ECON20003 QM2

Tutorial 4 Semester 1, 2022

Chin Quek

Department of Economics

March 2022

Objectives

- Comparing the Central Locations of Two Populations
  - Matched-pairs Z/t test
  - One-sample Sign test
  - One-sample Wilcoxon signed rank test
  - Two-independent-sample Z/t test (3 types)
  - Wilcoxon rank-sum test

Reminder: Always create an R Project and R script when starting on a new exercise.

Inference for comparison of two populations

We have to establish whether the data gathered for the two populations is:

- matched pair (paired-sample design), or come from
- independent samples (independent measures design).

What is the difference between matched pair and independent samples?

Inference for comparison of two populations

We have to establish whether the data gathered for the two populations is:

- matched pair (paired-sample design), or come from
- independent samples (independent measures design).

What is the difference between matched pair and independent samples?

- The SSK textbook defines independent samples as "samples that are **completely** unrelated to one another".

Inference for comparison of two populations

We have to establish whether the data gathered for the two populations is:

- matched pair (paired-sample design), or come from
- independent samples (independent measures design).

What is the difference between matched pair and independent samples?

- The SSK textbook defines independent samples as "samples that are **completely** unrelated to one another".

Matched pair data is defined as recognising a "natural pairing between one observation in the first sample and one observation in the second sample. If a natural pairing exists then the experiment involves matched pairs".

Inference for comparison of two populations

We have to establish whether the data gathered for the two populations is:

- matched pair (paired-sample design), or come from
- independent samples (independent measures design).

What is the difference between matched pair and independent samples?

- The SSK textbook defines independent samples as "samples that are **completely** unrelated to one another".

Matched pair data is defined as recognising a "natural pairing between one observation in the first sample and one observation in the second sample. If a natural pairing exists then the experiment involves matched pairs".

- In the matched pair case, we concentrate on the difference, D, between respective (matched) observations.
- With independent sample, we are looking at the difference between the population means, $\mu_1 - \mu_2$.

Review Question 1

A survey is conducted to be determine customer satisfaction in a Macca's store in George St., Sydney and another survey contains data (taken on the same day at the same time) of customer satisfaction for a Macca's store in Collins St., Melbourne.

Answers to questions relating to speed of service, food quality etc. were based on a 5 point Likert scale:

- 1 - Strongly Agree,
- 2 - Agree,
- 3 - Neither Agree nor Disagree,
- 4 - Disagree,
- 5 - Strongly Disagree

Is the data matched-pair or independent samples? What type of data is this?

Review Question 1

A survey is conducted to be determine customer satisfaction in a Macca's store in George St., Sydney and another survey contains data (taken on the same day at the same time) of customer satisfaction for a Macca's store in Collins St., Melbourne.

Answers to questions relating to speed of service, food quality etc. were based on a 5 point Likert scale:

- 1 - Strongly Agree,
- 2 - Agree,
- 3 - Neither Agree nor Disagree,
- 4 - Disagree,
- 5 - Strongly Disagree

Is the data matched-pair or independent samples? What type of data is this?

Since Macca's products are uniform across all stores in different geographic locations, this data can be thought of as being matched pair data.

IF however, we were conducting the same type of survey for an Italian restaurant in George Street, Sydney and another Italian restaurant in Collin Street, Melbourne, this data would be *independent*.

Rating data are of the ordinal scale, and qualitative.

More Review Questions

A lecturer wants to compare students' scores on the mid-semester and final exam. This is most often done by obtaining a sample of students and recording each student's mid-semester exam score and final exam score.

What is the experimental design?

A shoe company is studying how many shoes Italian men and women own. In one research study they take a random sample of 500 Italian adults and ask each individual if they identify as a man or women and how many pairs of shoes they own.

What is the experimental design?

In a second study the researchers use a different design. This time they take a random sample of 250 married couples in Italy. They record the number of shoes owned by each husband and each wife.

What is the experimental design?

Suppose that sample 1 size is 20, mean is 100 and $s_1^2 = 225$, sample 2 size is 15, mean is 125 and $s_2^2 = 150$. We are interested to conduct a hypothesis test of the differences in the means, $\mu_1 - \mu_2$. Assume equal population variances. What is the degrees of freedom for the relevant statistical test?

A more robust parametric alternative to the independent samples t-test is the:

a. Paired-sample t test.
b. Sign test.
c. Welch's t test.
d. Wilcoxon rank sum test.
e. Wilcoxon signed rank test.

A more robust parametric alternative to the independent samples t-test is the:

a. Paired-sample t test.
b. Sign test.
c. Welch's t test.
d. Wilcoxon rank sum test.
e. Wilcoxon signed rank test.

Recall in the tutorial handout that by default, R assumes that the population variances are different and performs an unequal variances t-test originally developed by Welch.

And that if there is reason to believe that the population variances are equal, we should be performing the equal variances t-test.

Parametric techniques for comparison of two populations

Explain the different formulae below. In what situation would you use one over the other?

Formula 1:

$$T = \frac{\bar{x} - \mu_D}{s_D / \sqrt{n_D}} \sim t_{n_D - 1}$$

Formula 2:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \sim t_{df}$$

where $df = \frac{\left(s^2_{\bar{x}_1 - \bar{x}_2}\right)^2}{\left(s^2_{\bar{x}_1}\right)^2/(n_1-1) + \left(s^2_{\bar{x}_2}\right)^2/(n_2-1)}$, $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Formula 3:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \sim N(0, 1)$$

where $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Formula 4:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \sim t_{df}$$

where $df = n_1 + n_2 - 2$ and $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Matched-pairs Z/t tests

The **matched-pairs** Z / t tests and the corresponding confidence interval for the difference (D) between the before and after population means are based on the following assumptions:

1. The data is a random sample of pairs of observations (i.e. the before and after samples are not independent of each other).
2. The variable of interest is quantitative and continuous.
3. The measurement scale is interval or ratio.
4. Either (Z test) the population standard deviation of the differences, σ_D, is known and the sample mean of the differences is at least approximately normally distributed, or (t-test) σ_D is unknown but the population of the differences is normally distributed (at least approximately).

Two-independent-sample Z/t test

The **two-independent-sample** Z / t test and the corresponding confidence interval estimator for the difference between two population means are based on the following assumptions:

i. The data consists of two independent random samples of independent observations (i.e. both the samples and the observations within each sample are independent).
ii. The variable of interest is quantitative and continuous.
iii. The measurement scale is interval or ratio.
iv. Either (Z test) the population standard deviations, σ_1 and σ_2, are known and the sample means are at least approximately normally distributed, or (t-test) σ_1 and σ_2 are unknown but the sampled populations are normally distributed (at least approximately).

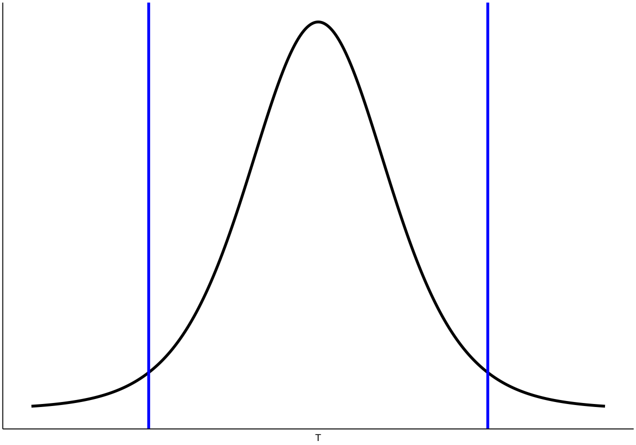
Application Question

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for 8 randomly selected subjects are shown in the table below. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated.

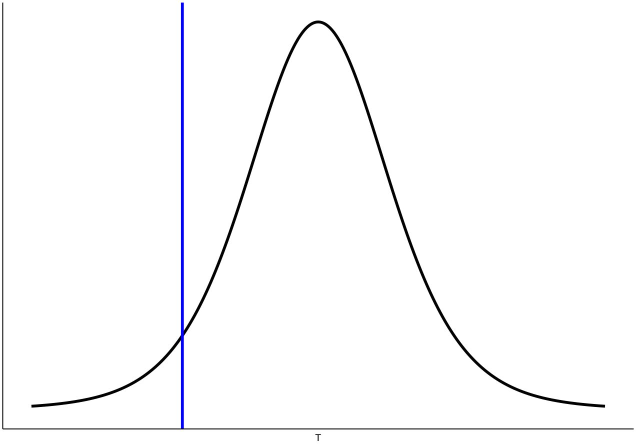
Subject	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Assume that the population of differences is **normally distributed**. Is there any statistical evidence that hypnotism is effective in reducing pain? Do we conduct a left-tail, right-tail or two-tail test?

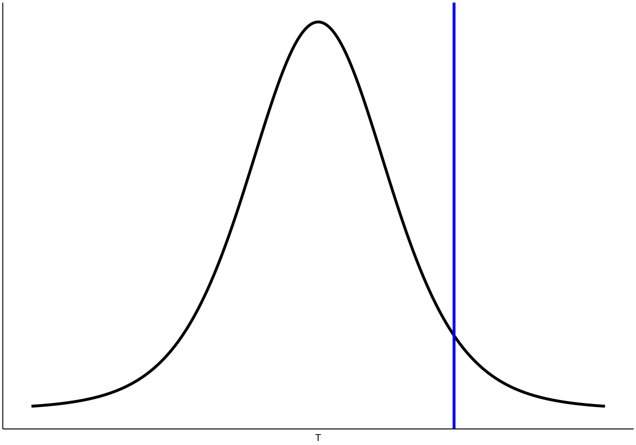
$$H_0 : \mu_{\text{before}} - \mu_{\text{after}} = 0 \qquad H_A : \mu_{\text{before}} - \mu_{\text{after}} \neq 0$$



$$H_0 : \mu_{\text{before}} - \mu_{\text{after}} = 0 \qquad H_A : \mu_{\text{before}} - \mu_{\text{after}} < 0$$



$$H_0 : \mu_{\text{before}} - \mu_{\text{after}} = 0 \qquad H_A : \mu_{\text{before}} - \mu_{\text{after}} > 0$$



Solving the application question

The null hypothesis is 0, i.e. the subject shows no improvement after hypnotism.

To investigate whether hypnotism is effective in reducing pain, we need to test if the sensory measurements are, on average, lower after hypnotism?

- In other words, if there is less pain felt after hypnotism, and the 'after' measurement should be lower than the 'before' measurement

Solving the application question

The null hypothesis is 0, i.e. the subject shows no improvement after hypnotism.

To investigate whether hypnotism is effective in reducing pain, we need to test if the sensory measurements are, on average, lower after hypnotism?

- In other words, if there is less pain felt after hypnotism, and the 'after' measurement should be lower than the 'before' measurement

So, if we define $\mu_D = \mu_{\text{after}} - \mu_{\text{before}}$, our hypotheses would be:

$$H_0 : \mu_D = 0 \qquad H_A : \mu_D < 0$$

Solving the application question

The null hypothesis is 0, i.e. the subject shows no improvement after hypnotism.

To investigate whether hypnotism is effective in reducing pain, we need to test if the sensory measurements are, on average, lower after hypnotism?

- In other words, if there is less pain felt after hypnotism, and the ‘after’ measurement should be lower than the ‘before’ measurement

So, if we define $\mu_D = \mu_{\text{after}} - \mu_{\text{before}}$, our hypotheses would be:

$$H_0 : \mu_D = 0 \qquad H_A : \mu_D < 0$$

If however, we define $\mu_D = \mu_{\text{before}} - \mu_{\text{after}}$, then our hypotheses would be:

$$H_0 : \mu_D = 0 \qquad H_A : \mu_D > 0$$

Define $D = \text{before} - \text{after}$

$$H_0 : \mu_D = 0 \qquad H_A : \mu_D > 0$$

before	after	D
6.6	6.8	-0.2
6.5	2.4	4.1
9.0	7.4	1.6
10.3	8.5	1.8
11.3	8.1	3.2
8.1	6.1	2.0
6.3	3.4	2.9
11.6	2.0	9.6

R output for a right-tail test

```
t.test(before, after, paired = TRUE , alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 3.0359, df = 7, p-value = 0.009478
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.174823      Inf
## sample estimates:
## mean of the differences
##                   3.125
```

Alternative and valid R output

```r
t.test(after, before, paired = TRUE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  after and before
## t = -3.0359, df = 7, p-value = 0.009478
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.174823
## sample estimates:
## mean of the differences
##                  -3.125
```

Experimental design and normality assumption

Instead of selecting 8 subjects for investigating effectiveness of hypnotism, would you prefer to use an independent measures design in which you select 16 subjects and divide them into two groups of 8?

Experimental design and normality assumption

Instead of selecting 8 subjects for investigating effectiveness of hypnotism, would you prefer to use an independent measures design in which you select 16 subjects and divide them into two groups of 8?

Suppose the normality assumption is not known, is there an alternative statistical test that you can perform?

Non-Parametric techniques for paired sample

Conditions for the sign test:

i. The data is a random sample of independent pairs of observations (i.e. the before and after samples are not independent of each other but the selected pairs are).
ii. The variable of interest is qualitative or quantitative.
iii. The measurement scale is at least ordinal.

Conditions for the Wilcoxon Signed Rank Sum test:

i. The data is a random sample of pairs of observations (i.e. the before and after samples are not independent of each other but the selected pairs are).
ii. The variable of interest is quantitative and continuous.
iii. The measurement scale is interval or ratio.
iv. The distribution of the differences is **symmetric**.

Exercise 2

Marketing strategists would like to predict consumers' response to new products and their accompanying promotional schemes. Consequently, studies that examine the differences between buyers and non-buyers of a product are of interest. One classic study conducted by Shuchman and Riesz (*Journal of Marketing Research*, Feb. 1975) was aimed at characterizing the purchasers and non-purchasers of Crest toothpaste.

The researchers demonstrated that both the mean household size (number of persons) and mean household income were significantly larger for purchasers than for non-purchasers.

A similar study utilized independent random samples of size 20 on the *age* of the *householder* primarily responsible for buying toothpaste. Householders were categorized as non-purchaser or purchaser of a particular brand of toothpaste coded as N and P, respectively. The data are saved in the **t4e2** file.

```
t4e2 <- read_excel("t4e2.xlsx")
```

a) Obtain and interpret a 90% confidence interval for the difference between the mean ages of purchasers and non-purchasers.

   Which formula do we use to obtain the 90% confidence interval?

Recall from our initial discussion that we have to choose the correct confidence interval estimator based on the information given in the question and also the data , if we have it - which in this case we do.

The population standard deviations are not given in the question which leaves two choices.

Are the population variances equal?

Recall from our initial discussion that we have to choose the correct confidence interval estimator based on the information given in the question and also the data , if we have it - which in this case we do.

The population standard deviations are not given in the question which leaves two choices.

Are the population variances equal?

```
by(t4e2$Age, t4e2$Householder, mean)
```

```
## t4e2$Householder: N
## [1] 47.2
## ------------------------------------------------------------
## t4e2$Householder: P
## [1] 39.8
```

```
by(t4e2$Age, t4e2$Householder, sd)
```

```
## t4e2$Householder: N
## [1] 13.62119
## ------------------------------------------------------------
## t4e2$Householder: P
## [1] 10.03992
```

The two sample variances are $s_1^2 = (13.621)^2 = 185.532$ and $s_2^2 = (10.040)^2 = 100.802$.

Their ratio, $s_1^2/s_2^2 = 1.84$, seems to be too big to assume that the corresponding population variances are equal, so let's follow the third scenario.

Remember that we are interested in the ratio of sample variances NOT sample standard deviations.

Assume that the sampled populations are not extremely non-normal.

The estimates of the variances of the sample means are:

$$s_{\bar{x}_1}^2 = \frac{s_1^2}{n_1} = \frac{185.32}{20} = 9.277,\ s_{\bar{x}_2}^2 = \frac{s_2^2}{n_1} = \frac{100.802}{20} = 5.040$$

and the estimate of the standard error of the difference between the two sample means is:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2} = \sqrt{9.277 + 5.040} = 3.784$$

The degrees of freedom for the t distribution is:

$$df = \frac{(s_{\bar{x}_1 - \bar{x}_2}^2)^2}{(s_{\bar{x}_1}^2)^2/(n_1 - 1) + (s_{\bar{x}_2}^2)^2/(n_2 - 1)} = \frac{(3.784)^2}{(9.277)^2/(19) + (5.040)^2/(19)} = 34.9 \approx 35$$

and the t reliability factor from the t-table is $t_{df,\alpha/2} = t_{35,0.05} = 1.690$.

The 90% CI estimate of the difference between the mean ages of purchasers and non-purchasers is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2} \times s_{\bar{x}_1 - \bar{x}_2} = (47.2 - 39.8) \pm 1.690 \times 3.784 = (1.005, 13.795)$$

It means that with 90% confidence the difference between the mean ages of purchasers and non-purchasers is somewhere between 1.0 and 13.8 years.

The 90% CI estimate of the difference between the mean ages of purchasers and non-purchasers is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2} \times s_{\bar{x}_1 - \bar{x}_2} = (47.2 - 39.8) \pm 1.690 \times 3.784 = (1.005, 13.795)$$

It means that with 90% confidence the difference between the mean ages of purchasers and non-purchasers is somewhere between 1.0 and 13.8 years.

What assumptions are necessary for constructing the CI?

Assumptions for a two-independent samples CI estimator

1. Data consists of two independent random samples
2. The variable of interest is quantitative and continuous
3. The measurement scale is interval or ratio
4. The population standard deviations are unknown but the sampled populations are normally distributed (at least approximately).

Assumptions for a two-independent samples CI estimator

1. Data consists of two independent random samples
2. The variable of interest is quantitative and continuous
3. The measurement scale is interval or ratio
4. The population standard deviations are unknown but the sampled populations are normally distributed (at least approximately).

First assumption: it was explicitly mentioned that the study was based on "independent random samples". The variable of interest is the age of the householder primarily responsible for buying toothpaste. It is a quantitative and continuous variable.

- BUT the actual observations are discrete values. So is the first assumption satisfied?

Assumptions for a two-independent samples CI estimator

1. Data consists of two independent random samples
2. The variable of interest is quantitative and continuous
3. The measurement scale is interval or ratio
4. The population standard deviations are unknown but the sampled populations are normally distributed (at least approximately).

First assumption: it was explicitly mentioned that the study was based on "independent random samples". The variable of interest is the age of the householder primarily responsible for buying toothpaste. It is a quantitative and continuous variable.

- BUT the actual observations are discrete values. So is the first assumption satisfied?

The actual observations are rounded to the nearest year, so they are discrete values, but since there are large number of possible values, for the purpose of hypothesis testing we can still treat this variable as being **continuous**.

> *Theoretically, age is a continuous variable since it can assume any real number, say, between 0 and 150 years. However, in practice it is always rounded so the data itself is discrete. In practice, continuous variables are always measured with limited precision, so the actual measurements are discrete. Still, if there are a large number of possible values, we treat the data as being continuous.*

Non-parametric technique for independent samples

What if we intend to compare the central locations of two populations which are very non-normal, or do not have means because they are measured on an ordinal scale, or do have means but we prefer to use the medians to measure their central locations?

- In these cases, we should use some non-parametric test for the difference between the population medians.

Non-parametric technique for independent samples

What if we intend to compare the central locations of two populations which are very non-normal, or do not have means because they are measured on an ordinal scale, or do have means but we prefer to use the medians to measure their central locations?

- In these cases, we should use some non-parametric test for the difference between the population medians.

Wilcoxon rank-sum test ???

- The Wilcoxon rank-sum test is **NOT** the same as the Wilcoxon signed ranks test!

Wilcoxon rank-sum test assumptions

The Wilcoxon rank-sum test is based on the following assumptions:

i. The data consists of two independent random samples of independent observations (i.e. both the samples and the observations within each sample are independent).
ii. The variable of interest is continuous ...
iii. and the measurement scale is at least ordinal.
iv. The two sampled populations that differ at most with respect to their central locations measured by the medians (i.e. they are identical in shape and spread).

Performing the Wilcoxon rank-sum test

Let sample 1 be the smaller (not bigger) sample, so that $n_1 \leq n_2$. To perform the Wilcoxon rank-sum test, we need to rank all available observations in the combined (pooled) sample from the smallest to the largest averaging the ranks of tied observations and calculate the rank sums of the two samples (T_1 and T_2). The test statistic is $T = T_1$.

The exact small sample ($n_1 \leq 10, n_2 \leq 10$) lower and upper critical values, T_L and T_U, are in Table 8, Appendix B (p.1086) of the Selvanathan book.

Performing the Wilcoxon rank-sum test

Let sample 1 be the smaller (not bigger) sample, so that $n_1 \leq n_2$. To perform the Wilcoxon rank-sum test, we need to rank all available observations in the combined (pooled) sample from the smallest to the largest averaging the ranks of tied observations and calculate the rank sums of the two samples (T_1 and T_2). The test statistic is $T = T_1$.

The exact small sample ($n_1 \leq 10, n_2 \leq 10$) lower and upper critical values, T_L and T_U, are in Table 8, Appendix B (p.1086) of the Selvanathan book.

We can reject H_0 if:

1. right tail test: $T \geq T_{U,\alpha}$
2. left tail test: $T \leq T_{L,\alpha}$
3. two tail test: $T \geq T_{U,\alpha/2}$ or $T \leq T_{L,\alpha/2}$

Large sample size normal approximation of T

For larger sample sizes the sampling distribution of T has a normal approximation with parameters

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2}, \qquad \sigma_T^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Note

T has a normal approximation is DIFFERENT to the data being normally distributed.

The hypotheses are

$$H_0 : \eta_1 - \eta_2 = 0 \qquad H_A : \eta_1 - \eta_2 \quad (\neq, >, <) \quad 0$$

Example

In a genetic inheritance study, samples of individuals from several ethnic groups were taken. Blood samples were collected from each individual and several variables measured. We shall compare the groups labelled "Native American" and "Caucasian" with respect to the variable `dispersion`.

The data is as follows:

Groups									
Native American	0.45	0.50	0.61	0.63	0.75	0.85	0.93		
Caucasian	0.44	0.45	0.52	0.53	0.56	0.58	0.58	0.65	0.79

How can we perform a test of differences between the two groups (i.e. Native Americans and Caucasians)?

Calculating rank sum

Ranks																
NA			.45	.50						.61	.63		.75		.85	.93
Ca	.44	.45			.52	.53	.56	.58	.58			.65		.79		
——	——	——	——	——	——	——	——	——	——	——	——	——	——	——	——	——
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tie-adj	1	2.5	2.5	4	5	6	7	8.5	8.5	10	11	12	13	14	15	16

Calculating rank sum

Ranks																
NA			.45	.50						.61	.63		.75		.85	.93
Ca	.44	.45			.52	.53	.56	.58	.58			.65		.79		
___	___	___	___	___	___	___	___	___	___	___	___	___	___	___	___	___
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tie-adj	1	2.5	2.5	4	5	6	7	8.5	8.5	10	11	12	13	14	15	16

$$T_{NA} = 2.5 + 4 + 10 + 11 + 13 + 15 + 16 = 71.5$$
$$T_{Ca} = 1 + 2.5 + 5 + 6 + 7 + 8.5 + 8.5 + 12 + 14 = 64.5$$