Topic 6: Basic Statistical Analysis

2023-08-08

In this topic, you will learn about test of mean:

- test of single mean
- independent t-test
- paired t-test
- one-way ANOVA

test of single mean

Test of Single Mean in R Programming

A test of a single mean in R is used to determine if a sample mean is significantly different from a hypothesized population mean. It helps to assess whether the observed sample mean provides enough evidence to reject or fail to reject the null hypothesis about the population mean.

Performing a Test of Single Mean:

1. One-Sample t-test (t.test()): The most commonly used test for a single mean in R is the one-sample t-test. It compares the sample mean to a hypothesized population mean and provides a p-value indicating the level of significance.

Example: One-Sample t-test

Suppose we have a sample of exam scores (out of 100) and want to test if the sample mean is significantly different from the population mean of 75.

```r
# Sample data
scores <- c(82, 78, 85, 90, 73, 88, 79, 84, 86, 81)

# Perform one-sample t-test
test_result <- t.test(scores, mu = 75)

# Print the test result
print(test_result)
```

```
##
##  One Sample t-test
##
## data:  scores
## t = 4.7295, df = 9, p-value = 0.001075
## alternative hypothesis: true mean is not equal to 75
## 95 percent confidence interval:
```

```
##  78.96487 86.23513
## sample estimates:
## mean of x
##      82.6
```

Interpreting the Output:

The output of the **t.test()** function includes the t-statistic, degrees of freedom, the observed sample mean, the hypothesized population mean (mu), and the p-value. The p-value indicates the probability of obtaining the observed sample mean or more extreme values under the assumption that the null hypothesis (population mean = hypothesized mean) is true.

If the p-value is below a pre-defined significance level (commonly 0.05), we reject the null hypothesis and conclude that there is significant evidence that the sample mean is different from the hypothesized population mean.

Alternative Form for One-Sample t-test:

2. t.test() with mu argument: You can also use the mu argument to specify the hypothesized population mean.

Example: One-Sample t-test with mu argument

```r
# Sample data
scores <- c(82, 78, 85, 90, 73, 88, 79, 84, 86, 81)

# Perform one-sample t-test with mu argument
test_result <- t.test(scores, mu = 75)

# Print the test result
print(test_result)
```

```
##
##  One Sample t-test
##
## data:  scores
## t = 4.7295, df = 9, p-value = 0.001075
## alternative hypothesis: true mean is not equal to 75
## 95 percent confidence interval:
##  78.96487 86.23513
## sample estimates:
## mean of x
##      82.6
```

Both of the above examples perform the same one-sample t-test, but the second example explicitly sets the hypothesized population mean using the mu argument.

Summary:

1. A test of a single mean in R helps to determine if a sample mean is significantly different from a hypothesized population mean.
2. The one-sample t-test is commonly used for this purpose in R, and it can be performed using the **t.test()** function.
3. The p-value obtained from the test helps to make a decision about rejecting or failing to reject the null hypothesis.

Independent t-test

Independent t-Test in R Programming

The independent t-test (also known as the two-sample t-test) is used to compare the means of two independent groups to determine if there is a significant difference between their sample means. It is commonly used when you have two groups, and you want to assess whether the means of a continuous variable are significantly different between these groups.

Performing an Independent t-Test:

1. Using **t.test()**: The **t.test()** function in R can be used to perform an independent t-test. It assumes that the two samples are independent and have approximately equal variances.

Example: Independent t-Test

Suppose we have two groups of exam scores (out of 100), one for males and another for females, and we want to compare whether the mean scores are significantly different between the two groups.

```r
# Sample data for two groups
scores_male <- c(82, 78, 85, 90, 73, 88, 79, 84, 86, 81)
scores_female <- c(78, 76, 80, 88, 70, 84, 77, 82, 85, 80)

# Perform independent t-test
test_result <- t.test(scores_male, scores_female)

# Print the test result
print(test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  scores_male and scores_female
## t = 1.1373, df = 17.997, p-value = 0.2703
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.203157  7.403157
## sample estimates:
## mean of x mean of y
##      82.6      80.0
```

Interpreting the Output:

The output of the **t.test()** function includes the t-statistic, degrees of freedom, the observed difference in sample means, and the p-value. The p-value indicates the probability of obtaining the observed difference in means or more extreme differences under the assumption that the two groups have equal means.

If the p-value is below a pre-defined significance level (commonly 0.05), we reject the null hypothesis and conclude that there is a significant difference between the means of the two groups.

Alternative Form for Independent t-Test:

2. Using **formula syntax in t.test()**: You can also use the formula syntax in **t.test()** to specify the grouping variable and the response variable.

Example: Independent t-Test with Formula Syntax

```r
# Create a data frame with two columns (Group and Scores)
data <- data.frame(
  Group = rep(c("Male", "Female"), each = 10),
  Scores = c(82, 78, 85, 90, 73, 88, 79, 84, 86, 81, 78, 76, 80, 88, 70, 84, 77, 82, 85, 80)
)

# Perform independent t-test using formula syntax
test_result <- t.test(Scores ~ Group, data = data)

# Print the test result
print(test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Scores by Group
## t = -1.1373, df = 17.997, p-value = 0.2703
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -7.403157  2.203157
## sample estimates:
## mean in group Female   mean in group Male
##                 80.0                 82.6
```

Both of the above examples perform the same independent t-test, but the second example uses the formula syntax to specify the grouping variable (Group) and the response variable (Scores).

Summary:

1. The independent t-test is used to compare the means of two independent groups to determine if there is a significant difference between their sample means.
2. The **t.test()** function in R can be used to perform an independent t-test, either with vector inputs or formula syntax.
3. The p-value obtained from the test helps to make a decision about rejecting or failing to reject the null hypothesis of equal means between the two groups.

Paired t-Test

Paired t-Test in R Programming

The paired t-test (also known as the dependent t-test) is used to compare the means of two related or paired samples to determine if there is a significant difference between their means. It is commonly used when you have two sets of observations that are linked or matched in some way.

Performing a Paired t-Test:

1. Using **t.test()**: The **t.test()** function in R can be used to perform a paired t-test. It assumes that the two samples are dependent (paired) and come from the same population.

Example: Paired t-Test

Suppose we have two sets of exam scores **(before and after a training program)** and want to determine if there is a significant improvement in scores after the program.

```r
# Sample data for before and after scores
scores_before <- c(82, 78, 85, 90, 73, 88, 79, 84, 86, 81)
scores_after <- c(88, 80, 90, 95, 78, 90, 82, 89, 92, 87)

# Perform paired t-test
test_result <- t.test(scores_before, scores_after, paired = TRUE)

# Print the test result
print(test_result)
```

```
##
##  Paired t-test
##
## data:  scores_before and scores_after
## t = -9, df = 9, p-value = 8.538e-06
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.631079 -3.368921
## sample estimates:
## mean difference
##            -4.5
```

Interpreting the Output:

The output of the **t.test()** function includes the t-statistic, degrees of freedom, the observed difference in paired sample means, and the p-value. The p-value indicates the probability of obtaining the observed difference in means or more extreme differences under the assumption that the two samples come from the same population.

If the p-value is below a pre-defined significance level (commonly 0.05), we reject the null hypothesis and conclude that there is a significant difference between the means of the paired samples.

Alternative Form for Paired t-Test:

2. The paired t-test can also be conducted in an alternative form using the **pair.t.test()** function in R. This function specifically focuses on paired t-tests and provides a more concise syntax for conducting the test for paired samples.

Function: t.test()

Syntax:

```r
t.test(x, y, alternative = "less")
```

- x: A numeric vector representing the measurements of the first group.
- y: A numeric vector representing the measurements of the second group.
- alternative: The alternative hypothesis. Options are **“two.sided” (default), “less”, or “greater”**.

Example: Paired t-Test using pair.t.test()

Using the same example of before and after scores:

```r
# Sample data: Before and after scores
before <- c(78, 85, 92, 76, 80)
after <- c(85, 90, 95, 82, 88)

# Perform paired t-test using t.test()
result <- t.test(before, after, paired = TRUE, alternative = "less")

# Print the results
print(result)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = -6.7424, df = 4, p-value = 0.001261
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##       -Inf -3.966116
## sample estimates:
## mean difference
##            -5.8
```

The t.test() function performs the paired t-test between the two vectors before and after with the specified alternative hypothesis. The output will include the t-statistic, degrees of freedom, p-value, and confidence interval.

The interpretation of the results and the significance testing process remains the same as in the previous example.

Summary:

1. The paired t-test is used to compare the means of two related or paired samples to determine if there is a significant difference between their means.
2. The **t.test()** function in R can be used to perform a paired t-test, either with vector inputs or formula syntax.
3. The p-value obtained from the test helps to make a decision about rejecting or failing to reject the null hypothesis of equal means for the paired samples.

One-Way ANOVA

One-Way ANOVA in R Programming

One-Way Analysis of Variance (ANOVA) is a statistical technique used to compare the means of three or more groups to determine if there are significant differences between them. It is commonly used when you have one categorical independent variable and one continuous dependent variable.

Performing One-Way ANOVA:

1. Using **aov()**: The **aov()** function in R is used to perform one-way ANOVA. It fits a linear model to the data and tests for differences between the means of the groups.

Example: One-Way ANOVA

Suppose we have three groups (A, B, and C) with exam scores and want to test if there are significant differences in scores between these groups.

```r
# Sample data for three groups
group_a <- c(82, 78, 85, 90, 73)
group_b <- c(88, 80, 90, 95, 78)
group_c <- c(79, 84, 86, 81, 82)

# Combine data into a single data frame
data <- data.frame(
  Score = c(group_a, group_b, group_c),
  Group = factor(rep(c("A", "B", "C"), each = 5))
)

# Perform one-way ANOVA
anova_result <- aov(Score ~ Group, data = data)

# Summarize the ANOVA result
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Group        2   60.4   30.20   0.908  0.429
## Residuals   12  399.2   33.27
```

Interpreting the Output:

The output of the **summary()** function on the **aov()** result provides the ANOVA table, which includes the F-statistic, degrees of freedom, and p-value. The p-value indicates the probability of obtaining the observed differences in means or more extreme differences under the assumption that the means of all groups are equal.

If the p-value is below a pre-defined significance level (commonly 0.05), we reject the null hypothesis and conclude that there are significant differences in means between at least two groups.

2. **Post-hoc Tests:**

When the ANOVA result is significant, post-hoc tests can be performed to determine which groups differ from each other. Popular post-hoc tests include Tukey's Honestly Significant Difference (HSD) and Bonferroni correction.

Example: Post-hoc Test using TukeyHSD()

```r
# Perform Tukey's HSD post-hoc test
posthoc_test <- TukeyHSD(anova_result)

# Print the post-hoc test results
print(posthoc_test)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Score ~ Group, data = data)
##
## $Group
##     diff        lwr       upr     p adj
## B-A  4.6  -5.131914 14.331914 0.4423167
## C-A  0.8  -8.931914 10.531914 0.9738916
## C-B -3.8 -13.531914  5.931914 0.5660912
```

Summary:

1. One-Way ANOVA is used to compare the means of three or more groups to determine if there are significant differences between them.
2. The **aov()** function in R is used to perform one-way ANOVA.
3. The p-value obtained from the ANOVA helps to make a decision about rejecting or failing to reject the null hypothesis of equal means for the groups.
4. Post-hoc tests can be performed when the ANOVA result is significant to identify which groups differ significantly from each other.