Homework 2

Your Name: Naga Satya Silpa Annadevara

Student ID: A20517818

---

1. **(35 points) Manually solve the problem below (do not use R):**

Note, if you need either z value or t value, you can find them by using this tool:
http://www.mathcracker.com/z_critical_values.php
http://www.mathcracker.com/t_critical_values.php
https://www.socscistatistics.com/pvalues/

A bank branch located in a commercial district of a city has the business objective of improving the process for serving customers during the noon to 1 PM (lunch period). The waiting time (defined as the time the customer enters the line until he or she reaches the teller window) of a random sample of 15 customers is collected, and the results are organized and stored as below:

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20
4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79

a) Calculate the mean and standard deviation, and find q1, q3 from the values above. Is the distribution symmetric? Why? [10]

ANSWER:

- The sample size: n = 15
- Sample mean = $\sum_{i=0}^{n} xi/n$ (dividing the sum of all values in a data set by the number of values.)
  - Mean = (4.21 + 5.55 + 3.02 + 5.13 + 4.77 + 2.34 + 3.54 + 3.20 + 4.50 + 6.10 + 0.38 + 5.12 + 6.46 + 6.19 + 3.79) / 15 = 64.3/15 = 4.28
  - Mean ($\overline{x}$ = 4.28)

- Standard deviation (S) = square root of the variance. ($\sqrt{S^2}$. Where is S^2variance.)
- So, to calculate the Standard deviation, first we need to calculate the variance of the sample (S^2)
  - Variance = $S^2 = \sum_{i=1}^{n} \frac{(xi-\overline{x})\text{^}2}{n-1}$
  - First, we need to calculate the mean (mean is already calculated above i.e., 4.28)
  1. Subtract the mean from each data point and square the result:
     - (4.21 - 4.28) ^2 = 0.0049
     - (5.55 - 4.28) ^2 = 1.6129

- (3.02 - 4.28) ^2 = 1.5876
- (5.13 - 4.28) ^2 = 0.7225
- (4.77 - 4.28) ^2 = 0.2401
- (2.34 - 4.28) ^2 = 3.7636
- (3.54 - 4.28) ^2 = 0.5476
- (3.20 - 4.28) ^2 = 1.1664
- (4.50 - 4.28) ^2 = 0.0484
- (6.10 - 4.28) ^2 = 3.3124
- (0.38 - 4.28) ^2 = 15.21
- (5.12 - 4.28) ^2 = 0.7056
- (6.46 - 4.28) ^2 = 4.7524
- (6.19 - 4.28) ^2 = 3.6481
- (3.79 - 4.28) ^2 = 0.2401

2. Calculate the average of the squared differences:

(0.0049+1.6129+1.5876+0.7225+0.2401+3.7636+0.5476+1.1664+0.0484+3.3124+15.21+0.7056+4.7524+3.6481+0.2401) ÷ 15-1

= 37.5626÷ 14 = 2.68

Therefore, The variance of the given sample data set: S^2 = 2.68

- The standard deviation = square root of the variance ($\sqrt{S^2}$)

$\sqrt{(2.68)}$ = 1.63

- To calculate Q1 , Q2 , Q3 , First we need to re-arrange the sample dataset in Ascending order:
  - 0.38, 2.34, 3.02, 3.20, 3.54, 3.79, 4.21, 4.50, 4.77, 5.12, 5.13, 5.55, 6.10, 6.19,6.46
    - Q1: (Lower quartile) = 3.20
    - Q2: (Median/Middle value) = 4.50
    - Q3: (Upper quartile) = 5.55

- Is the distribution symmetric?
  - Based on the central limit theorem, If the sample size(n) is less than 30, then it will not follow normal distribution.
  - Here the sample size (n) = 15 which is less than 30, The distribution is not symmetric. It follows t-distribution.
  - To tell whether it is symmetric or not, we can say if the mean and median are equal, then its symmetric. Here the mean is 4.28 & the median is 4.50.
  - Mean and median are not equal, hence we can say the distribution is not symmetric and it is skewed.
  - To say its skewness in detail, we can use a formula:

Skewness = (3 * (Mean - Median)) / Standard Deviation

= (3 * (4.28 – 4.50)) / 1.63

= (3 * –0.22) / 1.63

= - 0.66 / 1.63
= -0.40

Hence, we can say it is slightly left-skewed/ negative-skewed.

b) As a customer walks into the branch office during the lunch period. She asks the branch manager how long she can expect to wait. If you are the manager, answer this question by using 95% confidence. [10]

ANSWER:

To answer this question, we need to follow the below steps one by one:

1. Calculate sample statistics such as 'sample mean' and 'sample standard deviation'.

Sample mean: already calculated in the above question.

$\bar{y}$ = 4.28

Sample standard deviation: already calculated in the above question.

S = 1.63

2. As sample size (n = 15) is less than 30, it follows t- distribution, hence we need to use t-value in the margin of error.

3. Produce confidence interval [a,b] by using the formula :

a = sample estimate – margin of error (sample estimate i.e, sample mean)

$$\bar{y} - t\ \frac{\alpha}{2}\left(\frac{S}{\sqrt{n}}\right)$$

b = sample estimate + margin of error

$$\bar{y} + t\ \frac{\alpha}{2}\left(\frac{S}{\sqrt{n}}\right)$$

- To get the t- value we need to have 2 inputs:

1. alpha (α) = 1 – confidence level
= 1- 95%

(α) = 0.05

2.degree of freedom (df) = n – 1
= 15 – 1
= 14

- http://www.mathcracker.com/t_critical_values.php

from this given tool, the t- value is calculated from the t- table and observed as $t\ \frac{\alpha}{2} = 1.761$

Now,

- Let's calculate a = $\bar{y} - t\ \frac{\alpha}{2}\left(\frac{S}{\sqrt{n}}\right)$

$= 4.28 - 1.761\left(\frac{1.63}{\sqrt{15}}\right)$

$= 4.28 - 1.761\left(\frac{1.63}{3.87}\right)$

$= 4.28 - 1.761\ (0.421)$

$= 4.28 - 0.741$ (0.741 is the margin of error)

$= 3.53$

- Let's calculate b = $\bar{y} + t\ \frac{\alpha}{2}\left(\frac{S}{\sqrt{n}}\right)$

$= 4.28 + 1.761\left(\frac{1.63}{\sqrt{15}}\right)$

$= 4.28 + 1.761\left(\frac{1.63}{3.87}\right)$

$= 4.28 + 1.761\ (0.421)$

$= 4.28 + 0.741$

$= 5.02$

Conclusion: I have 95% confidence to say that the average waiting time (population mean) will fall in the interval [3.53, 5.02].

c) We were told the average waiting minute will be 5 minutes. But we think it could be more than 5 minutes. By using 90% as confidence level, validate the hypothesis. Show your steps and calculations [15]

ANSWER:

To answer this question, we need to follow the below steps one by one:

1. Collect sample data & Calculate sample statistics such as 'sample mean' and 'sample standard deviation':

- ✓ n = 15
- ✓ Sample mean $\bar{X}$ = 4.28
- ✓ Sample standard deviation = S = 1.63

2. State null hypothesis and alternative hypothesis:

$$H_0 : \mu = 5$$
$$H_a : \mu > 5$$

3. Based on Ha, decide it is one-tailed or two-tailed test:

- Based on alternative hypothesis H_a, it is one- tail test.
- As alternative hypothesis is larger than μ, the tail is on the right side. That means the reject region is always on the right.

4. Specify the desired level of significance:

α = 1 – confidence level.
= 1- 90%
= 1- 0.90
= 0.10

5. Determine the appropriate technique:

- ➢ σ is unknown and n = 15 which is less than 30.
- ➢ As sample size is less, this is a t- test.
- ➢ Hence use t- statistics to make the conclusion.

6. Calculate critical value:

- Since it is t- test, to calculate the t α / t- critical value, we need to have 2 inputs.
  - α = 0.10
  - df (degree of freedom) = n-1 = 15 – 1 = 14

- with the α value and df value available, the t- critical value/ t α value is calculated from the t- table and observed as follows:

$$t_c \; or \; t_\alpha = \; 1.345$$

7. Calculate the test statistic value (t-value) by using the formula: (t-score)

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{4.28 - 5}{\frac{1.63}{\sqrt{15}}}$$

$$= \frac{-\ 0.72}{\frac{1.63}{3.87}}$$

$$= \frac{-0.72}{0.42}$$

$$= -1.714$$

8.Compare the test- statistic value with the critical value / t α value. 6. Is the test statistic in the rejection region?

$$t_{STAT} < t_c \; or \; t_\alpha$$

$$-\ 1.714 < 1.345$$

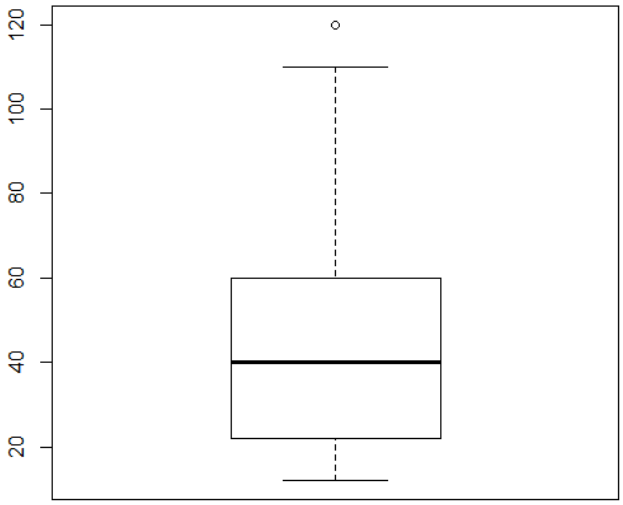
The test statistic is in the non-rejection region.

9. Draw the conclusion based on test statistic value:

- Since the test statistic falls in the non- rejection region, we fail to reject the null hypothesis. Therefore, we do not have enough evidence to conclude that the average waiting minute is more than 5 minutes with a 90% confidence level.
- In summary, based on the given data and calculations, we do not have enough evidence to support the hypothesis that the average waiting minute is more than 5 minutes with a 90% confidence level.

2. (55 points) Chicago Ventra Transit Card can be used on both CTA bus, metro and Pace buses. We are going to explore a resident's average monthly cost on CTA transportations. In this case, we performed a survey, and collect monthly cost on CTA transits from 30 people, their monthly cost can be listed as follows:
12, 12, 12, 15, 24, 35, 14, 12, 120, 55, 45, 30, 40, 40, 40, 60, 60, 40, 50, 22, 36, 28, 21, 50, 39, 60, 90, 100, 110, 100

1). [5] To further understand the distribution, we draw a boxplot as follows. Interpret the box plot.



ANSWER:
Interpretation of the boxplot:

- **The distribution:** The distribution is not symmetric; the data is not evenly distributed. The median is not exactly in the middle. From the visualization, the box plot clearly shows that it is positively skewed.

- **Skewness:** The box plot clearly follows a positive skew from the visualization.

    - To say its skewness in detail, we can use a formula:
      Skewness = (3 * (Mean - Median)) / Standard Deviation
      = (3 * (45.73 – 40)) / 30.18
      = (3 * 5.73) / 30.18
      = 17.19 / 30.18
      = 0.56
      Hence, we can say it is positively skewed / right skewed.

- **Potential Outliers:** From the visualization we can clearly see that there is a potential outlier towards the value 120.

    - To confirm the potential outliers, we need to follow the below steps:

- Calculate IQR (Interquartile range):
  - Q3 – Q1 = 60-22 = 38
- Upper bound value = Q3 + 1.5* IQR = 60 + (1.5 * 38) = 60 + 57 = 117

- Lower bound value = Q1 – 1.5* IQR = 22 – (1.5 * 38) = 22 – 57 = -35

❖ Any data values not falling between upper bound value and lower bound value are suspected to be potential outliers. Here 120 is the potential outlier that is not falling between the upper and lower bound values.

- **The variance.:** Variance is nothing but the variation or the spread of data. The variance in the boxplot is as follows.
  - ➢ Variance = $S^2 = \sum_{i=1}^{n} \frac{(xi-\overline{x})^2}{n-1}$

    Hence the variance from the above formula: $S^2 = 910.83$

- **Standard deviation:** 30.18 (it's the square root of variance- ($\sqrt{S^2}$)
- **Minimum value:** 12
- **Maximum value:** 110 (Since 120 is the outlier, it cannot be considered as the max value)
- **1st Quartile(Q1):** 22
- **2nd Quartile(Q2):** 40 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 60
- **Range:** (maximum value – minimum value) excluding the outliers.
  = 110 -12 = 98
- **Mean value/ Average value of the data:** mean = $\sum_{i=0}^{n} xi/n$
  (Dividing the sum of all values in a data set by the number of values.)
  = 45.733

- **Median value:** Median is the middle value. If the size of the data is odd there is 1 median, If the data size is even, then the sum of middle 2 numbers divided by 2.

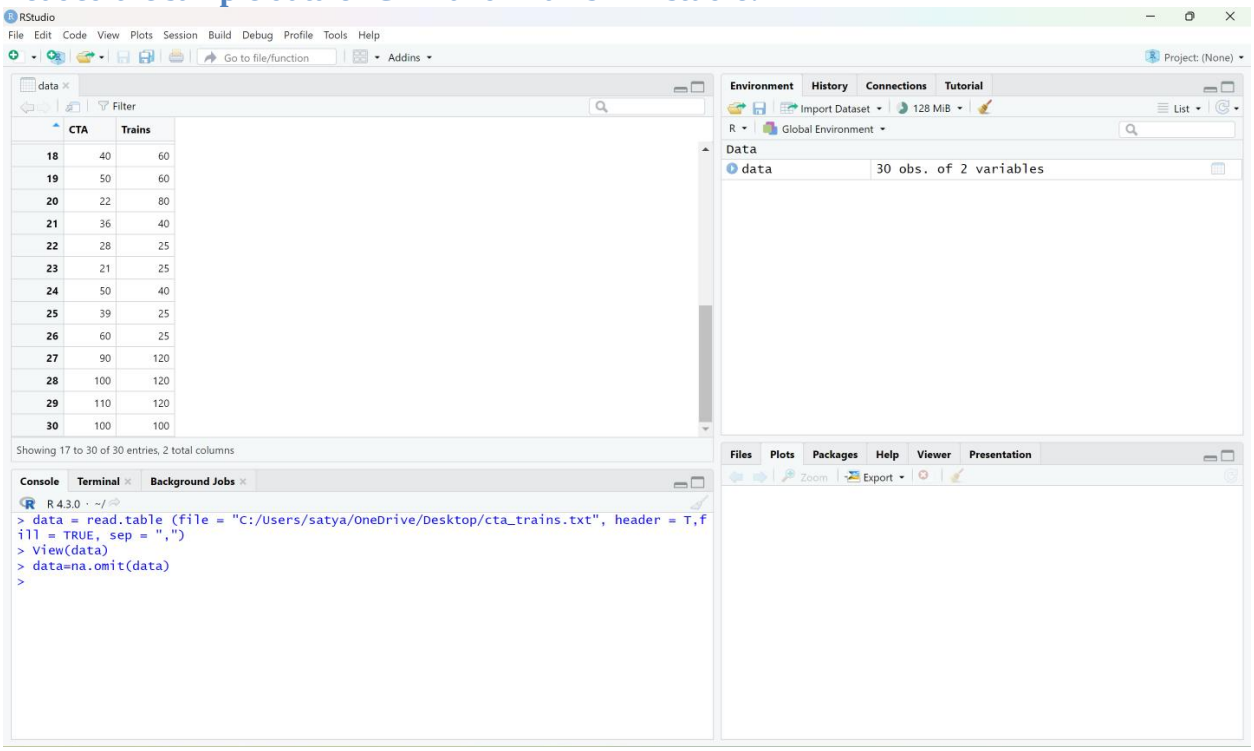
  Fist we need to arrange the data in Ascending order:
  12,12,12,12,14,15,21,22,24,28,30,35,36,39,40,40,40,40,45,50,50,55,60,60,60,90,100, 100,110,120

  Hence median = (40+40)/2 = 40 (Note: Median is also called Q2)

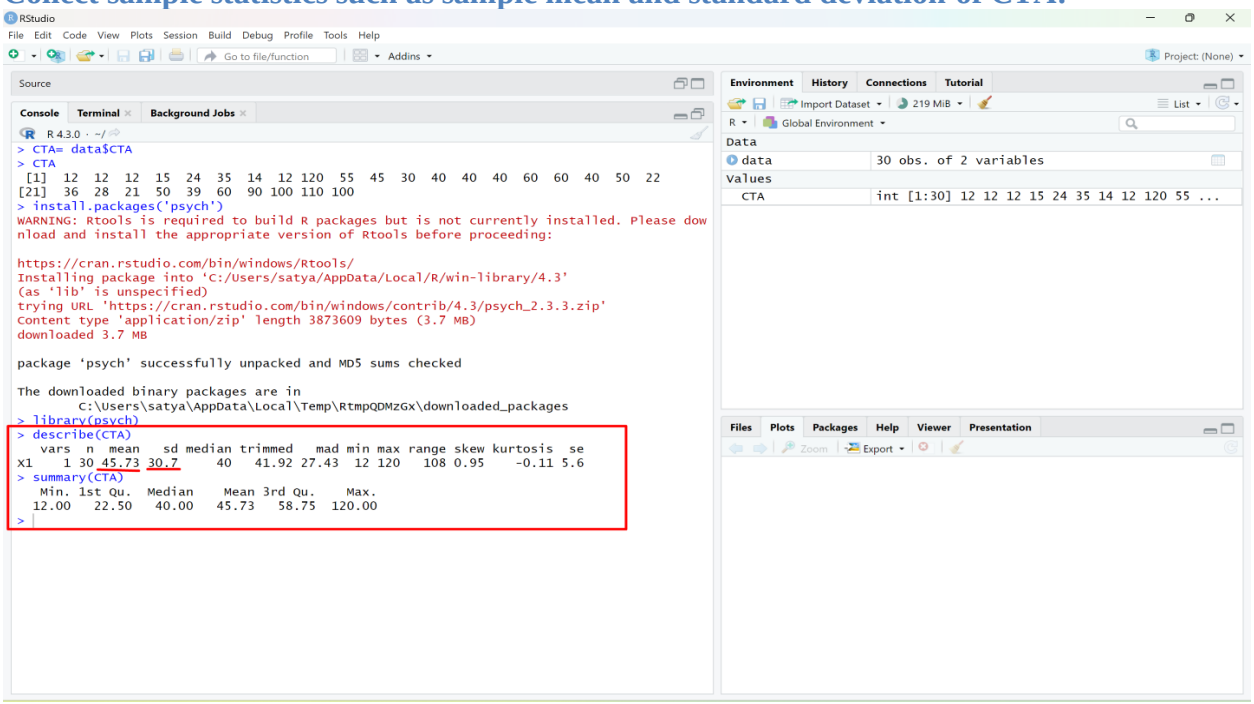
2). [15] Use the sample statistics to estimate the average monthly cost on CTA transits by Chicago residents by using 95% as the confidence level. Assume that we know the population variance is 4. Use R to solve the problem. paste your snapshot of R coding and outputs, also deliver your conclusions.

ANSWER:

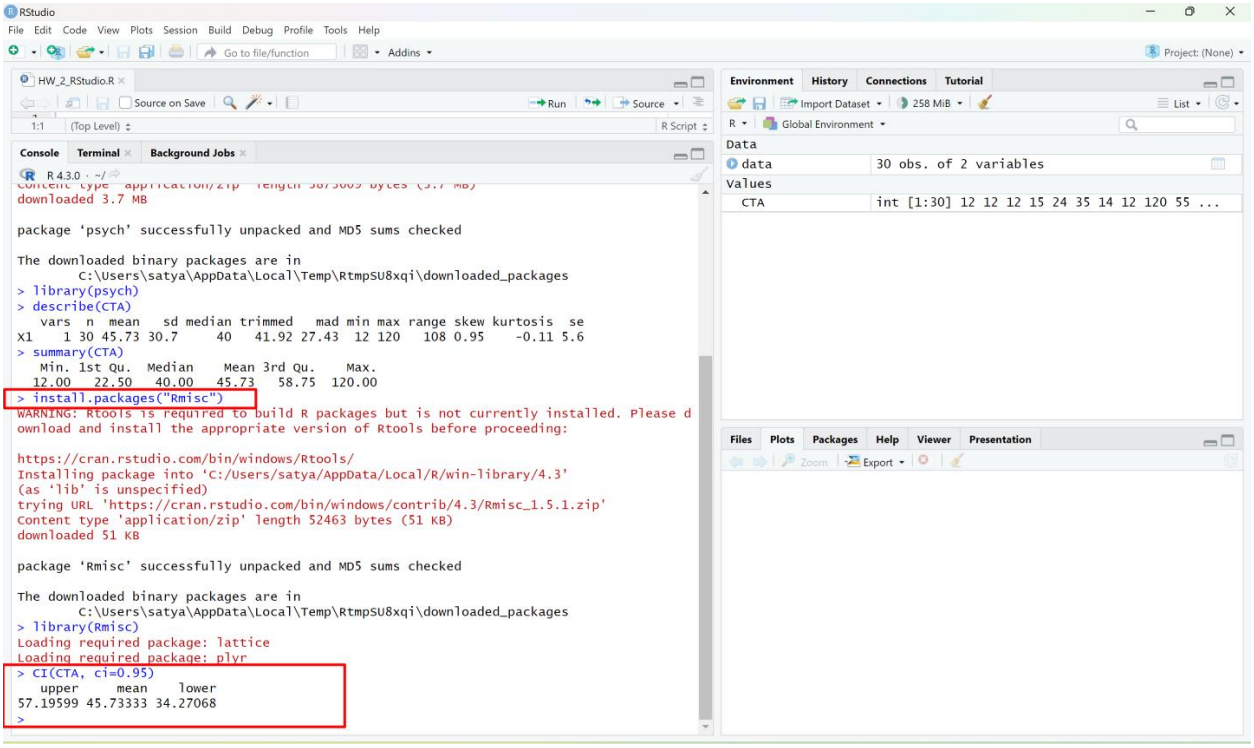
Loaded the sample data of CTA and Trains in r-studio:



Collect sample statistics such as sample mean and standard deviation of CTA:



Produce confidence interval: To Produce confidence interval, "Rmisc" package was installed & called to use the function CI in it.



Conclusion: I have 95% confidence to say that the average monthly cost on CTA transits by Chicago residents (population mean) will fall in the interval [34.27, 57.19].

4). In addition, they can use Chicago metro trains in their daily life. We ask the same group of 30 users to use Chicago metro trains only and record their monthly cost on trains. In this case, we get two groups of data as follows. We display it as two tables, since it is not able to put them on a single table. For each table, row 1 is the monthly cost by using CTA for each person, row2 is the monthly cost by using train for each person. Each column contains the costs by a same person but use different transportation (CTA vs Metro Trains) . Assume that we know they have the same population variance of 4.

user	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
row1	12	12	12	15	24	35	14	12	120	55	45	30	40	40	40
row2	10	16	13	14	28	41	16	10	80	40	75	25	41	29	40

user	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
row1	60	60	40	50	22	36	28	21	50	39	60	90	100	110	100
row2	50	50	60	60	80	40	25	25	40	25	25	120	120	120	100

It is told that there are no differences if they are going to use the CTA or Metro trains. However, we believe using CTA is more expensive than using Metro trains. We are going to use hypothesis testing to examine whether the costs by two different means should be the same or not. Assume we use 95% confidence level.

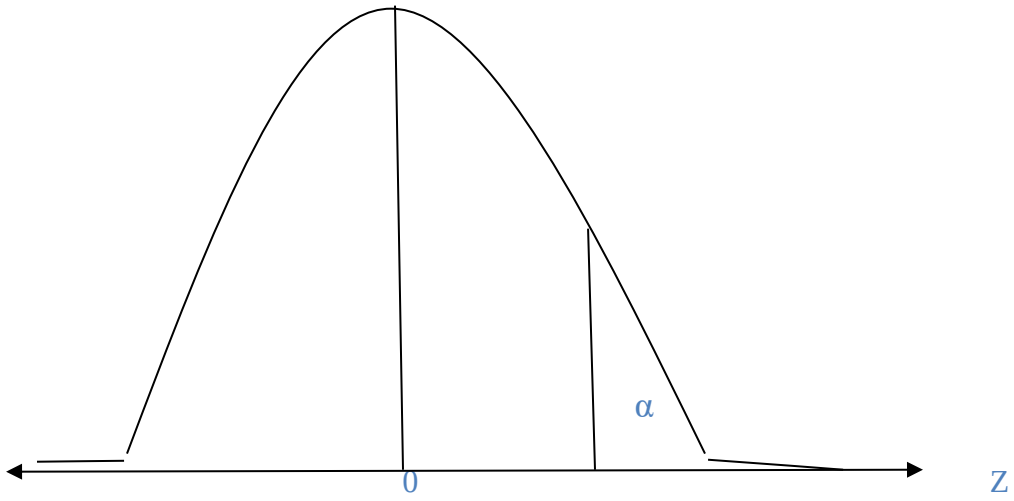
4.1), [15] write down your null and alternative hypothesis, and tell me is it a two-tailed or one-tailed test, and it is two independent or paired samples, why?

ANSWER:

-It is a 2- sample hypothesis testing. Sample 1: CTA (μ_1)
Sample 2: Metro trains. (μ_2)

Null & alternative Hypothesis representation:
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$

- It is one- tailed hypothesis test, because based on the alternative hypothesis (H_1 *or* H_a), it says that, CTA is more expensive than Metro trains. So, we can observe there is only 1 reject region (Z_α) towards the right. Z test because the sample size is 30 (n = 30) and the population variance is known.



- It is two – paired samples because the user (same group) was measured for the monthly cost twice by CTA & Metro trains. So, the 2 samples are paired & dependent on each other.

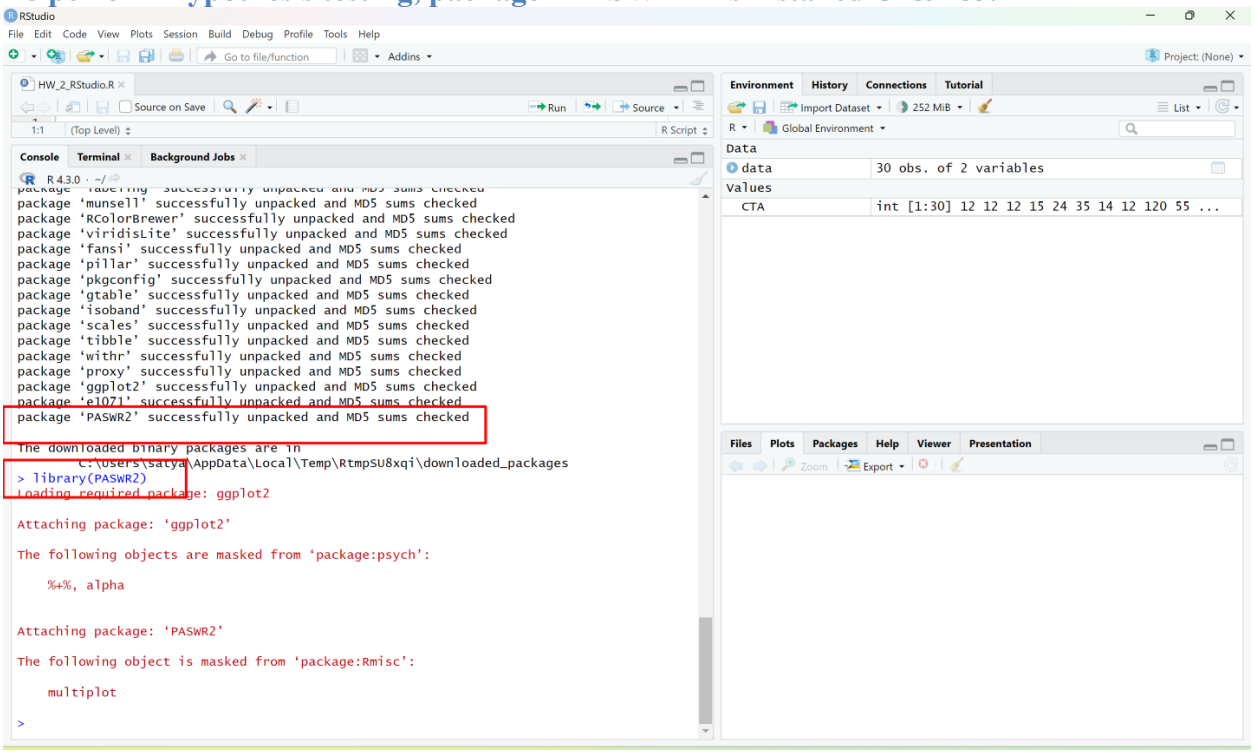
4.2), [20] Perform hypothesis testing to tell whether the costs by two different means are the same or not based on 95% confidence level, by using R. Again, give the R coding, snapshot, outputs, deliver your conclusions by referring to/explaining the outputs.

ANSWER: It is 2-sample paired hypothesis testing.

As it is 2- paired sample hypothesis, we can convert it into 1-sample and perform hypothesis testing. For that we need to calculate the difference between CTA & Trains columns & re-represent null and alternative hypothesis in 1- sample as follows:

$$\mu_{diff} = \mu_1 - \mu_2$$
$$H_0 : \mu_{diff} = 0$$
$$H_1 : \mu_{diff} > 0$$

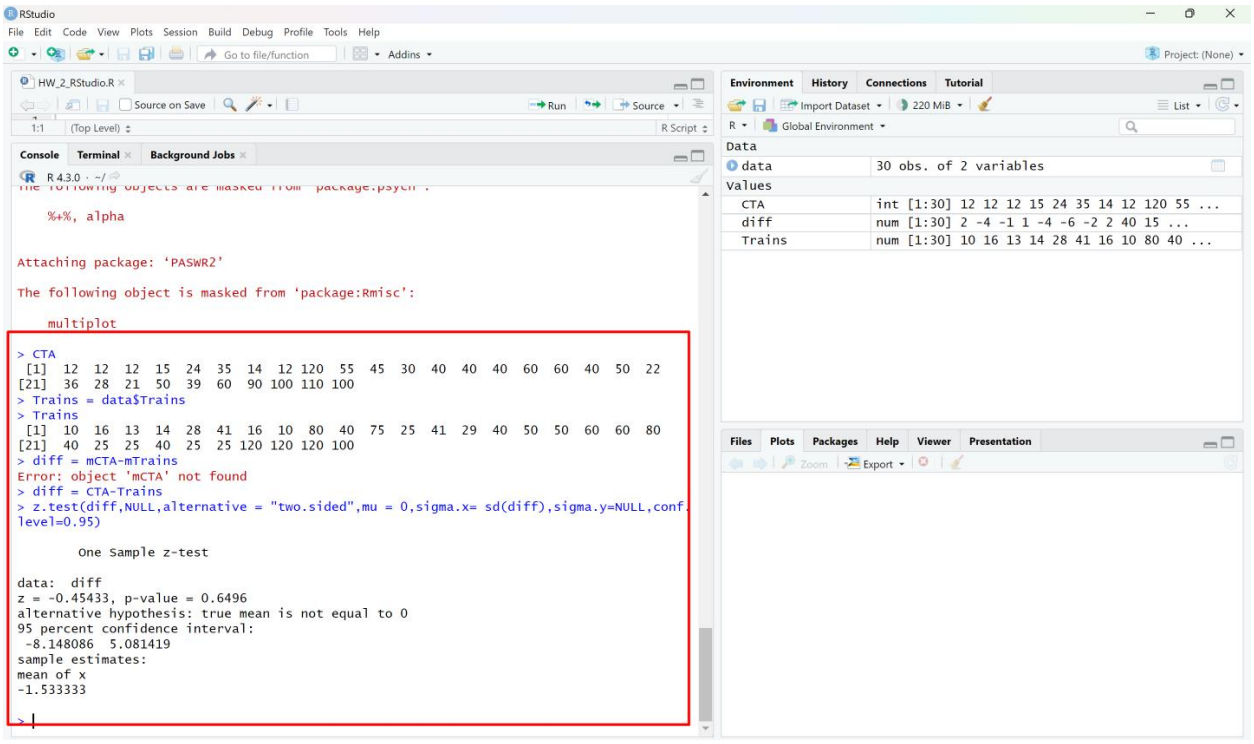
To perform hypothesis testing, package "PASWR2" is installed & called:



Now, let's perform hypothesis testing: as the sample size is 30 (n=30) which is large, we do z-test. We converted the 2-paired sample to 1-sample by calculating the difference between 2 columns.

Now we have only one column (1-sample) which is the 'diff' column. And we can perform hypothesis testing on that one 'diff' column,

Note: diff = CTA - Trains



Conclusion:

Based on the paired z-test performed at a 95% confidence level, we fail to reject the null hypothesis. The p-value is 0.6496, which is greater than the significance level of 0.05. Therefore, we do not have sufficient evidence to conclude that there is a significant difference in the monthly costs between using CTA and metro trains. The 95% confidence interval for the difference in means is -8.1480 to 5.0814, which includes zero. This further supports the conclusion that the means of the two transportation methods are likely to be equal.

Hence, we do not reject the null hypothesis based on 95% confidence level.

3. (10 points) The z-test and t-test we used for hypothesis testing (including both one-sample and two-sample hypothesis testing) are also known as "Parametric Test". Alternatively, there are other statistical tests which can be used as alternatives, and they are called "Non-Parametric Test". Search online or find other learning materials to answer the questions below.

1). (5 points) what are the differences between Parametric Test and Non-Parametric Test?

ANSWER:

Basis for comparison	Parametric Test	Non-Parametric Test
1. Definition	The parametric test is a hypothesis test that offers generalizations for claiming something about the parent population's mean. A commonly employed t-test in this context is one that is based on Student's t-statistic.	The nonparametric test is characterized as a hypothesis test without underlying assumptions, i.e., it does not call for the population's distribution to be represented by parameters.
2. Meaning	A parametric test is a statistical test in which assumptions about a population parameter are made.	Non-parametric test is a statistical test that is used when there are non-metric independent variables.
3. Population Information	Completely known	Not known
4. Applicability	Only on variables.	Both variables and attributes
5. Central tendency measurement	In general, the measure of central tendency in the parametric test is mean,	- While in the case of the nonparametric test is median. - As a result, it is also known as the distribution-free test.
6. Correlation test	The parametric test makes use of Pearson's coefficient of correlation to assess the degree of relationship between two quantitative variables.	while spearman's rank correlation is used in the nonparametric test.
7. Basis of test statistic	Distribution	Arbitrary

2). (5 points) what are the requirements/conditions to use z-test or t-test? If our data does not meet these requirements, which non-parametric test is the alternative for hypothesis testing?

ANSWER:

Requirements to use Z-test / t-test:

1) If the sample size is large (n >= 30) & population standard deviation is known => Z-test
2) If Sample size is small (n < 30) and population STD is unknown => t-test
3) If Sample size is large (n >= 30) and population STD is unknown => t-test
4) Either a Z-test or a T-test can be used to compare the means of two independent samples, such as two separate groups.

5) A paired T-test is often employed when we need to compare the means of paired samples (for instance, before and after measurements).
6) Df (degree of freedom must be known to carry t-test

- You can use non-parametric tests for hypothesis testing if your data does not fulfill these conditions or if you would prefer a non-parametric option. Non-parametric tests don't rely on presumptions about how the data are distributed.

Equivalent alternative non-parametric tests for hypothesis testing:

Parametric Test	**Non- Parametric Test**
1. Independent Sample t Test	Mann-Whitney U test: Utilizing this test, two independent groups' medians are compared. The independent samples t-test can be replaced with this method.
2. Paired samples t test	Wilcoxon signed Rank test: This test is used to determine if the median of one group deviates from a value that is hypothesized or to compare the medians of two groups that are related. It serves as a substitute for the paired samples t-test.
3. One way Analysis of Variance (ANOVA)	Kruskal Wallis Test: To compare the medians of three or more independent groups, apply this test. The one-way analysis of variance (ANOVA) test can be substituted with this method
4. One-way repeated measures Analysis of Variance	Friedman's ANOVA: The medians of three or more related groups are compared using this test. This test serves as an alternative to the repeated measures ANOVA.