r-statistics.co (/) by Selva Prabhakaran

Search..

•

Tutorial

R Tutorial (R-Tutorial.html)

ggplot2

ggplot2 Short Tutorial (ggplot2-Tutorial-With-R.html)

ggplot2 Tutorial 1 - Intro (Complete-Ggplot2-Tutorial-Part1-With-R-Code.html)

ggplot2 Tutorial 2 - Theme (Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html)

ggplot2 Tutorial 3 - Masterlist (Top50-Ggplot2-Visualizations-MasterList-R-Code.html)

ggplot2 Quickref (ggplot2-cheatsheet.html)

Foundations

Linear Regression (Linear-Regression.html)

Statistical Tests (Statistical-Tests-in-R.html)

Missing Value Treatment (Missing-Value-Treatment-With-R.html)

Outlier Analysis (Outlier-Treatment-With-R.html)

Feature Selection (Variable-Selection-and-Importance-With-R.html)

Model Selection (Model-Selection-in-R.html)

Logistic Regression (Logistic-Regression-With-R.html)

Advanced Linear Regression (Environments.html)

Advanced Regression Models

Advanced Regression Models (adv-regression-models.html)

Time Series

Time Series Analysis (Time-Series-Analysis-With-R.html)

Time Series Forecasting (Time-Series-Forecasting-With-R.html)

More Time Series Forecasting (Time-Series-Forecasting-With-R-part2.html)

High Performance Computing

Parallel computing (Parallel-Computing-With-R.html)

Strategies to Speedup R code (Strategies-To-Improve-And-Speedup-R-Code.html)

Useful Techniques

Association Mining (Association-Mining-With-R.html)

Multi Dimensional Scaling (Multi-Dimensional-Scaling-With-R.html)

Optimization (Profiling.html)

InformationValue package (Information-Value-With-R.html)

Stay up-to-date. Subscribe! (https://docs.google.com/forms/d/1xkMYkLNFU9U39Dd8S_2JC0p8B5t6_Yq6zUQjanQQJpY/viewform)

Chat! (https://docs.google.com/forms/d/13GrkCFcNa-TOIllQghsz2SIEbc-YqY9eJX02B19l5Ow/viewform)

Contents

Statistical Tests

> This chapter explains the purpose of some of the most commonly used statistical tests and how to implement them in R

1. One Sample t-Test

Why is it used?

It is a parametric test used to test if the mean of a sample from a normal distribution could reasonably be a specific value.

```
set.seed(100)
x <- rnorm(50, mean = 10, sd = 0.5)
t.test(x, mu=10) # testing if mean of x could be
#=> One Sample t-test
#=>
#=> data:  x
#=> t = 0.70372, df = 49, p-value = 0.4849
#=> alternative hypothesis: true mean is not equal to 10
#=> 95 percent confidence interval:
#=>   9.924374 10.157135
#=> sample estimates:
#=> mean of x
#=>  10.04075
```

How to interpret?

In above case, the p-Value is not less than significance level of 0.05, therefore the null hypothesis that the mean=10 cannot be rejected. Also note that the 95% confidence interval range includes the value 10 within its range. So, it is ok to say the mean of ‘x’ is 10, especially since ‘x’ is assumed to be normally distributed. In case, a normal distribution is not assumed, use wilcoxon signed rank test shown in next section.

Note: Use conf.level argument to adjust the confidence level.

2. Wilcoxon Signed Rank Test

Why / When is it used?

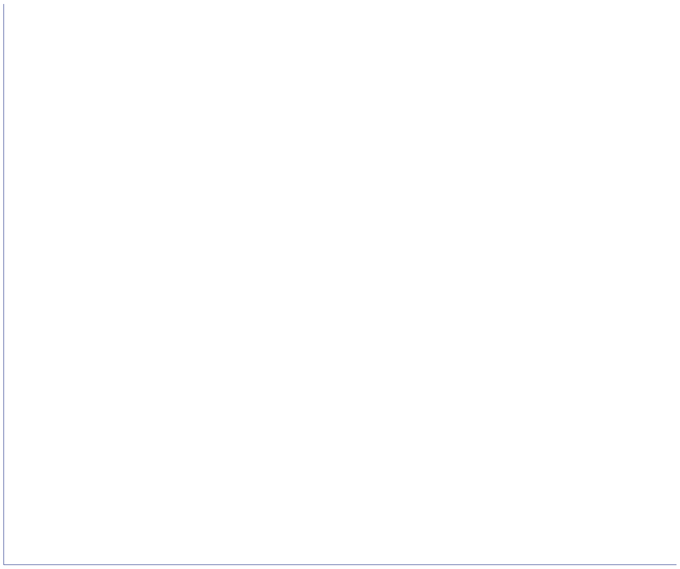
To test the mean of a sample when normal distribution is not assumed. Wilcoxon signed rank test can be an alternative to t-Test, especially when the data sample is not assumed to follow a normal distribution. It is a non-parametric method used to test if an estimate is different from its true value.

```
numeric_vector <- c(20, 29, 24, 19, 20, 22, 28, 23, 19, 19)
wilcox.test(numeric_vector, mu=20, conf.int = TRUE)
#>  Wilcoxon signed rank test with continuity correction
#>
#> data:  numeric_vector
#> V = 30, p-value = 0.1056
#> alternative hypothesis: true location is not equal to 20
#> 90 percent confidence interval:
#>  19.00006 25.99999
#> sample estimates:
#> (pseudo)median
#>       23.00002
```

How to interpret?

If p-Value < 0.05, reject the null hypothesis and accept the alternate mentioned in your R code’s output. Type example(wilcox.test) in R console for more illustration.

3. Two Sample t-Test and Wilcoxon Rank Sum Test



Both t.Test and Wilcoxon rank test can be used to compare the mean of 2 samples. The difference is t-Test assumes the samples being tests is drawn from a normal distribution, while, Wilcoxon’s rank sum test does not.

How to implement in R?

Pass the two numeric vector samples into the t.test() when sample is distributed ‘normal’y and wilcox.test() when it isn’t assumed to follow a normal distribution.

```
x <- c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46)
y <- c(1.15, 0.88, 0.90, 0.74, 1.21)
wilcox.test(x, y, alternative = "g")  # g for greater
#=> Wilcoxon rank sum test
#=> 
#=> data:  x and y
#=> W = 35, p-value = 0.1272
#=> alternative hypothesis: true location shift is greater than 0
```

With a p-Value of 0.1262, we cannot reject the null hypothesis that both x and y have same means.

```
t.test(1:10, y = c(7:20))      # P = .00001855
#=> Welch Two Sample t-test
#=> 
#=> data:  1:10 and c(7:20)
#=> t = -5.4349, df = 21.982, p-value = 1.855e-05
#=> alternative hypothesis: true difference in means is not equal to 0
#=> 95 percent confidence interval:
#=>   -11.052802  -4.947198
#=> sample estimates:
#=> mean of x mean of y
#=>       5.5      13.5
```

With p-Value < 0.05, we can safely reject the null hypothesis that there is no difference in mean.

What if we want to do a 1-to-1 comparison of means for values of x and y?

```
# Use paired = TRUE for 1-to-1 comparison of observations.
t.test(x, y, paired = TRUE) # when observations are paired, use 'paired' argument.
wilcox.test(x, y, paired = TRUE) # both x and y are assumed to have similar shapes
```

When can I conclude if the mean’s are different?

Conventionally, If the p-Value is less than significance level (ideally 0.05), reject the null hypothesis that both means are the are equal.

4. Shapiro Test

Why is it used?

To test if a sample follows a *normal distribution*.

```
shapiro.test(numericVector) # Does myVec follow a normal disbn?
```

Lets see how to do the test on a sample from a normal distribution.

```
# Example: Test a normal distribution
set.seed(100)
normaly_disb <- rnorm(100, mean=5, sd=1) # generate a normal distribution
shapiro.test(normaly_disb)  # the shapiro test.
#=> Shapiro-Wilk normality test
#=>
#=> data:  normaly_disb
#=> W = 0.98836, p-value = 0.535
```

How to interpret?

The null hypothesis here is that the sample being tested is normally distributed. Since the p Value is not less that the significane level of 0.05, we don’t reject the null hypothesis. Therefore, the tested sample is confirmed to follow a normal distribution (thou, we already know that!).

```
# Example: Test a uniform distribution
set.seed(100)
not_normaly_disb <- runif(100)  # uniform distribution.
shapiro.test(not_normaly_disb)
#=>     Shapiro-Wilk normality test

#=> data:  not_normaly_disb
#=> W = 0.96509, p-value = 0.009436
```

How to interpret?

If p-Value is less than the significance level of 0.05, the null-hypothesis that it is normally distributed can be rejected, which is the case here.

5. Kolmogorov And Smirnov Test

Kolmogorov-Smirnov test is used to check whether 2 samples follow the same distribution.

```
ks.test(x, y) # x and y are two numeric vector
```

```
# From different distributions
x <- rnorm(50)
y <- runif(50)
ks.test(x, y)  # perform ks test
#=> Two-sample Kolmogorov-Smirnov test
#=>
#=> data:  x and y
#=> D = 0.58, p-value = 4.048e-08
#=> alternative hypothesis: two-sided
```

```
# Both from normal distribution
x <- rnorm(50)
y <- rnorm(50)
ks.test(x, y)  # perform ks test
#=> Two-sample Kolmogorov-Smirnov test
#=>
#=> data:  x and y
#=> D = 0.18, p-value = .3959
#=> alternative hypothesis: two-sided
```

How to tell if they are from the same distribution ?

If p-Value < 0.05 (significance level), we reject the null hypothesis that they are drawn from same distribution. In other words, p < 0.05 implies x and y from different distributions

6. Fisher’s F-Test

Fisher’s F test can be used to check if two samples have same variance.

```
var.test(x, y)  # Do x and y have the same variance?
```

Alternatively fligner.test() and bartlett.test() can be used for the same purpose.

7. Chi Squared Test

Chi-squared test in R can be used to test if two categorical variables are dependent, by means of a contingency table.

Example use case: You may want to figure out if big budget films become box-office hits. We got 2 categorical variables (Budget of film, Success Status) each with 2 factors (Big/Low budget and Hit/Flop), which forms a 2 x 2 matrix.

```
chisq.test(table(categorical_X, categorical_Y), correct = FALSE)  # Yates continuity correction not applied
#or
summary(table(categorical_X, categorical_Y)) # performs a chi-squared test.
# Sample results
#=> Pearson's Chi-squared test
#=> data:  M
#=> X-squared = 30.0701, df = 2, p-value = 2.954e-07
```

How to tell if x, y are independent?

There are two ways to tell if they are independent:

1. **By looking at the p-Value**: If the p-Value is less that 0.05, we fail to reject the null hypothesis that the x and y are independent. So for the example output above, (p-Value=2.954e-07), we reject the null hypothesis and conclude that x and y are not independent.
2. **From Chi.sq value**: For 2 x 2 contingency tables with 2 degrees of freedom (d.o.f), if the Chi-Squared calculated is greater than 3.841 (critical value), we reject the null hypothesis that the variables are independent. To find the critical value of larger d.o.f contingency tables, use qchisq(0.95, n-1), where n is the number of variables.

8. Correlation

Why is it used?

To test the linear relationship of two continuous variables

The cor.test() function computes the correlation between two continuous variables and test if the y is dependent on the x. The null hypothesis is that the true correlation between *x* and *y* is zero.

```
cor.test(x, y) # where x and y are numeric vectors.
```

```
cor.test(cars$speed, cars$dist)
#=> Pearson's product-moment correlation
#=>
#=> data:  cars$speed and cars$dist
#=> t = 9.464, df = 48, p-value = 1.49e-12
#=> alternative hypothesis: true correlation is not equal to 0
#=> 95 percent confidence interval:
#=>   0.6816422 0.8862036
#=> sample estimates:
#=>       cor
#=> 0.8068949
```

How to interpret?

If the p Value is less than 0.05, we reject the null hypothesis that the true correlation is zero (i.e. they are independent). So in this case, we reject the null hypothesis and conclude that *dist* is dependent on *speed*.

9. More Commonly Used Tests

```
fisher.test(contingencyMatrix, alternative = "greater")  # Fisher's exact test to test independence of rows and columns in contingency table
friedman.test()  # Friedman's rank sum non-parametric test
```

There are more useful tests available in various other packages.

The package `lawstat` has a good collection. The outliers package has a number of test for testing for presence of outliers.