent Samples**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | | |
| 1. A simple random sample of differences selected from the target population of differences. 2. The population of differences is continuous. | | | |
| ***Step 2: State the null and alternative hypotheses.*** | | | |
| **Right-sided Test** | **Left-sided Test** | | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | | |
| **Test** | | **Two-Sided Test with SAS** | |
| {sum of positive ranks} | |  | |
| ***Step 4: Determine the p-value.*** | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *p-value:*   |  |  |  |  | | --- | --- | --- | --- | | **Test** | **P-Value** | **Test** | **P-Value** | | **Left-Sided** |  | **Right-Sided** |  | | **Two-Sided** |  | **Two-Sided (SAS)** |  |   *Calculate the p-value from either or.*  *Let* |

|  |
| --- |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

Using SAS to Perform Inference Using the Wilcoxon Signed-Rank Test

* We will use PROC UNIVARIATE to determine normality non-normality, and then use it to perform inference (CI and Hypothesis Tests).

**DATA** **NEW\_DATASET**;

SET **DATASET**;

diff = **VAR1** – **VAR2**;

**RUN**;

**PROC** **UNIVARIATE** DATA=**NEW\_DATASET** NORMAL CIPCTLDF;

VAR diff;

**RUN**;

Using R to Perform Inference Using the Wilcoxon Signed-Rank Test

* If the data are separate into their own columns, you can use the following **wilcox.test** procedure. Make sure to choose *TRUE* for **paired**, and *FALSE* for **exact**, so that it will still calculate values, even when there are differences that are 0:

**wilcox.test(X, Y, alternative="less/greater/two.sided", paired=T, exact=F)**

**Example 4.4:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Field | Brand A | Brand B | Field | Brand A | Brand B |
| 1 | 211.4 | 186.3 | 11 | 208.9 | 183.6 |
| 2 | 204.4 | 205.7 | 12 | 208.7 | 188.7 |
| 3 | 202.0 | 184.4 | 13 | 213.8 | 188.6 |
| 4 | 201.9 | 203.6 | 14 | 201.6 | 204.2 |
| 5 | 202.4 | 180.4 | 15 | 201.8 | 181.6 |
| 6 | 202.0 | 202.0 | 16 | 200.3 | 208.7 |
| 7 | 202.4 | 181.5 | 17 | 201.8 | 181.5 |
| 8 | 207.1 | 186.6 | 18 | 201.5 | 208.7 |
| 9 | 203.6 | 205.7 | 19 | 212.1 | 186.8 |
| 10 | 216.0 | 189.1 | 20 | 203.4 | 182.9 |

A city park developer compared a new formulation of a fertilizer, Brand A, to the previously used fertilizer Brand B, on each of 20 different softball fields. Each field was divided in half, with brand A randomly assigned to one half and brand B to the other. The effect of the fertilizer on the grass grown at each field was measured by the weight (in pounds) of grass clippings produced by mowing the grass at the fields over a 1-month period.

Evaluate whether Brand A tends to produce more grass than Brand B at the 5% significance level.

|  |
| --- |
| SAS Code: (Create Data Set fertilizer with the *diff* variable)  **PROC** **UNIVARIATE** DATA=fertilizer Normal CIPCTLDF;  VAR diff;  **RUN**; |
| R Code: (Create Data Set *fertilizer*)  **diff=fertilizer$BrandA-fertilizer$BrandB**  **shapiro.test(diff)**  **wilcox.test(fertilizer$BrandA,fertilizer$BrandB,alternative="greater",paired=T, exact=F)** |

Further Asides:

Before we jump into the next section, we will address a few extra concepts that fit well with the introduction of the ***Signed-Rank Test***.

|  |
| --- |
| Aside 1: *The Signed-Rank Test can work as a one-sample test as well.* |

While we introduced the Signed-Rank Test in the two-sample-procedure realm, note we are testing the differences, and so we are again drawing back to one-sample procedures.

Thus, ***if the distribution of a sample is symmetric***, then we can opt to perform the Signed-Rank Test on the null hypothesis . More on why it is optimal to do so in *Aside 3*.

|  |
| --- |
| Aside 2: *The Signed-Rank Test can actually test for the mean.* |

Typically, for nonparametric tests, we don’t really look at parametric values like the mean and standard deviation. However, since the Signed-Rank Test relies on the distribution being *symmetric*, we can rely on an old property about distributions:

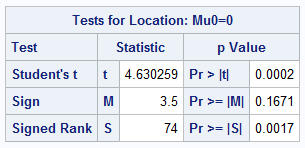
* When a distribution is symmetric, the mean and median are equal.

The property above allows us to provide a little more information when performing this test, since implies that testing is actually simultaneously test whether

|  |
| --- |
| Aside 3: *Guidance on Usage for the t-test, sign test, and signed-rank test.* |

In your introductory statistics classes, much of the discussion of t-tests and the like were used for the intention to make inference on the mean. However, just because we use it for such doesn’t necessarily imply that the mean is the only thing for which the t-test is used.

As mentioned when we use PROC UNIVARIATE in SAS, the t-test, the sign test, and the signed-rank test are all tests for the *location* of the distribution, focusing on the central value (the mean or median) to help detect the center of the distribution. However, depending on the assumptions provided, we end up using different tests. To help with our decision, let’s look at the SAS results for the three tests for the previous example:



As you can see, the p-values differ a bunch, and these differences are impacted by the strength of the test. As our assumptions begin to narrow the types of distributions we have, the *power* of the test improves.

As mentioned in Chapter 2, a more powerful test will be able to detect difference when ones exist.

* At , we can reject for the t-test and signed-rank test, but not the sign test.
* At, we can reject for the t-test but not the signed-rank test or the sign test.

This situation reinforces that the assumptions of our distribution (or the increase of our sample size) plays an important factor in performing our hypothesis tests, especially between our parametric and nonparametric tests. Without confirming these assumptions, tests like the t-test provide results that could be worthless (and provide huge Type-I errors), and we need to rely on the tests like the sign test that works regardless of assumptions beyond a continuous distribution. However, with each improvement in our set of assumptions, we get to reject the null more often, which is ideal.

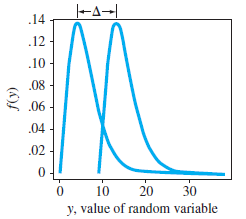
Thus, when an assumption has been met, we want to use a more-powerful test in our statistical analysis. In summary, we will rely on the table below when performing tests that are either one-sample or dependent two-sample:

|  |  |  |
| --- | --- | --- |
| **Situation** | **If the distribution is …** | **Then use the following test:** |
| **1** | Normal (or a large sample) | t-test |
| **2** | Symmetric, but non-normal | Signed-rank test |
| **3** | Non-symmetric or unknown | Sign test |

**4.6 Rank Sum Test**

We will now delve into the situation where we have ***independent samples***. When normality could be assumed, we were able to use the two-sample t-tests. However, when normality cannot be assumed for both populations, we have to resort to nonparametric procedures.

Recall our discussion with *shift in populations* when we discussed one-sample procedures in Section 4.5. When we test for *difference in means*, we do so because we consider the means to be the centers of our population (yay for Normal distributions!). When the mean isn’t reliable to use, we need to take a step back and think about that shift with some new terminology.



**Definition:**

Source:  
 Ott & Longnecker

A distribution is ***shifted by a factor***  to a variable two variables share the same distribution shape, but with a different location.

When testing for a shift of distribution , we can utilize the ***Wilcoxon Rank Sum Test***. This test either goes by the names or has analogous versions called the ***Mann-Whitney U Test*** and ***Mann-Whitney-Wilcoxon Test***.

Note that our null hypothesis would be that the populations are the same, and so , while different implies .

For Population 1 to be shifted to the right of Population 2 suggests and to the left implies .

For populations with sample sizes and , set

The procedure for the Wilcoxon Rank Sum Test follows as such.

1. List the data values in the combined data set from smallest to largest.
2. In the next column, assign the numbers 1 to to the largest value. These are the ***ranks*** of the observations.
3. If there are ties in the combined data set, average the corresponding ranks and assign that average to the given observations.
4. Let denote the sum of the ranks for the observations from Population 1.

Let’s look at the following example, if we had the following differences from Step 2:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sample 1** | | | **Sample 2** | | | |
| **Obs.** | **1** | **2** | **3** | **1** | **2** | **3** | **4** |
| **Values** |  |  |  |  |  |  |  |

Then, we can take their absolute values and rank them accordingly.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs.** | **2** | **2** | **3** | **3** | **1** | **1** | **4** |
| **Ordered Values** |  |  |  |  |  |  |  |
| **Ranks** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Let {sum of ranks from population 1} and {sum of ranks from population 2}.

So, in the above table, , and .

We will then use the ***test statistic*** {sum of ranks from population 1}.

Now, if all the observations from Sample 1 were *smaller* than Sample 2, the first ranks would come from the first sample, and so then we’d get

Consequently, we get a maximum for at .

Now, under the null hypothesis, there is no difference between the two populations, and so we would expect that when we combine our samples, the observations would be mixed. Thus, we believe that the average sum for would be:

(If the sample sizes and are the same, then the mean is just half the sum of the ranks.)

We can obtain confidence intervals by using a nonparametric procedure called the ***Hodges-Lehmann*** confidence interval, which uses the sampling distribution obtained by our method.

The ***Wilcoxon Rank Sum Test*** in general is a test that measures a *shift in location*, which would imply that there is a difference with the distribution. We use the test statistic to determine if the ranks are radically different from the expected value when then distributions are identical.

Note:

To help determine whether the assumption that the distributions are identical in shape but shifted in location, we can utilize prior information or graphical utilities like boxplots and histograms.

Of course, the two programming languages cannot provide us with the same test statistic – *that would be too easy*! While SAS produces , this time, R is the jerk and doesn’t give you . Instead, it gives you something analogous:

where is the *smallest* that could be (if all observations from Population 1 were smaller than Population 2), shifting our values so that starts at 0. (This statistic is not the same as from the Shapiro-Wilk Test.)

(*Sometimes… convention is not exactly convenience…*)

To determine our p-values, we can use the ***Exact p-value***, but for a larger number of observations, the computing capacity gets a bit disruptive. Practice has shown that if *and* , we can take our test statistic, *normalize it to a z-score*, and get an **approximate** **p-value** through a normal approximation. We can now summarize our test below:

**Wilcoxon Rank Sum Test About a Shift in Location, 2 Independent Samples**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** | | | |
| 1. Two independent Simple Random Samples. 2. The distributions for each population are identical in shape, but shifted in location. | | | |
| ***Step 2: State the null and alternative hypotheses.*** | | | |
| **Right-sided Test** | **Left-sided Test** | | **Two-sided Test** |
| ***Step 3: Compute the value of the test statistic.*** | | | |
| SAS Test Statistic | | R Test Statistic | |
| {sum of ranks from Population 1} | |  | |
| ***Step 4: Determine the p-value.*** | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *p-value:*   |  |  |  |  | | --- | --- | --- | --- | | **Test** | **SAS P-Value** | **Test** | **R P-Value** | | **Left-Sided** |  | **Left-Sided** |  | | **Right-Sided** |  | **Right-Sided** |  | | **Two-Sided** |  | **Two-Sided** |  |   *Calculate the p-value from .*  *or*  – Use Exact Test  *and*  – Use Normal Approximation (- |

|  |
| --- |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

Using SAS to Perform Inference Using the Wilcoxon Signed-Rank Test

* We will use **PROC UNIVARIATE** to determine normality non-normality, and then use **PROC NPAR1WAY** to perform inference. We will use the Hodges-Lehmann CI using the **HL** option.

**PROC** **NPAR1WAY** DATA=**DATA\_SET** WILCOXON HL CORRECT=no;

CLASS **VARIABLE\_FOR\_GROUP\_TYPE**;

VAR **VARIABLE DESIRED**;

EXACT Wilcoxon;

**RUN**;

[Note: We will use the **Z p-values** when and , and **Exact p-values** otherwise.]  
  
Using R to Perform Inference Using the Wilcoxon Signed-Rank Test

* You can use the following **wilcox.test** procedure. Make sure to choose *FALSE* for **paired**.
  + If the data are separated into their own columns/row vectors, use **X, Y**.
  + If the categories are merged into one variable, you can use **Y~X** instead.

**wilcox.test(X, Y, alternative="less/greater/two.sided", paired=FALSE, exact=T/F,** **conf.int =0.95)**

**wilcox.test(Y~X, alternative="less/greater/two.sided", paired=FALSE, exact=T/F,** **conf.int =0.95)**

[Note: For the option **exact**, we use *TRUE* (*T*) when and , and FALSE (*F*) otherwise.]

**Example 4.5:**

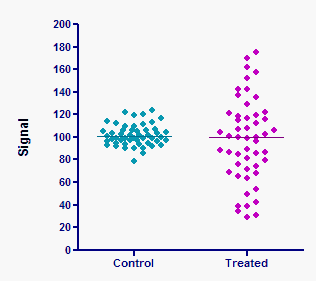
An investigator designed a test to study the effect of alcohol on driving reaction times. Ten participants consumed a specified amount of alcohol and another group of ten participants consumed the same amount of a non-alcoholic drink, a placebo. The 2 participants’ average reaction times (in seconds) to a series of simulated driving situations are reported in the following table:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Placebo | 0.90 | 0.37 | 1.63 | 0.83 | 0.95 | 0.78 | 0.86 | 0.61 | 0.38 | 1.97 |
| Alcohol | 1.46 | 1.45 | 1.76 | 1.44 | 1.11 | 3.07 | 0.98 | 1.27 | 2.56 | 1.32 |

Does it appear that alcohol consumption increases reaction times?

|  |
| --- |
| SAS Code: (Create Data Set *reaction* with *group* for placebo/alcohol, and *time* for the observation)  **PROC** **UNIVARIATE** DATA=reaction NORMAL;  CLASS group;  VAR time;  ODS SELECT TestsForNormality;  **RUN**;  **PROC** **NPAR1WAY** DATA=reaction WILCOXON HL CORRECT=no;  CLASS group;  VAR time;  EXACT Wilcoxon;  **RUN**; |
| R Code: (Create Data Set *reaction*)  **shapiro.test(reaction$placebo)**  **shapiro.test(reaction$alcohol)**  **wilcox.test(reaction$placebo, reaction$alcohol, alternative="less", paired=F, exact=T)** |

|  |
| --- |
| **Beware:**  The procedure we are performing only works if the distributions are identical in shape but shifted in a given direction. *This procedure does not detect whether the distributions**differ*. This distinction is important if the centers are the same, but the spreads differ. If such is the case, and you wish to test whether the distributions *differ*, you will need to perform a ***Kolmogorov-Smirnov Test for two samples.*** |

**

Brief Aside – Kolmogorov-Smirnov Test

Source: Graphpad

We will not spend much time on this situation, but from the warning above, note that the Rank Sum test is typically good for detecting shifts in location, but not a distribution in general.

For example, in the photo on the right, the two distributions are clearly different, but the Rank Sum test would not detect a difference in location. Instead, we would test using the **Kolmogorov-Smirnov (KS) Test**. This test simply uses the distributions of the data to determine if they differ.

Here are some guidelines for choosing between the two tests:

* The KS test is sensitive to ***any*** differences in the two distributions. Substantial differences in shape, spread or median will result in a small p-value, and so it will reject the null hypothesis more often.
* In contrast, the Rank Sum test is mostly sensitive to changes in the median or location, and ***should*** be used if you are looking to check for differences in those factors.
* The Rank Sum test is used more often and is recognized by more people, and has been extended to handle tied values. The KS test does not handle ties so well.

**Kolmogorov-Smirnov Test About a Difference in Distributions, 2 Independent Samples**

|  |
| --- |
| ***Step 1: State the assumptions and make sure they are satisfied.*** |
| 1. Two independent Simple Random Samples. 2. The distributions for each population are continuous. |
| ***Step 2: State the null and alternative hypotheses.*** |
|  |
| ***Step 3: Compute the value of the test statistic.*** |
|  |
| ***Step 4: Determine the p-value.*** |

|  |  |  |
| --- | --- | --- |
| *p-value:*   |  | | --- | | **P-Value** | |  |   *Calculate the p-value from .* |

|  |
| --- |
| ***Step 5: Based on p-value, make your conclusion.*** |
| *P-value:*   * If , *Reject* . * If , *Fail to Reject .* |

Using SAS to Perform Inference Using the Kolmogorov-Smirnov Test

* We will use **PROC UNIVARIATE** to determine normality non-normality, and then use **PROC NPAR1WAY** to perform inference.

**PROC** **NPAR1WAY** DATA=**DATA\_SET** EDF CORRECT=no;

CLASS **VARIABLE\_FOR\_GROUP\_TYPE**;

VAR **VARIABLE DESIRED**;

EXACT EDF;

**RUN**;

Using R to Perform Inference Using the Kolmogorov-Smirnov Test

* You can use the following **ks.test** procedure.
  + If the data are separated into their own columns/row vectors, use **X, Y**.
  + If the categories are merged into one variable, you can use **Y~X** instead.

**ks.test(X, Y, alternative="less/greater/two.sided", exact=NULL)**

**ks.test(Y~X, alternative="less/greater/two.sided", exact=NULL)**