# Simple statistical tests

# Các kiểm định thống kê cơ bản

This page demonstrates how to conduct simple statistical tests using **base** R, **rstatix**, and **gtsummary**.

* T-test
* Shapiro-Wilk test
* Wilcoxon rank sum test
* Kruskal-Wallis test
* Chi-squared test
* Correlations between numeric variables

…many other tests can be performed, but we showcase just these common ones and link to further documentation.

Chương này sẽ trình bày cách để thực hiện các phép kiểm định thống kê cơ bản bằng cách sử dụng **base** R, **rstatix**, và **gtsummary**.

* Kiểm định t
* Kiểm định Shapiro-Wilk
* Kiểm định tổng thứ hạng Wilcoxon
* Kiểm định Kruskal-Wallis
* Kiểm định Chi-squared (Chi bình phương)
* Tương quan giữa các biến định lượng

... nhiều kiểm định khác có thể được thực hiện, nhưng chúng tôi chỉ trình bày các kiểm định thông dụng và kết nối với các phần khác trong cuốn sổ tay này

Each of the above packages bring certain advantages and disadvantages:

* Use **base** R functions to print a statistical outputs to the R Console
* Use **rstatix** functions to return results in a data frame, or if you want tests to run by group
* Use **gtsummary** if you want to quickly print publication-ready tables

Mỗi gói lệnh được đề cập bên trên đều có một số ưu điểm và khuyết điểm nhất định

* Sử dụng các câu lệnh của **base** R để cho kết quả đầu ra trong R Console
* Sử dụng các câu lệnh của **rstatix** để cho kết quả dưới dạng cấu trúc số liệu hoặc khi muốn thực hiện các kiểm định theo nhóm
* Sử dụng các câu lệnh của **gtsummary** khi muốn kết quả là các bảng biểu có thể sử dụng được ngay

## Preparation

## Các bước chuẩn bị

### Load packages

### Tải các gói lệnh

This code chunk shows the loading of packages required for the analyses. In this handbook we emphasize p\_load() from **pacman**, which installs the package if necessary and loads it for use. You can also load installed packages with library() from **base** R. See the page on [R basics](#r-basics) for more information on R packages.

Đoạn mã này hiển thị việc tải các gói lệnh cần thiết cho phân tích. Trong cuốn sổ tay này, chúng tôi nhấn mạnh đến hàm p\_load()trong gói lệnh **pacman**, cài đặt gói lệnh nếu cần thiết và tải để sử dụng. Các gói lệnh đã cài đặt cũng có thể được tải bằng library()từ **base** R. Xem thêm thông tin các gói lệnh của R trong chương [Cơ](#r-basics) bản về R.

### Import data

### Nhập số liệu

We import the dataset of cases from a simulated Ebola epidemic. If you want to follow along, [click to download the “clean” linelist](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds) (as .rds file). Import your data with the import() function from the **rio** package (it accepts many file types like .xlsx, .rds, .csv - see the [Import and export](#import-and-export) page for details).

The first 50 rows of the linelist are displayed below.

Chúng ta nhập bộ số liệu của các ca bệnh về một vụ dịch Ebola mô phỏng. Để tiện theo dõi, [nhấp vào đây để tải số liệu các ca bệnh theo dòng (linelist](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds)) đã được làm sạch (dưới dạng .rds). Nhập số liệu bằng hàm import() từ gói lệnh **rio** (nó chấp nhận nhiều loại tập tin như .xlsx, .rds, .csv – xem chi tiết trong chương [Nhập và xuất số liệu](#import-and-export)).

50 hàng đầu tiên của tập tin linelist được hiển thị bên dưới.

## **base** R

## Các kiểm định trong **base** R

You can use **base** R functions to conduct statistical tests. The commands are relatively simple and results will print to the R Console for simple viewing. However, the outputs are usually lists and so are harder to manipulate if you want to use the results in subsequent operations.

Các lệnh trong **base** R có thể được sử dụng để thực hiện các kiểm định thống kê. Các câu lệnh tương đối đơn giản và kết quả sẽ hiển thị trong bảng điều khiển R Console. Tuy nhiên, kết quả đầu ra thường dưới dạng liệt kê, vì thế sẽ khó thao tác hơn nếu muốn sử dụng kết quả trong các thao tác tiếp theo.

### T-tests

### Kiểm định t

A [t-test](https://en.wikipedia.org/wiki/Student%27s_t-test), also called “Student’s t-Test”, is typically used to determine if there is a significant difference between the means of some numeric variable between two groups. Here we’ll show the syntax to do this test depending on whether the columns are in the same data frame.

**Syntax 1:** This is the syntax when your numeric and categorical columns are in the same data frame. Provide the numeric column on the left side of the equation and the categorical column on the right side. Specify the dataset to data =. Optionally, set paired = TRUE, and conf.level = (0.95 default), and alternative = (either “two.sided”, “less”, or “greater”). Enter ?t.test for more details.

**Syntax 2:** You can compare two separate numeric vectors using this alternative syntax. For example, if the two columns are in different data sets.

You can also use a t-test to determine whether a sample mean is significantly different from some specific value. Here we conduct a one-sample t-test with the known/hypothesized population mean as mu =:

Một [kiểm định t](https://en.wikipedia.org/wiki/Student%27s_t-test), hay cũng được gọi là “Student’s t-Test”, thường được sử dụng để xác định có sự khác biệt có ý nghĩa thống kê giữa giá trị trung bình của hai nhóm. Bên dưới là cú pháp để thực hiện kiểm định này tùy thuộc vào các cột có trong cùng một khung số liệu hay không.

**Cú pháp 1:** Đây là cú pháp khi cột của biến liên tục và phân loại nằm trong cùng một khung số liệu. Đặt biến liên tục bên trái và biến phân loại bên phải của phương trình. Ghi rõ bộ số liệu sau data =. Các tùy chọn khác như số liệu bắt cặp, viết thêm paired = TRUE, khoảng tin cậy, viết thêm conf.level = (mặc định là 0.95), và giả thuyết thay thế alternative = (hai đuôi - “two.sided”, hoặc một đuôi nhỏ hơn hay lớn hơn - “less”, or “greater”). Viết lệnh ?t.test để biết thêm chi tiết.

**Cú pháp 2:** Đây là cú pháp khi so sánh hai véc tơ dạng số. Ví dụ như hai cột nằm trong hai bộ số liệu khác nhau.

Kiểm định t cũng được sử dụng để xác định có sự khác biệt có ý nghĩa thống kê giữa giá trị trung bình của mẫu với một số giá trị cụ thể. Đây là phép kiểm kịnh t cho một mẫu với trung bình dân số giả thuyết/đã biết như mu =:

### Shapiro-Wilk test

### Kiểm định Shapiro-Wilk

The [Shapiro-Wilk test](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test) can be used to determine whether a sample came from a normally-distributed population (an assumption of many other tests and analysis, such as the t-test). However, this can only be used on a sample between 3 and 5000 observations. For larger samples a [quantile-quantile plot](https://ggplot2.tidyverse.org/reference/geom_qq.html) may be helpful.

[Kiểm định Shapiro-Wilk](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test) có thể được sử để xác định xem một mẫu có phân bố bình thường/phân bố chuản hay không (một giả định của nhiều kiểm định khác, ví dụ như kiểm định t). Tuy nhiên, phép kiểm định này chỉ có thể được sử dụng cho một mẫu có từ 3 đến 5000 quan sát. Đối với cỡ mẫu lớn hơn, nên sử dụng biểu đồ [quantile-quantile.](https://ggplot2.tidyverse.org/reference/geom_qq.html)

### Wilcoxon rank sum test

### Kiểm định tổng thứ hạng Wilcoxon

The Wilcoxon rank sum test, also called the [Mann–Whitney U test](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test), is often used to help determine if two numeric samples are from the same distribution when their populations are not normally distributed or have unequal variance.

Kiểm định tổng thứ hạng Wilcoxon, hay còn gọi kiểm định [Mann–Whitney U](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test), thường được sử dụng để giúp xác định xem hai mẫu có cùng phân bố hay không khi dân số của nó không có phân bố chuẩn hoặc có phương sai không bằng nhau.

### Kruskal-Wallis test

### Kiểm định Kruskal-Wallis

The [Kruskal-Wallis test](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance) is an extension of the Wilcoxon rank sum test that can be used to test for differences in the distribution of more than two samples. When only two samples are used it gives identical results to the Wilcoxon rank sum test.

Kiểm định [Kruskal-Wallis](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance) là một phần mở rộng của kiểm định tổng thứ hạn Wilcoxon mà có thể được sử dụng để kiểm định sự khác biệt trong phân bố của hơn hai mẫu. Khi có hai mẫu được sử dụng, nó cho kết quả giống như của kiểm định tổng thứ hạn Wilcoxon.

### Chi-squared test

### Kiểm định Chi bình phương

[Pearson’s Chi-squared test](https://en.wikipedia.org/wiki/Chi-squared_test) is used in testing for significant differences between categorical croups.

Kiểm định [Chi bình phương của Pearson](https://en.wikipedia.org/wiki/Chi-squared_test) được sử dụng trong kiểm tra sự khác biệt có ý nghĩa thống kê giữa các biến phân loại.

## **rstatix** package

## **Gói lệnh trong “rstatix** package”

The **rstatix** package offers the ability to run statistical tests and retrieve results in a “pipe-friendly” framework. The results are automatically in a data frame so that you can perform subsequent operations on the results. It is also easy to group the data being passed into the functions, so that the statistics are run for each group.

Gói lệnh trong “**rstatix** package” cho phép thực hiện các kiểm định thống kê và truy xuất kết quả “dễ sử dụng cho các tính toán tiếp theo”. Có nghĩa là kết quả xuất tự động thành một cấu trúc của bảng số liệu để có thể thực hiện các thao tác tiếp theo. Nó cũng cho phép nhóm số liệu trong các hàm để thực hiện các thống kê theo nhóm.

### Summary statistics

### Tóm tắt thống kê

The function get\_summary\_stats() is a quick way to return summary statistics. Simply pipe your dataset to this function and provide the columns to analyse. If no columns are specified, the statistics are calculated for all columns.

By default, a full range of summary statistics are returned: n, max, min, median, 25%ile, 75%ile, IQR, median absolute deviation (mad), mean, standard deviation, standard error, and a confidence interval of the mean.

You can specify a subset of summary statistics to return by providing one of the following values to type =: “full”, “common”, “robust”, “five\_number”, “mean\_sd”, “mean\_se”, “mean\_ci”, “median\_iqr”, “median\_mad”, “quantile”, “mean”, “median”, “min”, “max”.

Hàm get\_summary\_stats() là một cách thực hiện tóm tắt thống kê nhanh. Chỉ cần đưa bộ số liệu và chỉ định các cột muốn phân tích vào hàm này. Nếu không có cột nào được cụ thể, tóm tắt thống kê sẽ tính toán cho tất cả các cột.

Tóm tắt thống kê đầy đủ sẽ cho kết quả mặc định như sau: số quan sát (n), giá trị nhỏ nhất, giá trị lớn nhất, trung vị, giá trị tứ phân vị thứ nhất, giá trị tứ phân vị thứ ba, khoảng tứ phân vị, độ lệch tuyệt đối của trung vị (mad), trung bình, độ lệch chuẩn, sai số chuẩn và khoảng tin cậy của trung bình.

Có thể tóm tắt một số giá trị thông kế bằng cách cung cấp một trong số các giá trị sau đến type =: “full”, “common”, “robust”, “five\_number”, “mean\_sd”, “mean\_se”, “mean\_ci”, “median\_iqr”, “median\_mad”, “quantile”, “mean”, “median”, “min”, “max”.

It can be used with grouped data as well, such that a row is returned for each grouping-variable:

You can also use **rstatix** to conduct statistical tests:

Nó cũng có thể được sử dụng để nhóm số liệu, như một hàng cho kết quả của mỗi biến nhóm.

**Rstatix** cũng được sử dụng để thực hiện các kiểm định thống kê:

### T-test

### Kiểm định t

Use a formula syntax to specify the numeric and categorical columns:

Or use ~ 1 and specify mu = for a one-sample T-test. This can also be done by group.

If applicable, the statistical tests can be done by group, as shown below:

Sử dụng cú pháp để chỉ định cột biến liên tục và cột biến phân loại:

Hoặc sử dụng ~ 1 và ghi rõ mu = cho kiểm định t một mẫu. Cú pháp này có thể sử dụng để thực hiện cho nhóm.

Nếu có thể, các kiểm định thống kê có thể thực hiện theo nhóm, như được trình bày bên dưới.

### Shapiro-Wilk test

### Kiểm định Shapiro-Wilk

As stated above, sample size must be between 3 and 5000.

Như đã đề cập bên trên, cỡ mẫu phải nằm trong khoảng từ 3 đến 5000.

### Wilcoxon rank sum test

### Kiểm định tổng thứ hạng Wilcoxon

### Kruskal-Wallis test

### Kiểm định Kruskal-Wallis

Also known as the Mann-Whitney U test.

Cũng được biết như kiểm định Mann-Whitney U

### Chi-squared test

### Kiểm định Chi bình phương

The chi-square test function accepts a table, so first we create a cross-tabulation. There are many ways to create a cross-tabulation (see [Descriptive tables](#descriptive-tables)) but here we use tabyl() from **janitor** and remove the left-most column of value labels before passing to chisq\_test().

Many many more functions and statistical tests can be run with **rstatix** functions. See the documentation for **rstatix** [online here](https://github.com/kassambara/rstatix) or by entering ?rstatix.

Hàm kiểm định Chi bình phương chấp nhận một bảng, vì vậy đầu tiên là tạo một bảng chéo. Có nhiều cách để tạo một bảng chéo (xem chương [Các bảng mô tả số liệu](#descriptive-tables)) nhưng ở đây chúng ta sử dụng hàm tabyl() từ **janitor** và bỏ cột ngoài cùng bên trái của nhãn giá trị trước khi đưa vào hàm chisq\_test().

Có rất nhiều hàm và kiểm định thống kê có thể được thực hiện bằng các hàm trong gói lệnh **rstatix**. Xem thêm tài liệu trực tuyến về **rstatix** [ở đây](https://github.com/kassambara/rstatix) hoặc gõ lệnh ?rstatix.

## gtsummary package

## Gói lệnh gtsummary

Use **gtsummary** if you are looking to add the results of a statistical test to a pretty table that was created with this package (as described in the **gtsummary** section of the [Descriptive tables](#tbl_gt) page).

Performing statistical tests of comparison with tbl\_summary is done by adding the add\_p function to a table and specifying which test to use. It is possible to get p-values corrected for multiple testing by using the add\_q function. Run ?tbl\_summary for details.

Sử dụng gói lệnh **gtsummary** cho phép truy xuất các kết quả của một kiểm định thống kê vào một bảng phù hợp (như được mô tả tại chương [Các bảng mô tả số liệu](#tbl_gt)trong mục **gtsummary**).

Khi thực hiện các kiểm định so sánh bằng hàm tbl\_summary, dùng thêm hàm add\_p để đưa cột giá trị p và kiểm định được sử dụng vào bảng. Có thể xuất nhiều giá trị p mà được hiệu chỉnh cho nhiều kiểm định bằng cách dùng thêm hàm add\_q. Gõ lệnh ?tbl\_summary để biết thêm chi tiết.

### Chi-squared test

### Kiểm định Chi bình phương

Compare the proportions of a categorical variable in two groups. The default statistical test for add\_p() when applied to a categorical variable is to perform a chi-squared test of independence with continuity correction, but if any expected call count is below 5 then a Fisher’s exact test is used.

Được sử dụng để so sánh các tỷ lệ của một biến phân loại trong hai nhóm. Kiểm định thống kê mặc định cho biến phân loại trong hàm add\_p()là kiểm định Chi bình phương về tính độc lập với hiệu chỉnh liên tục, nhưng nếu có bất kỳ giá trị kỳ vọng nào nhỏ hơn 5 thì kiểm định chính xác của Fisher sẽ được sử dụng.

### T-tests

### Kiểm định t

Compare the difference in means for a continuous variable in two groups. For example, compare the mean age by patient outcome.

Được sử dụng để so sánh sự khác biệt về trung bình của một biến trung bình trong hai nhóm. Ví dụ như so sánh tuổi trung bình với kết cục của bệnh nhân.

### Wilcoxon rank sum test

### Kiểm định tổng thứ hạng Wilcoxon

Compare the distribution of a continuous variable in two groups. The default is to use the Wilcoxon rank sum test and the median (IQR) when comparing two groups. However for non-normally distributed data or comparing multiple groups, the Kruskal-wallis test is more appropriate.

Được dùng để so sánh sự phân bố của một biến liên tục trong hai nhóm. Kiểm định mặc định là kiểm định tổng thứ hang Wilcoxon và trung vị (khoảng tứ phân vị IQR) khi so sánh hai nhóm. Tuy nhiên, đối với số liệu không có phân bố chuẩn hoặc so sánh nhiều nhóm, kiểm định Kruskal-wallis là kiểm định thích hợp hơn.

### Kruskal-wallis test

### Kiểm định Kruskal-wallis

Compare the distribution of a continuous variable in two or more groups, regardless of whether the data is normally distributed.

Được sử dụng để so sánh sự phân bố của một biến liên tục trong hai hay nhiều nhóm, bất kể số liệu có phân bố chuẩn hay không.

## Correlations

## Tương quan

Correlation between numeric variables can be investigated using the **tidyverse**  
**corrr** package. It allows you to compute correlations using Pearson, Kendall tau or Spearman rho. The package creates a table and also has a function to automatically plot the values.

Mối tương quan giữa các biến định lượng có thể được kiển bằng cách sử dụng lệnh **corrr từ gói lệnh tidyverse. Lệnh này cũng cho phép tính các hệ số tương quan bằng phương pháp** Pearson, Kendall hoặc Spearman. Gói lệnh này tạo ra một bảng kết quả và cũng có chức năng tự động vẽ các giá trị.

## Resources

## Nguồn

Much of the information in this page is adapted from these resources and vignettes online:

[gtsummary](http://www.danieldsjoberg.com/gtsummary/articles/tbl_summary.html) [dplyr](https://dplyr.tidyverse.org/articles/grouping.html) [corrr](https://corrr.tidymodels.org/articles/using-corrr.html) [sthda correlation](http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r)

Phần lớn thông tin trong phần này được phỏng theo các nguồn sau:

[gtsummary](http://www.danieldsjoberg.com/gtsummary/articles/tbl_summary.html) [dplyr](https://dplyr.tidyverse.org/articles/grouping.html) [corrr](https://corrr.tidymodels.org/articles/using-corrr.html) [sthda correlation](http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r)