Tests for One and for Two Samples

Introduction

- We have mentioned in Topic 4 – Describing Numerical Data (slide 6), that not only descriptive statistics (in the form of numerical or graphical) can be used to make inference about data but **inferential statistics** (such as hypothesis testing or significance test) also can be used for the same purpose.
- In this course, hypothesis testing has been roughly described in Topic 6 – Categorical Data Analysis (chi-square test for 2×2 contingency table).
- Just introduce the tests and how to perform these tests using software might not be very helpful in helping students to understand the meaning of the test, hence a quick introduction about hypothesis testing will be given.
- For students that have learnt details about steps to perform a test, you can ignore this theory part.
- In general, two situations will be introduced, that is when we have only one sample (or so called **one-sample data**) and when we have two samples (or **two-sample data**).

Hypothesis

- In many studies, we want to check if our data support a certain statement about a population. These statements are hypotheses about the population.
- **Definition:** A hypothesis is a statement about population, usually claiming that a parameter takes a particular numerical value or falls in a certain range of values.
- First example: In the up coming election in US, we may be interested in making a statement about the proportion (p) of residents in a specific state who will vote for Republic. A random sample of n voters from this state is collected and asked which party they would vote for. In this case, we want to see which one of the following hypotheses our data provide evidence for:

$$p = 1/2 \quad \text{vs} \quad p > 1/2.$$

- Second example: A random sample of 47 newborn's weight was recorded. Our aim is to estimate the population mean weight of newborns, μ. We may be interested in deciding between the following two statements

$$\mu = 3.3 \quad \text{vs} \quad \mu \neq 3.3.$$

The Null and Alternative Hypotheses

- The first hypothesis is called the **null hypothesis** H_0 and the other is called the **alternative hypothesis** H_A or H_1. Both are the statements about **population parameter**.
- H_0 usually is a statement with equality, or no effect.
- In the Example 1 in previous slide, H_0 is $p = 1/2$, and H_1 is $p > 1/2$, this is called "one sided test" and the side is right side.
- For our aim, H_1 can be the complement of H_0 (two sided) or one sided left, or one side right. However, in software, by default, H_1 is the complement of H_0, meaning if you let R know that your H_0 is $p = 1/2$ then by default the result of the test that R gives will be for two sided test with $H_1 : p \neq 1/2$.
- The decision rule has typically **two possible conclusions**: either **reject** H_0 or **do not reject** H_0.
- Note that, when H_0 is not rejected, many people usually say "accept" H_0 instead of "do not reject H_0", but in fact this way of saying ("accept" H_0) is not exactly.

Terminology

- The decision rule (of reject or do not reject H_0) is based on a **test statistic**.
- The set of values of a test statistic that leads to rejection of H_0 is called the **rejection region** or **critical region**.
- The set of values that leads to "“acceptance" H_0 is called the **acceptance region**.
- The probability distribution of the test statistic when H_0 is true is call the **null distribution**.

General Steps of a Significance Test

There are 5 steps in general for a significance test:

1. Assumptions: state the assumptions that the data should satisfy in order for the result of the test to be reliable.
2. Hypotheses: state the hypotheses of the test.
3. Test statistic: measure how far the estimate value falls from the value given under H_0.
4. Calculate p-value: to quantify how far the estimate value falls from the value given under H_0.
5. Conclusion: based on p-value to make conclusion.

Roughly, we'll assume that H_0 is true in the population. If our data (represent the population) resemble H_0 then we do not reject H_0, however if our data do not resemble H_0 then we tend to reject H_0.

Parametric Tests vs Nonparametric Tests

- **Parametric tests** are the significance tests that assume some form of distribution for the sample (or population) to follow. Example, a parametric test can be used when sample approximately follow normal distribution (which could be t-test).

- Conversely, **nonparametric tests** are the significance tests that do not assume any form of distribution for the sample. For example, when sample(s) is/are categorical (may be ordinal), then using parametric tests is not appropriate; or for some situation when sample is quantitative however the sample size is too small to appropriately approximate a distribution for it.

1 Introduction

2 Tests for One-Sample Data
- Parametric Tests
- Nonparametric Tests

3 Tests for Two-Sample Data
- Tests for Two Independent Samples
- Tests for Two Dependent/Matched Samples

One-Sample Data

- There are many situations where our data contain only one sample.
- Example 1: a random sample of 500 voters in a study to predict if Worker Party is winning at Aljunied GRC in the upcoming General Election; In this sample, we want to see if the population proportion (voting for WP) is larger than 0.5 or not.
- Example 2: a random sample of weight of 47 newborns. We'll perform tests based on this sample.
- Example 3: a random sample of weight of 50 male students is given. Another random sample of weight of 50 female students from the same population is given. We'll perform **tests based on these two samples** to draw some conclusion. **This is NOT the situation of one-sample data**.

Significance Tests About Mean

- For quantitative variable, significance tests often refer to population mean μ.
- We shall consider an example where we had a random sample of 47 babies' weights born from smoking mothers.
- Suppose that the average baby born in population is 3.3 kg. We'll perform a test to see if the mean weight of newborn from smoking mothers μ is lighter than mean weight of newborn in population (3.3 kg).
- What we are about to learn is known as the **one sample t-test**.

t-Test (1)

- Step 1. Assumption: sample must come from **randomization**; sample follows or **approximately follows normal distribution**. If the distribution differs from normal then the sample size should be large enough (at least 30); For the baby weight data `babyweights.csv`, you can plot the histogram, the summary to check the normality, which we can approximate the distribution of sample to be normal.

```
> data = read.csv("C:/Data/babyweights.csv", header = TRUE)
> attach(data)
> summary(weight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.041   2.862   3.202   3.208   3.572   4.141
> var(weight)
[1] 0.2564027
```

- Step 2. Hypotheses: population mean (mean weight of babies born from smoking mothers) is denoted as μ.

$$\mu = 3.3 \quad \text{vs} \quad \mu \neq 3.3.$$

t-Test (2)

- Step 3. Test statistic:

$$T = \frac{\bar{X} - \mu_0}{\text{s.e}(\bar{X})} = \frac{3.208 - 3.3}{\sqrt{0.2564/47}} = -1.2456,$$

  where μ_0 is the value of parameter μ under H_0.
- Under H_0, test statistic T follows a t distribution with df $n - 1 = 46$, hence **null distribution** is t_{46}.

- Step 4. Hence, p-value for the (2 sided test) is the two tail areas, which is 0.218.

- Step 5. (Conclusion) The p-value is not small, that means data do not provide strong evidence (or provide very weak evidence) against H_0.

t-Test in R

```
> t.test(weight, mu = 3.3)

        One Sample t-test

data:  weight
t = -1.2488, df = 46, p-value = 0.2181
alternative hypothesis: true mean is not equal to 3.3
95 percent confidence interval:
 3.059091 3.356438
sample estimates:
mean of x
 3.207764
```

t-Test in Python

```
In [28]: import pandas as pd
         import numpy as np
         import statistics as st
         from statistics import mean
         from statistics import median
         from statistics import variance
         from scipy import stats
         import math
         data = pd.read_csv (r"C:\Data\babyweights.csv")
         weight = data['weight']

         stats.ttest_1samp(weight, popmean=3.3)
```

```
Out[28]: Ttest_1sampResult(statistic=-1.2487860392703891, pvalue=0.2180614156412892)
```

t-Test in SAS (1)

Importing data; find few summaries of variable weight; and then perform a t-test:

```
* IMPORTING DATA FROM A TEXT/CSV FILE:;
FILENAME REFFILE '/home/u59061977/babyweights.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=DLM
    OUT=WORK.baby;
    DELIMITER=",";
    GETNAMES=YES;
    DATAROW=2;
RUN;

/* To find few summaries (mean, sd...) of variable weight:  */
proc means data=baby;
var weight;
run;

/* Test H_0: mu = 3.3 against H_1: mu != 3.3;  */
/* the ouput of code below will include sign test and signed rank test also   */

proc univariate data=baby mu0=3.3;
var weight;
run;
```

More detail on the code can be found here:
`https://v8doc.sas.com/sashtml/proc/z0146803.htm`

t-Test in SAS (2)

The output

The UNIVARIATE Procedure
Variable: weight

Moments			
N	47	Sum Weights	47
Mean	3.20776402	Sum Observations	150.764909
Std Deviation	0.50636218	Variance	0.25640265
Skewness	-0.2897971	Kurtosis	-0.3490705
Uncorrected SS	495.412773	Corrected SS	11.7945221
Coeff Variation	15.7855183	Std Error Mean	0.07386051

Basic Statistical Measures			
Location		Variability	
Mean	3.207764	Std Deviation	0.50636
Median	3.202360	Variance	0.25640
Mode	3.120713	Range	2.10013
		Interquartile Range	0.73935

Tests for Location: Mu0=3.3				
Test	Statistic		p Value	
Student's t	t	-1.24879	Pr > |t|	0.2181
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	-89	Pr >= |S|	0.3517

Nonparametric Tests

- In order to use t-test, we need some assumptions including the assumption that the sample is approximately normally distributed (or if not normally distributed then sample size should be large).
- However, we might have many situations where the sample does not follow normal distribution nor having large size.
- That means, the sample is skewed. In these cases, the suitable test is about median rather than the mean.
- The suitable nonparametric tests are: sign test and Wilcoxon signed rank test where the WSR test is much stronger.

Rough Idea of Sign Test

- A sample of size n is randomly collected from population: $x_1, x_2, ..., x_n$ that is skewed. We want to test

$$H_0 : \text{population median} = m_0 \quad \text{vs} \quad H_1 : \text{population median} \neq m_0$$

- We'll assign the sign ($+$ or -) to each data point: if $x_i > m_0$ then x_i has $+$ sign; if $x_i < m_0$ then x_i has - sign; if $x_i = m_0$ then no sign is given and the sample size will be reduced.
- The total number of positive sign is counted, $V+$; and the total number of negative sign is $V-$.
- We then take the test statistic $V = \min(V+, V-)$.
- This test statistic follow Binomial distribution $Bin(n^*, 0.5)$. $n^* = n$ - number of data points that are equal to m_0.
- p-value is then calculated (2 tails probability for the 2 sided test).

- Hypotheses

$$H_0 : \text{population median} = m_0 \quad \text{vs} \quad H_1 : \text{population median} \neq m_0$$

x_1	$+$
x_2	$+$
x_3	-
$\vdots$	$\vdots$
x_n	$+$

- If $x_i \neq m_0, \forall i$ then the test statistic: $V = \min(V+, V-) \sim Bin(n, 0.5)$

The Idea of Wilcoxon Signed Rank Test

- Hypotheses:

$$H_0 : \text{population median} = m_0 \quad \text{vs} \quad H_1 : \text{population median} \neq m_0$$

- The difference of data point and m_0 is calculated $x_i - m_0$ and give the rank and sign accordingly, they are $d_i(+)$ or $d_i(-)$.

- $V+$ is the sum of all the positive ranks. $V-$ is the sum of all the negative rank.

- The idea behind the test is that if $V+ \approx V-$ then we have evidence supporting H_0.

- If $V+$ is much greater (lesser) than $V-$ in absolute value, then we have evidence that the median is greater (lesser) than the hypothesised value.

- This test is much stronger than the sign test, hence only the WSR test is presented in detail.

Wilcoxon Signed Rank Test in R

$$H_0 : \text{population median} = 3.3 \quad \text{vs} \quad H_1 : \text{population median} < 3.3$$

```
> weight.non.0 = (weight[weight!=3.3])
> wilcox.test(weight.non.0, mu=3.3, alternative="less")

        Wilcoxon signed rank test with continuity correction

data:  weight.non.0
V = 475, p-value = 0.1745
alternative hypothesis: true location is less than 3.3
```

- The p-value is 0.1745, which is quite close to the parametric t-test (one-sided left test with p-value of 0.109).

Wilcoxon Signed Rank Test in Python

```
In [39]: scipy.stats.wilcoxon(x = weight-m0, y=None, zero_method='wilcox', correction=False, alternative='less')
Out[39]: WilcoxonResult(statistic=475.0, pvalue=0.1731072831261078)
```

- The p-value for the one sided test is 0.173, similar as the Wilcoxon signed rank test conducted in R.

Wilcoxon Signed Rank Test in SAS

- The signed rank test in SAS is produced together with t-test, however by default the output is for two sided alternative.

- For the baby weight data, the signed rank test for a two sided alternative has p-value 0.3517.

Independent and Dependent Samples

- Most analyses will require **independent samples** for the groups.
- The observations in one sample give us no clue about the values in the other sample.
- When we have two groups comprises the same subjects, then we have **dependent samples**.
- Independent samples can arise in a few ways:
  - In an experimental study, study units are assigned randomly to different treatments.
  - In an observational study, we draw a random sample from the population, and then observe an explanatory variable, like the weight of a person
  - In an observational study, we draw a random sample from a group (say smokers), and then a random sample from another group (say non-smokers), we will also have a random sample
- Dependent samples **usually** arise in the following way: If we were to measure weight lost of a group of 50 individuals before and after a weight loss program, then we would have two dependent groups.
  - The weight measurements *before* the program.
  - The weight measurements *after* the program.

Protein Level and Weight Gain

Example (Level of Protein in Diet and Weight Gain)

- In a randomized study, 30 rats were randomly assigned to a high protein diet, and 30 rats were randomly assigned to a low protein diet.
- The response variable, was the amount of weight gained in grams.
- We are interested in assessing if the level of protein in the diet is associated with the amount of weight gained.

For comparing two independent samples, we could have parametric or non parametric test.

5 Steps of a t-Test (Parametric) for Comparing Two Means

Step 1: Assumptions Here are the assumptions required for the PARAMETRIC test comparing population means:

- Independent samples, either from random sampling or randomized experiment.
- The population variance of each group is the same. We shall check this by a test (so called equal-variance test). If the equal-variance test suggests that the two variances are not equal then we have to resort to the *unequal variance* comparing means test.
- The population distribution of each group is **approximately normal**. This assumption is most crucial when n is small.

Step 2: Hypothesis

- The null hypothesis of a test for comparing two means has the following form:

$$H_0 : \mu_1 = \mu_2$$

  where μ_1 is the population mean of group 1, and μ_2 is the population mean of group 2.
- A two-sided alternative hypothesis would be

$$H_1 : \mu_1 \neq \mu_2$$

- One-sided alternatives are also possible,

$$H_1 : \mu_1 < \mu_2 \text{ or } H_1 : \mu_1 > \mu_2$$

Step 3: Test Statistic (Point Estimates)

- Let us represent the sample from group 1 as $X_1, X_2, \ldots, X_{n_1}$ and the sample from group 2 as $Y_1, Y_2, \ldots, Y_{n_2}$.
- We shall denote the sample mean from group 1 as $\bar{X}$ and the sample mean from group 2 as $\bar{Y}$. Note that n_1 and n_2 need not be equal.
- The point estimate of the difference between the population means is

$$\bar{X} - \bar{Y}$$

- Let us denote the sample variance from group 1 as s_1^2 and the sample variance from group 2 as s_2^2.
- Recall that this test, we have assumed that the population variances of the two groups are equal. If we denote this common value by σ^2, then we can use the data from both samples to estimate it. We shall call this the pooled estimate of the common variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Step 3: Test Statistic (Formulas)

- The test statistic is the distance between the point estimate and the null hypothesis value of the difference between population means, which is 0.
- This distance is measured in terms of standard error:

$$T = \frac{(\bar{X} - \bar{Y}) - 0}{se}$$

where the standard error is computed as

$$se = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- If H_0 is true, then T follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Steps 4 and 5: p-Value and Conclusion

- For a two-sided test, the p-value is the two-tail probability of a t distribution for values more extreme than the observed T.
- Smaller p-values give stronger evidence against H_0 in support of H_1.
- Interpret the p-value in the context of the experiment.
- If a decision is needed, reject H_0 if the p-value is less than a pre-decided significance level, e.g. 0.05.

Confidence Intervals for the Difference

Usually, the software will also provide us with a confidence interval for the mean difference. This is computed as

$$(\bar{X} - \bar{Y}) \pm t_{n_1+n_2-2,1-\alpha/2} \times se$$

- If 0 falls within the interval, it means that 0 is a plausible value for $\mu_1 - \mu_2$. This means we cannot rule out the possibility that $\mu_1 = \mu_2$.
- If both numbers of the interval are positive values, the confidence interval suggests that $\mu_1 - \mu_2$ is positive. We would infer that μ_1 is larger than μ_2.
- If both numbers of the interval are negative values, the confidence interval suggests that $\mu_1 - \mu_2$ is negative. We would infer that μ_1 is smaller than μ_2.

t-Test for Comparing Two Means in R (1)

- Take the data `protein_and_weight_gain.csv` as an example.
- We would compare the mean of weight gain for the 2 levels of protein intake (high and low).

```
> data = read.csv("C:/Data/protein_and_weight_gain.csv", header = TR
> attach(data)
> x = weight_gain[which(level == "high")]
> y = weight_gain[which(level == "low")]
> var.test(x,y) # # equal-variance test

        F test to compare two variances

data:  x and y
F = 0.97878, num df = 29, denom df = 29, p-value = 0.9543
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4658651 2.0564142
sample estimates:
ratio of variances
         0.9787807
```

t-Test for Comparing Two Means in R (2)

- We can test to see if the variances from the two samples are the same, this test has the null hypothesis that the two variances are the same.
- From the output, this equal-variance test has large p-value, suggests that the two samples have the similar variance.

```
> t.test(x,y, mu = 0, var.equal = TRUE) # if variances are equal

        Two Sample t-test

data:  x and y
t = 3.7553, df = 58, p-value = 0.0004033
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.78646 22.28021
sample estimates:
mean of x mean of y
 95.13333  80.60000

> #t.test(x,y, mu = 0, var.equal = FALSE) # if variances are NOT equ
```

t-Test for Comparing Two Means in R (3)

- Equivalently, we can use the following command to perform the t-test to compare the 2 populations means:

```
> t.test(weight_gain~level, mu = 0,var.equal = TRUE)

        Two Sample t-test

data:  weight_gain by level
t = 3.7553, df = 58, p-value = 0.0004033
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.78646 22.28021
sample estimates:
mean in group high  mean in group low
          95.13333           80.60000

> #t.test(weight_gain~level, mu = 0, var.equal = FALSE)
> # if variances are NOT equal
```

t-Test for Comparing Two Means in Python

```
In [54]: data = pd.read_csv (r"C:/Data/protein_and_weight_gain.csv")
         data = pd.DataFrame(data)
```

```
In [60]: # extracting the weight gain for the "low" and "high":
         high_data = data[(data['level'] =="high")]
         x = high_data["weight_gain"]

         low_data = data[(data['level'] =="low")]
         y = low_data["weight_gain"]
```

```
In [61]: scipy.stats.bartlett(x,y)  #Bartlett test to test if variances are equal
Out[61]: BartlettResult(statistic=0.0032784285184831435, pvalue=0.9543400246981172)
```

```
In [62]: scipy.stats.ttest_ind(x, y, axis=0, equal_var=True) # 2 samples t test
Out[62]: Ttest_indResult(statistic=3.7552732458950753, pvalue=0.00040332283036763155)
```

t-Test for Comparing Two Means in SAS (Code)

```
*  ANOTHER DATASET: PROTEIN AND WEIGHT GAIN;
* IMPORTING DATA FROM A TEXT/CSV FILE:;
FILENAME REFFILE '/home/u59061977/protein_and_weight_gain.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=DLM
    OUT=WORK.weightgain;
    DELIMITER=",";
    GETNAMES=YES;
    DATAROW=2;
RUN;

/* t test to compare means of weight gain, classified by level   */
PROC TTEST data = weightgain; *sides = L or U;
 var weight_gain;
class level;
 run;
```

More detail about the code can be found here:
`https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_ttest_syntax01.htm&docsetVersion=15.1&locale=en`.

t-Test for Comparing Two Means in SAS (Output)

The TTEST Procedure

Variable: weight_gain

level	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
high		30	95.1333	14.9083	2.7219	56.0000	120.0
low		30	80.6000	15.0690	2.7512	49.0000	107.0
Diff (1-2)	Pooled		14.5333	14.9889	3.8701		
Diff (1-2)	Satterthwaite		14.5333		3.8701		

level	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
high		95.1333	89.5665	100.7	14.9083	11.8731	20.0415
low		80.6000	74.9731	86.2269	15.0690	12.0011	20.2575
Diff (1-2)	Pooled	14.5333	6.7865	22.2802	14.9889	12.6886	18.3157
Diff (1-2)	Satterthwaite	14.5333	6.7864	22.2802			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	58	3.76	0.0004
Satterthwaite	Unequal	57.993	3.76	0.0004

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	1.02	0.9543

The t-test for equal and unequal assumption are both performed.
The test for equal variance is perform (the last table of the output) where it is not against the hypothesis of equal variance with p-value of 0.954.

Nonparametric Tests for Two Independent Samples

- The assumptions of two-sample t-test do not meet (data are not normally distributed and the sample size is small); In some situations, the data even be categorical with order.

Example (Math Scores)

- Suppose you are a secondary school math teacher, and you believe that students will score better in their weekly quiz if they have had breakfast.
- This week, you compare the scores of four students who ate a healthy breakfast with four students who did not have breakfast.

Ate Breakfast	Skipped Breakfast
87	93
96	83
92	79
84	73

Example (Tanning Methods)

- A student - Allison investigated in comparing the quality of two tanning methods without exposure to the sun: "tanning lotion" (which uses a bronze tanning lotion applied twice over a two-day period), and "tanning-studio" (where the person is exposed to UV light in a studio).
- Five females were recruited to join the study. Three of them were randomly chosen to apply the tanning lotion where the other two used the tanning studio.
- Allison then ranked the 5 females in terms of the quality of their tans. The ranks went from 1 to 5 with 1 = most natural looking and 5 = least natural looking.

Lotion	Studio
1	3
2	4
5	

Ranking Data (Math Scores)

Value	Rank
73	1
79	2
83	3
84	4
87	5
92	6
93	7
96	8

- All the values are grouped together and ranked; the smallest value is assigned rank 1 and the largest is assigned rank 8.
- The group who ate breakfast (in red) appears to have higher ranks.
- If there were no difference between the groups, we would expect the 8 ranks to be spread equally between the two groups.
- Similar idea is applied to the Tanning Methods data, however this data already have its observations be the rank from 1 to 5.

Mann-Whitney U-Test (1)

- Mann-Whitney test is also can be called as **Wilcoxon Rank Sum test**.
- Let $X_1, ..., X_n$ be IID with cdf F,
  and $Y_1, ..., Y_m$ IID with cdf G.
- We consider the null hypothesis $H_0 : F = G$.
- We are interested in whether X values are on the whole larger than the Y values or vice–versa.

Mann-Whitney U-Test (2)

There are few different definition on how the test statistic is calculated. In R, the test statistic is computed by: number of pairs $(X[i], Y[j])$ for which $Y[j]$ is not greater than $X[i]$.
Another way is presented below (Mathematical Statistics, 3rd edt by John A. Rice).

- All the $(n+m)$ observations are grouped together: $Z_1, ..., Z_{n+m}$ as pooled sample, and we assume the values are distinct. We define:

$$\text{Rank}(Z) = i$$

  if Z is the ith smallest value within the pooled sample.

- Define the rank sum scores

$$R_X = \sum_{i=1}^{n} \text{Rank}(X_i), \quad R_Y = \sum_{i=1}^{m} \text{Rank}(Y_i)$$

- The idea is that under $H_0 : F = G$, the ranks are uniformly distributed from $\{1, ..., m+n\}$, so the rank sums should not be too small or large.

Mann-Whitney U-Test (3)

- Note that $R_X + R_Y = (m+n)(m+n+1)/2$ is fixed. So looking at one is equivalent to looking at the other.
- We reject if either R_X or R_Y is too small or large.
- We take the smaller sample, suppose of size $n = \min(n, m)$, and compute the sum ranks R from that sample.
- Let $R' = n(m+n+1) - R$. The Mann-Whitney test statistic is

$$W = \min(R, R').$$

- We reject H_0 if W is too small.

Mann-Whitney U-Test: Examples

Math Scores Example:

- Let X denote the score of student who ate breakfast. Then

$$R = R_X = 4 + 5 + 6 + 8 = 23, \quad R' = 4 \times (4 + 4 + 1) - 23 = 13$$

$$W = \min(R, R') = 13$$

- Using the file `Mann_W_Table_1.jpg` in the "Statistical Tables" folder on Luminus, the rejection region at $\alpha = 0.05$ for 2-sided test is $W \leq 10$ ($n_1 = n_2 = 4$). Hence, when $W = 13$ we do not reject H_0.

Tanning Methods Example:

- Let X denote the rank of 3 females using tanning lotion. Then

$$R = 3 + 4 = 7, \quad R' = 2 \times (2 + 3 + 1) - 7 = 5$$

$$W = \min(R, R') = 5$$

- There is no rejection region if the test is 2-sided test with $n_1 = 2; n_2 = 3$, performed at $\alpha = 0.05$. Hence, at $\alpha = 0.05$, we do not reject H_0.

Mann-Whitney U-Test With Ties

- Assume that in the Math Scores example, there are two students who shared the same score of 84 (student with the score 83 is actually also has score 84).
- These 2 students will share the same rank that is 3.5, calculated by $(3 + 4)/2 = 3.5$.

Mann-Whitney U-Test in R (1)

```
> bf = c(87,96,92,84) # with breakfast
> no.bf = c(93,83,79,73) #without breakfast
> wilcox.test(bf,no.bf)

        Wilcoxon rank sum exact test

data:  bf and no.bf
W = 13, p-value = 0.2
alternative hypothesis: true location shift is not equal to 0
> lotion = c(1,2,5)
> studio = c(3,4)
> wilcox.test(lotion, studio)

        Wilcoxon rank sum exact test

data:  lotion and studio
W = 2, p-value = 0.8
alternative hypothesis: true location shift is not equal to 0
```

Mann-Whitney U-Test in R (2)

We still can apply the Mann-Whitney U-Test for the 2 independent samples in the Protein level and weight gain data.

```
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 682.5, p-value = 0.0006001
alternative hypothesis: true location shift is not equal to 0
```

The p-value is quite close to the p-value obtained by t-test.

Mann-Whitney U-Test in Python

Mann-Whitney U-Test for the 2 independent samples in the Protein level and weight gain data:

```
In [6]: scipy.stats.mannwhitneyu(x, y, use_continuity=True, alternative='two-sided')
Out[6]: MannwhitneyuResult(statistic=682.5, pvalue=0.00060011971204658 19)
```

- The p-value (from Mann Whitney U test) in Python is the same as the p-value (from Wilcoxon rank sum test) in R.

Mann-Whitney U-Test in SAS

```
44 *Perform the Mann-Whitney U Test;
45 proc npar1way data=weightgain wilcoxon;
46 class level;
47 var weight_gain;
48 run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable weight_gain Classified by Variable level

level	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
high	30	1147.50	915.0	67.607729	38.250
low	30	682.50	915.0	67.607729	22.750

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	Z	Pr > Z	Pr > |Z|	t Approximation Pr > Z	t Approximation Pr > |Z|
1147.500	3.4316	0.0003	0.0006	0.0006	0.0011

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
11.8264	1	0.0006

More detail about the idea of SAS code above can be found here:
`https://support.sas.com/rnd/app/stat/procedures/npar1way.html`
or here:
`https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_npar1way_syntax01.htm&docsetVersion=15.1&locale=en`.

Dependent Samples

- There are many situations where each subject receives both treatments.
- Each subject could have been measured in the absence of drug and after receiving the drug.
- The response of a subject in the control and treatment groups would no longer be independent.
- 2-sample t-test cannot be used since the 2 groups are no longer independent.
- A paired t-test can be used if the differences between before and after treatments follow (or approximately follow) normal distribution.
- If the sample size is too small or it is not appropriate to assume a normal distribution for the differences between before and after, then nonparametric tests (Wilcoxon signed rank test) can be used.

Example: Reaction Time

Example

- Consider a study on a sample of 32 drivers.
- In a simulation of driving situations, a target flashed red or green at irregular periods. Drivers pressed a brake button as soon as they detected a red light, and their reaction time was measured.
- The study was repeated twice for each driver – at one of the repetitions, the individual carried on a phone conversation; at the other, they listened to the radio.
- It is of interest to determine if cell phone usage is associated with a slower reaction time.

Example: Smoking and Platelet Aggregation

Example

- Platelets are involved in the formation of blood clots and it is known that smokers suffer more often from disorders involving blood clots than do nonsmokers.
- A study on a sample of 11 individuals before and after they smoked a cigarette and measured the extent to which the blood platelets aggregated.
- Data gives the maximum percentage of all the platelets that aggregated after being exposed to a stimulus.

Before	25	25	27	44	30	67	53	53	52	60	28
After	27	29	37	56	46	82	57	80	61	59	43
Difference	2	4	10	12	16	15	4	27	9	-1	15

Paired t-Test (1)

- Denote the pairs as $(X_i, Y_i), i = 1, ..., n$.
- X's and Y's have the means μ_X and μ_Y and variance σ_X^2 and σ_Y^2.
- The difference is denoted as D with the random sample of $D_i = X_i - Y_i, i = 1, ..., n$.
- Different pairs are independently distributed. It follows that the differences $D_i = X_i - Y_i$ are independent and $E(D_i) = \mu_X - \mu_Y$.
- We can estimate $\mu_D = \mu_X - \mu_Y$ by $\bar{D} = \bar{X} - \bar{Y}$.
- If the sample of differences is normally distributed (or approximately) then we can use paired t-test to test

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D \neq 0$$

Paired t-Test (2)

- When the population variance of the differences is unknown, the test statistic is $t = \dfrac{\bar{D} - 0}{s_{\bar{D}}}$ which follows a t distribution with $df = n - 1$, where $s_{\bar{D}}$ is the standard error of $\bar{D}$ - the sample mean of differences.

- A $100(1-\alpha)\%$ CI for μ_D is $\bar{D} \pm t_{n-1}(\alpha/2)s_{\bar{D}}$.

- A two-sided level α test of $H_0 : \mu_D = 0$ (i.e., no treatment effect) has the rejection region $|\bar{D}| > t_{n-1}(\alpha/2)s_{\bar{D}}$.

Paired t-Test for Smoking and Platelet Aggregation Data

- $\bar{D} = 10.27$, and $s_{\bar{D}} = 2.4 = \text{Var(Difference)}/11$.

- 90% CI:

$$\bar{D} \pm t_{10}(0.05) s_{\bar{D}} = (5.9, 14.6).$$

- Two–sided level 0.01 hypothesis test:

$$t = \frac{10.27}{2.4} = 4.28 > 3.169 = t_{10}(0.005) \quad \Rightarrow p - \text{value} < 0.01.$$

Paired t-test in R

```
> before = c(25, 25,27, 44,30,67, 53, 53,52, 60, 28)
> after = c(27, 29, 37, 56, 46, 82, 57, 80,61,59,43)
> t.test(after, before, mu = 0, paired = TRUE, conf.level = 0.9)

        Paired t-test

data:  after and before
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
  5.913967 14.631488
sample estimates:
mean of the differences
                10.27273
```

Paired t-test in Python

```
In [12]: scipy.stats.ttest_rel(after,before, axis=0, nan_policy='propagate')
```

```
Out[12]: Ttest_relResult(statistic=4.271608818429545, pvalue=0.0016328499219996722)
```

scipy.stats.ttest_rel

scipy.stats.**ttest_rel**(*a, b, axis=0, nan_policy='propagate'*) [source]

Calculate the t-test on TWO RELATED samples of scores, a and b.

This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values.

Parameters: **a, b** : *array_like*

The arrays must have the same shape.

axis : *int or None, optional*

Axis along which to compute test. If None, compute over the whole arrays, *a*, and *b*.

nan_policy : *{'propagate', 'raise', 'omit'}, optional*

Defines how to handle when input contains nan. The following options are available (default is 'propagate'):

- 'propagate': returns nan
- 'raise': throws an error
- 'omit': performs the calculations ignoring nan values

Paired t-test in SAS

```
101 * CREATE DATA;
102 data platelet;
103 input before after;
104 datalines;
105 25 27
106 25 29
107 27 37
108 44 56
109 30 46
110 67 82
111 53 57
112 53 80
113 52 61
114 60 59
115 28 43
116 ;
117 PROC TTEST DATA=platelet;
118     PAIRED after*before;
119 RUN;
```

The TTEST Procedure

Difference: after - before

N	Mean	Std Dev	Std Err	Minimum	Maximum
11	10.2727	7.9761	2.4049	-1.0000	27.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
10.2727	4.9143	15.6311	7.9761	5.5730	13.9975

DF	t Value	Pr > |t|
10	4.27	0.0016

Nonparametric Tests for Paired Samples

Example

Consider an experiment that each subject tries each of the two drugs. The time span to pain relief is measured:

Subject	1	2	3	4	5	6	7	8
Drug A	20	40	30	45	19	27	32	26
Drug B	18	36	32	46	15	22	29	25
Difference (A - B)	2	4	-2	-1	4	5	3	1

- The Wilcoxon signed rank test can be used for the sample of differences (last row). The idea of how the tests are conducted is the same as for the case of one-sample data.

Paired Samples: Wilcoxon Signed Rank Test

```
> #####   Wilcoxon Signed Rank Test:
> drugA = c(20, 40, 30, 45, 19, 27, 32, 26)
> drugB = c(18, 36, 32, 46, 15, 22, 29, 25)
> diff = drugA - drugB
> wilcox.test(diff)

        Wilcoxon signed rank test with continuity correction

data:  diff
V = 31, p-value = 0.07895
alternative hypothesis: true location is not equal to 0
```

Paired Samples: nonparametric test in Python

```
drugA = np.array([20, 40, 30, 45, 19, 27, 32, 26])
drugB = np.array([18, 36, 32, 46, 15, 22, 29, 25])
diff = drugA - drugB
print(diff)
```

```
array([ 2,  4, -2, -1,  4,  5,  3,  1])
```

```
#Wilcoxon Signed Rank test for diff:
scipy.stats.wilcoxon(x = diff, zero_method='wilcox', correction=True, alternative='two-sided')
```

```
WilcoxonResult(statistic=5.0, pvalue=0.07894833771600107)
```

Paired Samples: nonparametric test in SAS

```
134 * CREATE DATA DRUG;
135 data drug;
136 input DrugA DrugB;
137 datalines;
138 20 18
139 40 36
140 30 32
141 45 46
142 19 15
143 27 22
144 32 29
145 26 25
146 ;
147 data drug;
148 set drug; /* the variable in dataset drug will be used  */
149 diff = DrugA - DrugB; /* diff is created by DrugA and DrugB */
150 run;
151
152 proc univariate data=drug mu0=0;
153 var diff;
154 run;
```

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.256304	Pr > |t|	0.0587
Sign	M	2	Pr >= |M|	0.2891
Signed Rank	S	13	Pr >= |S|	0.0859