**1. Cơ sở lý thuyết**

**1.1 Thống kê là gì?**

**A/ Định nghĩa**

- Thống kê là một nhánh của toán học ứng dụng liên quan đến việc thu thập, mô tả, phân tích và suy luận các kết luận từ dữ liệu định lượng. Các lý thuyết toán học đằng sau thống kê chủ yếu dựa vào phép tính vi phân và tích phân, đại số tuyến tính và lý thuyết xác suất.

- Trên thực tế, thống kê là ý tưởng chúng ta có thể tìm hiểu về các tính chất của bộ lớn của các đối tượng hoặc các sự kiện (một quần thể) bằng cách nghiên cứu các đặc điểm của một số nhỏ các đối tượng hoặc các sự kiện (tương tự mẫu ). Bởi vì trong nhiều trường hợp việc thu thập dữ liệu toàn diện về toàn bộ dân số là quá tốn kém, khó khăn hoặc không thể thực hiện được, thống kê bắt đầu với một mẫu có thể được quan sát một cách thuận tiện hoặc có thể chi trả được.

- Các nhà thống kê, những người làm công tác thống kê, đặc biệt quan tâm đến việc xác định cách đưa ra kết luận đáng tin cậy về các nhóm lớn và các hiện tượng chung từ các đặc điểm quan sát được của các mẫu nhỏ chỉ đại diện cho một phần nhỏ của nhóm lớn hoặc một số trường hợp hạn chế của một hiện tượng chung.

- Thống kê được sử dụng hầu như trong tất cả các ngành khoa học như khoa học vật lý và xã hội, cũng như trong kinh doanh, nhân văn, chính phủ và sản xuất.

- Lợi ích của thống kê:

* Thống kê sử dụng bằng chứng số để rút ra kết luận hợp lệ
* Các nhà thống kê biết cách tránh những sai lầm thường gặp

- Hai lĩnh vực chính của thống kê là:

* Thống kê mô tả.
* Thống kê suy diễn.

**a. Thống kê mô tả**

-Mô tả thống kê là hệ số mô tả ngắn gọn rằng tóm tắt một tập dữ liệu nhất định, đó có thể là một đại diện của toàn bộ hoặc một mẫu của một dân số. Thống kê mô tả được chia thành các thước đo về xu hướng trung tâm và các thước đo về độ biến thiên (lan truyền). Các phép đo xu hướng trung tâm bao gồm giá trị trung bình, giá trị trung bình và phương thức, trong khi các phép đo độ biến thiên bao gồm độ lệch, phương sai, các biến số tối thiểu và tối đa, [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp) và [độ lệch](https://www.investopedia.com/terms/s/skewness.asp) .

- Thống kê mô tả được sử dụng để mô tả những đặc tính cơ bản của [dữ liệu](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) thu thập được từ nghiên cứu thực nghiệm qua các cách thức khác nhau. Thống kê mô tả và [thống kê suy luận](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_suy_lu%E1%BA%ADn) cùng cung cấp những tóm tắt đơn giản về [mẫu](https://vi.wikipedia.org/wiki/M%E1%BA%ABu_(th%E1%BB%91ng_k%C3%AA)) và các thước đo. Cùng với phân tích đồ họa đơn giản, chúng tạo ra nền tảng của mọi phân tích định lượng về số liệu. Để hiểu được các hiện tượng và ra quyết định đúng đắn, cần nắm được các phương pháp cơ bản của mô tả dữ liệu. Có rất nhiều kỹ thuật hay được sử dụng. Có thể phân loại các kỹ thuật này như sau:

* Biểu diễn dữ liệu bằng đồ họa trong đó các đồ thị mô tả dữ liệu hoặc giúp so sánh dữ liệu;
* Biểu diễn dữ liệu thành các bảng số liệu tóm tắt về dữ liệu;
* Thống kê tóm tắt (dưới dạng các giá trị thống kê đơn nhất) mô tả dữ liệu.

Khi tạo các trị thống kê mô tả, người ta có thể nhằm 2 mục tiêu:

* Chọn một [trị thống kê](https://vi.wikipedia.org/w/index.php?title=Tr%E1%BB%8B_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1) để chỉ ra những đơn vị có vẻ giống nhau thực ra có thể khác nhau thế nào. Các giáo trình thống kê gọi một giải pháp đáp ứng mục tiêu này là thước đo [khuynh hướng trung tâm](https://vi.wikipedia.org/w/index.php?title=Khuynh_h%C6%B0%E1%BB%9Bng_trung_t%C3%A2m&action=edit&redlink=1).
* Chọn một trị [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) khác cho thấy các đơn vị khác nhau thế nào. Loại trị thống kê này thường được gọi là một thước đo [phân tán thống kê](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%A1n_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1).

Khi tóm tắt một lượng như độ dài, cân nặng hay tuổi tác, nói chung người ta hay dùng các trị thống kê như số [trung bình cộng](https://vi.wikipedia.org/wiki/Trung_b%C3%ACnh_c%E1%BB%99ng), [trung vị](https://vi.wikipedia.org/wiki/S%E1%BB%91_trung_v%E1%BB%8B); hay trong trường hợp một phân bố đơn mốt (mode - số trung phương), người ta thường dùng [mốt](https://vi.wikipedia.org/wiki/Mode_(th%E1%BB%91ng_k%C3%AA)). Đôi khi, người ta chọn lựa những giá trị đặc thù từ [hàm phân bố tích lũy](https://vi.wikipedia.org/wiki/H%C3%A0m_ph%C3%A2n_ph%E1%BB%91i_t%C3%ADch_l%C5%A9y) gọi là các [tứ phân vị](https://vi.wikipedia.org/wiki/T%E1%BB%A9_ph%C3%A2n_v%E1%BB%8B).

Các thước đo chung nhất về mức độ phân tán của dữ liệu lượng là [phương sai](https://vi.wikipedia.org/wiki/Ph%C6%B0%C6%A1ng_sai), giá trị căn bậc 2 của nó, tức là [độ lệch chuẩn](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99_l%E1%BB%87ch_chu%E1%BA%A9n); [khoảng](https://vi.wikipedia.org/w/index.php?title=Kho%E1%BA%A3ng_(th%E1%BB%91ng_k%C3%AA)&action=edit&redlink=1); [khoảng cách giữa các tứ phân vị](https://vi.wikipedia.org/w/index.php?title=Kho%E1%BA%A3ng_c%C3%A1ch_gi%E1%BB%AFa_c%C3%A1c_t%E1%BB%A9_ph%C3%A2n_v%E1%BB%8B&action=edit&redlink=1); và [độ lệch bình quân tuyệt đối](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BB%99_l%E1%BB%87ch_b%C3%ACnh_qu%C3%A2n_tuy%E1%BB%87t_%C4%91%E1%BB%91i&action=edit&redlink=1).

Khi thực hiện một trình diễn đồ họa để tóm tắt một bộ dữ liệu, cũng có thể áp dụng cả hai mục tiêu nói trên. Một ví dụ đơn giản về kỹ thuật đồ họa là [đồ thị phân bố](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BB%93_th%E1%BB%8B_ph%C3%A2n_b%E1%BB%91&action=edit&redlink=1), thứ đồ thị phơi bày cả khuynh hướng trung tâm lẫn độ [phân tán thống kê](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%A1n_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1).

**b. Thống kê suy diễn**

- Thống kê suy diễn là công cụ mà nhà thống kê sử dụng để rút ra kết luận về các đặc điểm của một quần thể từ các đặc điểm của mẫu và quyết định mức độ chắc chắn của chúng về độ tin cậy của các kết luận đó.

- Dựa trên kích thước mẫu và sự phân bố của dữ liệu mẫu, các nhà thống kê có thể tính toán xác suất để các thống kê đo lường xu hướng trung tâm, độ biến thiên, phân bố và mối quan hệ giữa các đặc điểm trong mẫu dữ liệu, cung cấp bức tranh chính xác về các thông số tương ứng của toàn bộ tổng thể từ đó mẫu được rút ra.

- Thống kê suy luận được sử dụng để khái quát hóa về các nhóm lớn, chẳng hạn như ước tính nhu cầu trung bình đối với một sản phẩm bằng cách khảo sát mẫu thói quen mua hàng của người tiêu dùng hoặc để cố gắng dự đoán các sự kiện trong tương lai, chẳng hạn như dự đoán lợi nhuận trong tương lai của một loại chứng khoán hoặc tài sản về lợi nhuận trong một khoảng thời gian mẫu.

**B/ Tầm quan trọng của thống kê**

Ngày nay, thống kê đóng vai trò quan trọng trong nhiều lĩnh vực:

- Học qua máy và khai thác dữ liệu: có hai ứng dụng cho học qua máy móc và khai thác dữ liệu: quản lý dữ liệu và phân tích dữ liệu. Các công cụ thống kê cần thiết cho việc phân tích dữ liệu.

- Thống kê trong xã hội học: thống kê được áp dụng cho một loạt các môn học, bao gồm cả khoa học tự nhiên và xã hội, chính trị và kinh doanh. Thống kê tư vấn có thể giúp các tổ chức và công ty không có chuyên môn trả lời những thắc mắc.

- Lĩnh vực chuyên môn: Các kỹ thuật thống kê được sử dụng trong một loạt các nghiên cứu khoa học và xã hội, bao gồm: ngành sinh học, tính toán sinh học, tính toán xã hội học, hệ thống sinh học, khoa học xã hội và nghiên cứu xã hội. Những ngành này bao gồm:

* Khoa học tính toán bảo hiểm
* Ứng dụng thông tin kinh tế
* Thiên văn học
* Sinh học
* Thống kê kinh doanh
* Hóa học (phân tích dữ liệu hóa học)
* Khai thác dữ liệu (áp dụng thống kê và nhận dạng mẫu để khám phá tri thức từ dữ liệu)
* Dân số học
* Kinh tế học (phân tích thống kê các số liệu kinh tế)
* Thống kê năng lượng
* Thống kê kỹ thuật
* Khoa học nghiên cứu bệnh dịch (phân tích thống kê của bệnh)
* Địa lý và hệ thống thông tin địa lý, đặc biệt trong phân tích không gian
* Xử lý hình ảnh
* Thống kê y tế
* Thống kê về hành vi, tâm lý
* Độ bền cơ khí
* Thống kê xã hội

- Thống kê áp dụng cho toán học hay nghệ thuật

**1.2 Thực hiện thống kê trong R như thế nào?**

**A/ Giới thiệu về ngôn ngữ lập trình R**

R là một công cụ rất mạnh cho học máy, thống kê và phân tích dữ liệu. Nó là một ngôn ngữ lập trình. Ngôn ngữ R là một platform-independent do đó chúng ta có thể sử dụng nó cho bất kỳ hệ điều hành nào. Việc cài đặt R cũng miễn phì vì thế chúng ta có thể sử dụng mà không cần phải mua bản quyền.

Ngôn ngữ R là mã nguồn mở nên ai trong chúng ta cũng có thể phân tích source code để hiểu được chính xác cách R vận hành. Bất kỳ ai cũng có thể thêm tính năng và fix bug mà không cần chờ nhà phát hành ra bản vá. Đồng thời, R có thể tích hợp được với ngôn ngữ khác (C,C++). Nó cũng cho phép chúng ta tương tác với nhiều nguồn dữ liệu và các gói thống kê (SAS, SPSS). R có một cộng đồng phát triển mạnh mẽ.

**B/ Giới thiệu về thư viện ggplot2**

ggplot2 là một package hỗ trợ visualization rất mạnh trong R. Dựa trên package này ta có thể vẽ được các đồ thị dạng barchart, line, plot, density, candle chart,pie,… và rất nhiều các đồ thị khác.Ngoài ra ggplot2 còn cho phép người dùng tùy chỉnh màu sắc, kích cỡ, theme, title, … để đồ thị được đẹp hơn. Cấu trúc của ggplot2 được chia rõ ràng làm 2 phần chính.

* ggplot(): phần này qui định đồ thị sẽ sử dụng data nào làm đầu vào. Lưu ý data phải có dạng data.frame. Dạng vector sẽ không được support.
* geom\_(aes(x,y)): Phần này qui định kiểu đồ thị và các trục tọa độ từ dữ liệu đầu vào. Nếu chỉ có ggplot() mà không thêm geom\_() thì chúng ta chỉ nhận được background mà không có đồ thị mặc dù data đã được khai báo. Trong geom\_() chúng ta phải khai báo thêm trục tọa độ vào các arguments x và y của aes() chẳng hạn như geom\_point(aes(x=bienx,y=bieny)).

Các kiểu đồ thị chính: geom\_line(): biển diễn line geom\_point(): biểu diễn point geom\_chart(): biểu diễn chart geom\_density(): biểu diễn dưới dạng density geom\_vertical(): thêm trục vertical geom\_abline(): thêm trục horizontal, vertical và diagonal geom\_qq(): quantile and quantile plot geom\_contour(): vẽ các đường đồng mức 2d và 3d geom\_label() geom\_text(): hiển thị text geom\_raster(), geom\_tile(), geom\_rect(): biểu diễn dạng màu sắc mật độ.

**2. Cơ sở dữ liệu và tài liệu tham khảo**

- Cơ sở dữ liệu: Cổng Light Novel- Trang đọc Light Novel lớn nhất Việt Nam

Đường link: <https://ln.hako.re/>

-Nguồn tham khảo:

<https://www.statology.org/ggplot-pie-chart/>

<http://r-statistics.co/ggplot2-Tutorial-With-R.html>

<https://www.statology.org/best-ggplot2-themes/>

<https://rpubs.com/phamdinhkhanh/381826>

**3. Mô tả về dữ liệu sử dụng crawl lấy về**

**a. Số lượng dòng và cột dữ liệu:**

- 8 cột x 1934 hàng

**b. Danh sách các trường dữ liệu (tên cột), kiểu dữ liệu (số, văn bản…)**

- Các trường dữ liệu: artist, author, gernes, heart, name, star, view, words.

- Kiểu dữ liệu:

+ Kiểu văn bản: artist, author, gernes, name

+ Kiểu số: heart, star, view, words

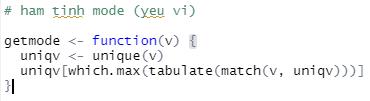


*Hình 1: 20 dòng đầu tiên của dữ liệu.*

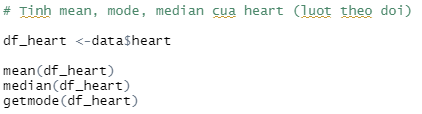
**3. Phân tích thống kê mô tả số liệu (sử dụng R để tính toán và vẽ biểu đồ)**

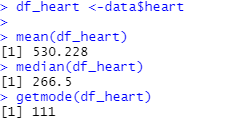
**a. Phân tích thống kê mô tả của số liệu:**

\* Tạo một hàm để tính mode (yếu vị):

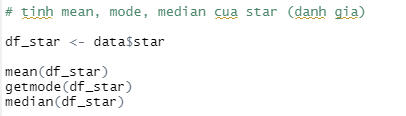


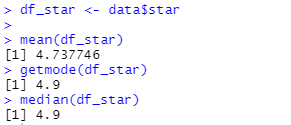
- Tính mean, mode, median của Heart (lượt theo dõi)



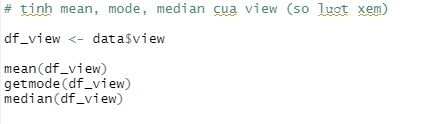


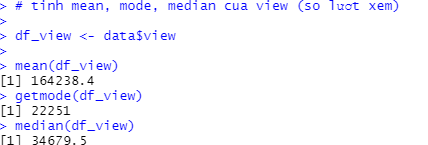
**- Tính mean, mode, median của star (đánh giá)**



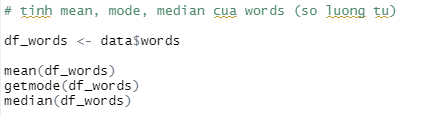


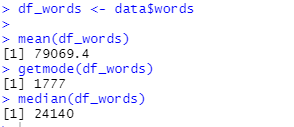
- Tính mean, mode, median của view (số lượt xem):



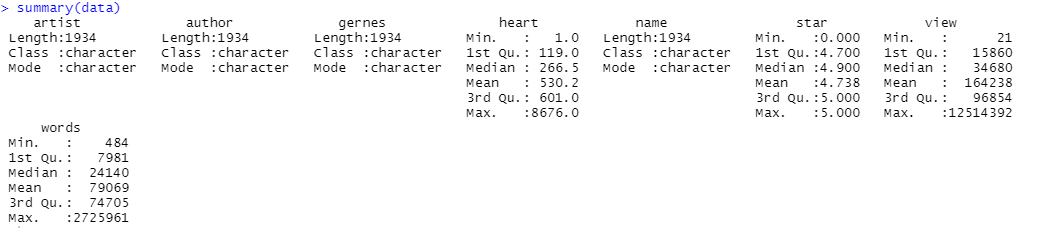


**- Tính mean, mode, median của words (số lượng chữ)**





\* Tổng quan về dữ liệu (sử dụng hàm summary)



b. Vẽ biểu đồ