ĐẠI HỌC ĐÀ NẴNG

TRƯỜNG ĐẠI HỌC BÁCH KHOA

CÁC CÔNG CỤ TOÁN HỌC TRONG KHDL

Chương 2

Khoa Công nghệ thông tin
TS. Phạm Công Thắng

Tài liệu tham khảo

- Jake VanderPlas, Python Data Science Handbook, O'Reilly Media, Inc., 548 pages, 2016.
- Peter Bruce, Andrew Bruce, Practical Statistics for Data Scientists: 50 Essential Concepts - 1st Edition, O'Reilly Media; 1st edition, 90 pages, 2017
- Andreas C. Müller, Sarah Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition, 392 pages, 2016
- Lillian Pierson, Data Science For Dummies - 2nd Edition, John Wiley & Sons Inc., 385 pages, 2017.
- David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Statistics for Business and Economics, South-Western College Pub., 880 pages, 2010.
- Các nguồn internet khác.

Nội dung

- Phân tích dữ liệu khám phá (Exploratory Data Analysis, EDA)
- Phân bố dữ liệu và phân bố lấy mẫu (Data and Sampling Distributions)
- Kiểm định giả thiết
- Suy diễn Bayes

Nội dung

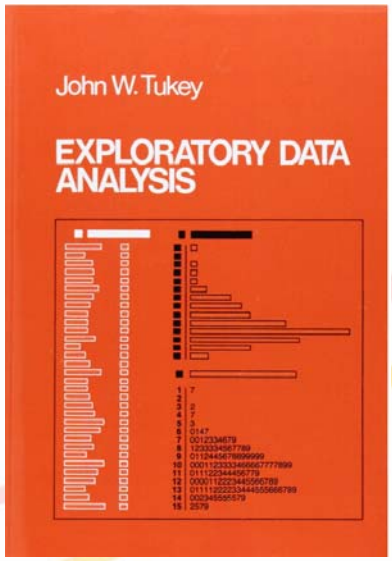
- ***Phân tích dữ liệu khám phá (Exploratory Data Analysis, EDA)***
- Dữ liệu và phân bố lấy mẫu (Data and Sampling Distributions)
- Kiểm định giả thiết
- Suy diễn Bayes

EDA

- EDA được John Tukey thúc đẩy bắt đầu từ những năm 1960.

(1915 – 2000)

- **Author: John Tukey**
- **Publisher :** Pearson; 1st edition (January 1, 1977)
- **Language: :** English
- **Paperback :** 712 pages



EDA

- Trong thống kê, EDA là một cách tiếp cận để phân tích các tập dữ liệu để:
  - Tóm tắt các đặc điểm chính của tập dữ liệu, thường bằng các phương pháp trực quan.
  - Mô hình thống kê có thể được sử dụng hoặc không, nhưng vấn đề của EDA là để xem dữ liệu có thể cho chúng ta biết gì ngoài nhiệm vụ mô hình hóa chuẩn hoặc kiểm định giả thuyết.

EDA

- EDA là bước rất quan trong trong bất kỳ quá trình phân tích dữ liệu.
  - Phân tích bộ dữ liệu dựa trên các phương pháp số và công cụ đồ họa trực quan khác nhau như bar graphs, pie charts, histograms, time plots, scatterplots…
  - Khám phá dữ liệu về các mẫu, xu hướng, cấu trúc cơ bản, độ lệch so với xu hướng, sự bất thường và cấu trúc lạ.

EDA

- Tối đa hóa thông tin chi tiết về tập dữ liệu.
- Khám phá cấu trúc cơ bản.
- Trích xuất các biến quan trọng.
- Phát hiện những điểm bất thường và bất thường.
- Kiểm tra các giả định cơ bản.
- Phát triển các mô hình phù hợp.
- Xác định cài đặt hệ số tối ưu

EDA

- Các bước của EDA:
  - Tạo các yêu cầu cần nghiên cứu về dữ liệu
  - Tái cấu trúc dữ liệu: có thể cần tạo các biến mới từ các biến hiện có, thay vì sử dụng hai biến, thu thập tỷ lệ hoặc tỷ lệ phần trăm của chúng
  - Dựa trên các yêu cầu nghiên cứu, sử dụng các công cụ đồ họa thích hợp và thu được số liệu thống kê mô tả: cố gắng hiểu cấu trúc dữ liệu, các mối quan hệ, sự bất thường, các hiện tượng không mong muốn.
  - Cố gắng xác định các biến gây nhiễu, quan hệ tương tác và đa cộng tuyến (multicollinearity) nếu có.
  - Xử lý các đối tượng quan sát được bị thiếu
  - Quyết định sự cần thiết về chuyển đổi các biến.
  - Quyết định giả thuyết dựa trên các yêu cầu nghiên cứu đặt ra.

EDA

- Sau khi thực hiên EDA:
  - Phân tích dữ liệu xác nhận (Confirmatory Data Analysis): Xác minh giả thuyết bằng phân tích thống kê.
  - Đưa ra kết luận và trình bày kết quả của một cách rõ ràng.

EDA – Phân loại

- Phân tích dữ liệu thăm dò thường được phân loại chéo theo hai cách.
  - Phương pháp là phi đồ họa hoặc sử dụng đồ họa.
  - Phương pháp là đơn biến hoặc đa biến (thường chỉ là hai biến).

EDA – Phân loại

- Các phương pháp phi đồ họa thường liên quan đến việc tính toán các thống kê tổng hợp.
- Các phương pháp đồ họa rõ ràng là tóm tắt dữ liệu theo các sơ đồ, biểu đồ hoặc hình ảnh.
- Các phương pháp đơn biến xem xét một biến (cột dữ liệu) tại một thời điểm, trong khi các phương pháp đa biến xem xét hai hoặc nhiều biến cùng một lúc để khám phá các mối quan hệ.
    - Thông thường, EDA đa biến sẽ là lưỡng biến (chính xác hai biến), nhưng đôi khi nó sẽ liên quan đến ba hoặc nhiều biến.
        - Cần thực hiện EDA đơn biến trên từng thành phần của EDA đa biến trước khi thực hiện EDA đa biến

Các yếu tố của dữ liệu có cấu trúc (Elements of Structured Data)

- Có hai loại dữ liệu có cấu trúc cơ bản:
  - Số (numeric)
  - Dữ liệu phân loại theo phạm trù hay danh mục (categorical)

Các yếu tố của dữ liệu có cấu trúc

- Dữ liệu dạng số có hai loại:
  - Liên tục: tốc độ gió hoặc khoảng thời gian…
  - Rời rạc: chẳng hạn như số lần xuất hiện của một sự kiện…
- Dữ liệu phân loại theo phạm trù chỉ chiếm một tập hợp cố định các giá trị, chẳng hạn như loại màn hình TV (plasma, LCD, LED,…)
  - Dữ liệu nhị phân là một trường hợp đặc biệt quan trọng của dữ liệu phân loại theo phạm trù chỉ nhận một trong hai giá trị, chẳng hạn như **0/1, yes / no** hoặc **true / false**.
  - Một loại đặc biệt của dữ liệu phân loại theo phạm trù là dữ liệu thứ có tự trong đó các danh mục được sắp xếp theo thứ tự.

Rectangular Data

- Dữ liệu hình chữ nhật
  - Đối tượng dữ liệu hình chữ nhật, giống như bảng tính hoặc bảng cơ sở dữ liệu
  - Một ma trận 2 chiều với các hàng biểu thị các bản ghi (các trường hợp) và các cột biểu thị các đặc trưng (biến).

Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	1
Automotive	US	3115	7	Tue	0.01	0.01	1

Estimates of Location

- Các biến có dữ liệu được đo hoặc đếm có thể có hàng nghìn giá trị khác nhau.
  - Bước cơ bản trong việc khám phá dữ liệu là nhận được "giá trị điển hình" cho mỗi đặc trưng (biến): ước tính vị trí của hầu hết dữ liệu (tức là xu hướng trung tâm của chúng).
  - Một số các đánh giá ước lượng:
    - Mean
    - Trimmed Mean
    - Weighted Mean…

Estimates of Location

- Mean: Ước tính cơ bản nhất về vị trí là giá trị trung bình hoặc giá trị trung bình

$$\text{Mean} = \bar{x} = \frac{\sum_i^N x_i}{N}$$

Estimates of Location

- Trimmed Mean:
  - Một biến thể của giá trị trung bình, tính giá trị trung bình đã được rút gọn.
  - Được tính bằng cách bỏ một số giá trị được sắp xếp cố định ở mỗi đầu và sau đó lấy giá trị trung bình của các giá trị còn lại.
    - Biểu diễn các giá trị được sắp xếp theo x1,x2,..xN trong đó x1 là giá trị nhỏ nhất và xN, công thức tính Trimmed Mean với p giá trị nhỏ nhất và lớn nhất bị bỏ qua là

$$\text{Trimmed Mean} = \bar{x} = \frac{\sum_{i=p+1}^{N-p} x_{(i)}}{N - 2p}$$

Estimates of Location

- Weighted Mean
  - Giá trị trung bình có trọng số, được tính bằng cách nhân mỗi giá trị dữ liệu xi bằng một trọng lượng wi và chia tổng của các trọng số.

$$\text{Weighted Mean} = \bar{x}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i}^{N} w_i}$$

Standard Deviation and Related Estimates

- Mean absolute deviation

$$\text{Mean Absolution Deviation} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

- Phương sai là giá trị trung bình của bình phương độ lệch
- Độ lệch chuẩn là căn bậc hai của phương sai.

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{N - 1}$$

$$\text{Standard Deviation} = s = \sqrt{\text{Variance}}$$

Estimates Based on Percentiles

- Một cách tiếp cận để ước tính độ phân bố dữ liệu dựa trên việc xem xét mức độ trải rộng của dữ liệu đã được sắp xếp. Thống kê dựa trên dữ liệu được sắp xếp (xếp hạng) được gọi là thống kê có thứ tự (*order statistics)*.
  - Thước đo cơ bản nhất là phạm vi (*range)*: hiệu số giữa giá trị lớn nhất và nhỏ nhất.
  - Các giá trị tối thiểu và tối đa đều hữu ích trong việc xác định các giá trị ngoại lệ.

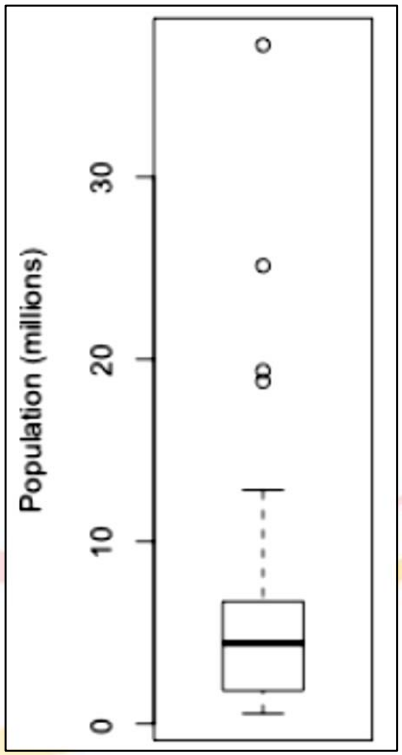
Estimates Based on Percentiles

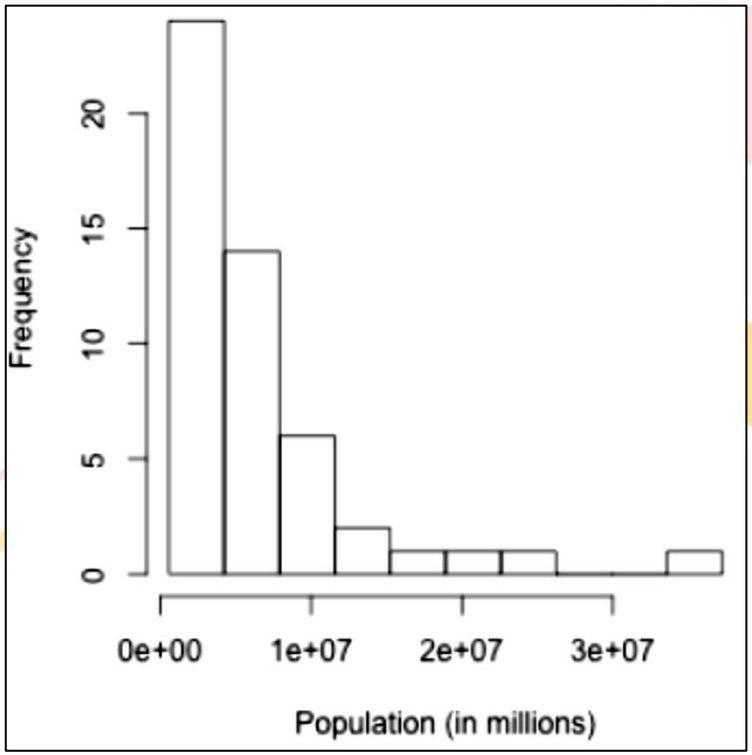
- Để tránh ‘nhạy cảm’ với các giá trị ngoại lệ
  - Chúng ta có thể xem phạm vi dữ liệu sau khi loại bỏ các giá trị từ mỗi đầu.
  - Các loại ước tính này dựa trên sự khác biệt giữa các **tỷ lệ phần trăm**.
  - Trong tập dữ liệu, phần trăm thứ P là giá trị sao cho ít nhất P phần trăm giá trị nhận giá trị này trở xuống và ít nhất (100-P) phần trăm giá trị nhận giá trị này trở lên.

Exploring the Data Distribution

- Mỗi ước tính tổng hợp dữ liệu trong một số duy nhất để mô tả vị trí hoặc sự thay đổi của dữ liệu: rất hữu ích để khám phá cách dữ liệu được phân phối tổng thể.
    - Percentiles
    - Boxplots
    - Histograms

5%	25%	50%	75%	95%
1.60	2.42	4.00	5.55	6.51

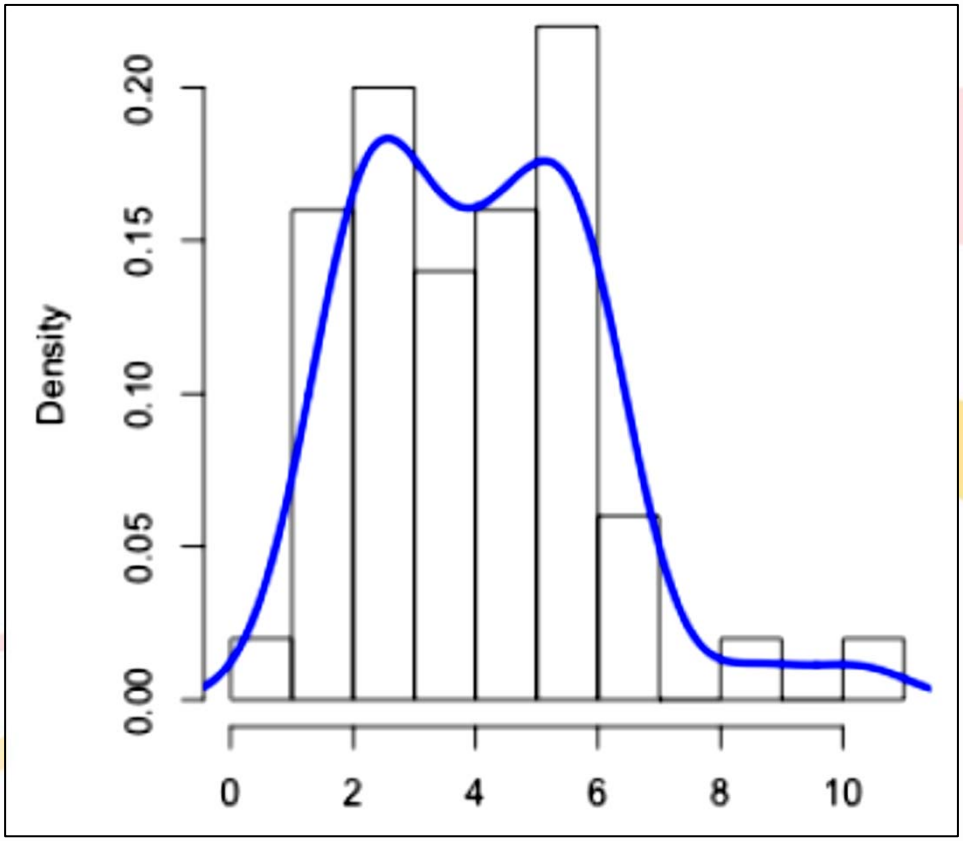




Exploring the Data Distribution

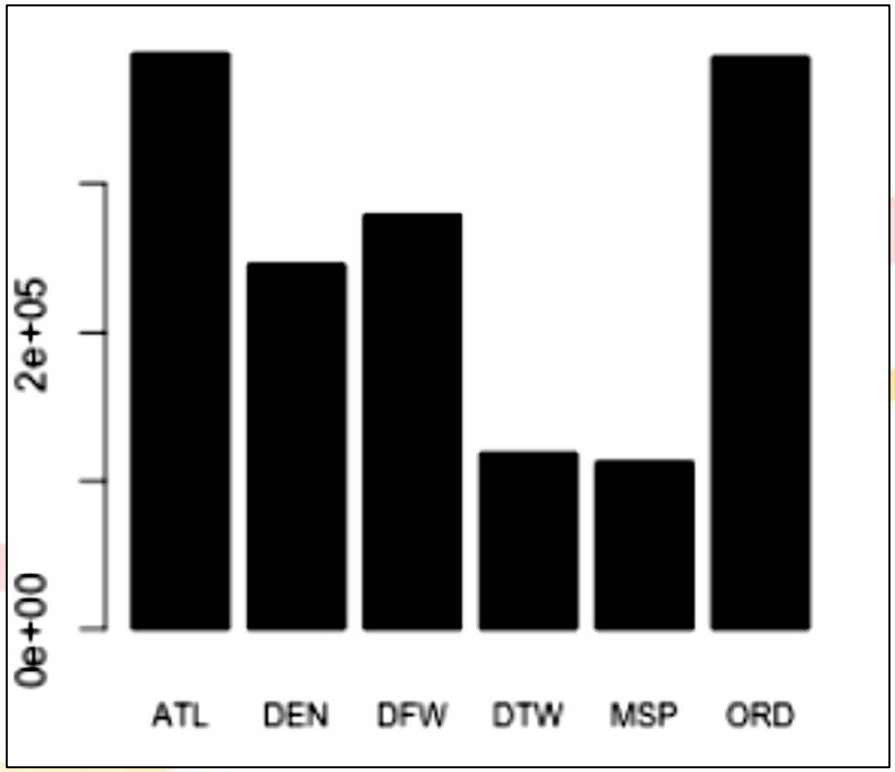
- Frequency Table (bảng tần xuất): chia phạm vi biến thành các đoạn cách đều nhau và chỉ rõ rằng có bao nhiêu giá trị rơi vào mỗi đoạn.
- Density Estimates: Phân phối của giá trị dữ liệu là đường liên tục. Nó có thể được coi là một biểu đồ được làm mịn (smoothed histogram), dù được tính toán trực tiếp từ dữ liệu bằng cách sử dụng ước tính mật độ kernal

BinNumber	BinRange	Count	States
1	563,626-4,232,658	24	WY,VT,ND
2	4,232,659-7,901,691	14	KY,LA,SC,
3	7,901,692-11,570,724	6	VA,NJ,NC,
4	11,570,725-15,239,757	2	PA,IL
5	15,239,758-18,908,790	1	FL
6	18,908,791-22,577,823	1	NY
7	22,577,824-26,246,856	1	TX
8	26,246,857-29,915,889	0	
9	29,915,890-33,584,922	0	
10	33,584,923-37,253,956	1	CA



Exploring Binary and Categorical Data

- Bar charts (biểu đồ cột) là một công cụ trực quan phổ biến để hiển thị một biến theo phạm trù duy nhất, thường thấy trên báo chí phổ biến.
  - Các biến theo phạm trù được liệt kê trên trục x và tần số hoặc tỷ lệ trên trục y
  - Có thể dùng biểu đồ hình tròn (pie charts)



Expected Value

- Một loại đặc biệt của dữ liệu theo phạm trù là dữ liệu trong đó các danh mục (phạm trù) đại diện hoặc có thể được ánh xạ tới các giá trị rời rạc trên cùng một tỷ lệ.
  - Ví dụ: một nhà tiếp thị cung cấp hai mức dịch vụ, một mức giá 300$/tháng và một mức 50$/ tháng.
    - Nhà tiếp thị cung cấp hội thảo trên webinars để tạo khách hàng tiềm năng và công ty thống kê rằng 5% người tham dự sẽ đăng ký dịch vụ $ 300, 15% cho dịch vụ $ 50 và 80% sẽ không đăng ký bất kỳ thứ gì.
    - Những dữ liệu này có thể được tổng hợp, vì mục đích tài chính, trong một ***"giá trị kỳ vọng" (expected value)*** duy nhất, là một dạng giá trị trung bình có trọng số ***trong đó trọng số là xác suất***.

Expected Value

- Cách tính giá trị kỳ vọng:
  - Nhân mỗi kết quả với xác suất xảy ra của nó.
  - Tính tổng các giá trị này.
    - Ví dụ: giá trị kỳ vọng của một người tham dự hội thảo trên webinars là 22,50 $/tháng:

$$EV = 0.05*300 + 0.15*50 + 0.*800 = 22.\ 5$$

- Giá trị kỳ vọng thực sự là một dạng của giá trị trung bình có trọng số (weighted mean):
  - Bổ sung các ý tưởng về kỳ vọng trong tương lai và trọng số xác suất, thường dựa trên phán đoán chủ quan.
  - Giá trị kỳ vọng là một khái niệm cơ bản trong định giá doanh nghiệp và lập ngân sách vốn: giá trị lợi nhuận kỳ vọng trong 5 năm từ một thương vụ mua lại mới hoặc tiết kiệm chi phí dự kiến…

Correlation

- Phân tích dữ liệu khám phá trong nhiều dự án mô hình hóa (cho dù trong khoa học dữ liệu hay trong nghiên cứu)
  - Liên quan đến việc kiểm tra mối tương quan giữa các yếu tố dự báo, giữa các yếu tố dự báo và một biến mục tiêu.
  - Các biến X và Y (mỗi biến có dữ liệu đo được) được cho là có tương quan thuận nếu giá trị cao của X đi với giá trị cao của Y và giá trị thấp của X đi với giá trị thấp của Y.
  - Nếu giá trị cao của X đi với giá trị thấp của Y và ngược lại, các biến có tương quan nghịch.

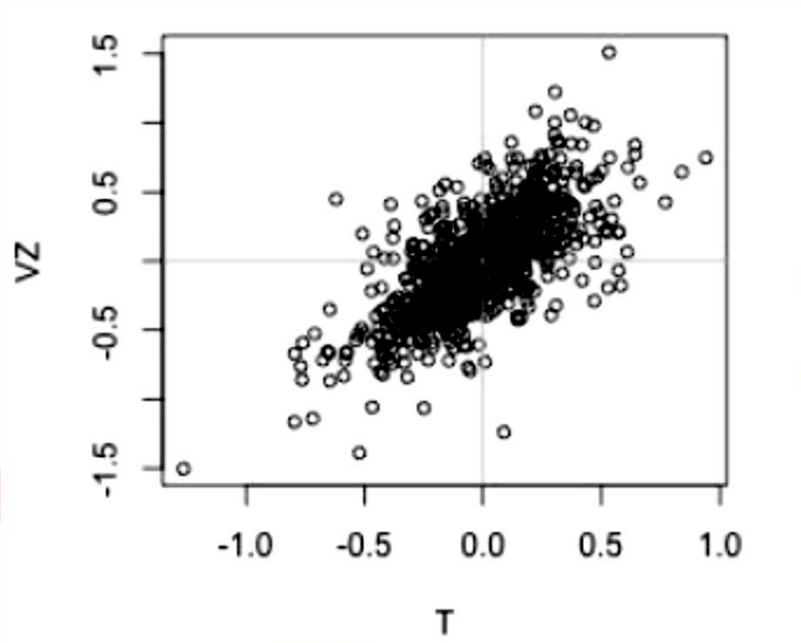
Correlation

- Hệ số tương quan (*correlation coefficient)* đưa ra ước tính về mối tương quan giữa hai biến luôn nằm trên cùng một thang đo.
- Hệ số tương quan của Pearson (*Pearson's correlation coefficient)* được tính bằng cách nhân độ lệch từ giá trị trung bình cho biến 1 với biến 2 và chia cho tích của độ lệch chuẩn ($s_x s_y$):

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Scatterplots

- Cách chuẩn để hình dung mối quan hệ giữa hai biến dữ liệu được đo lường là sử dụng biểu đồ phân tán hay phân bố (Scatterplots)
  - Trục x đại diện cho một biến
  - Trục y là biến khác
  - Mỗi điểm trên biểu đồ là một bản ghi.

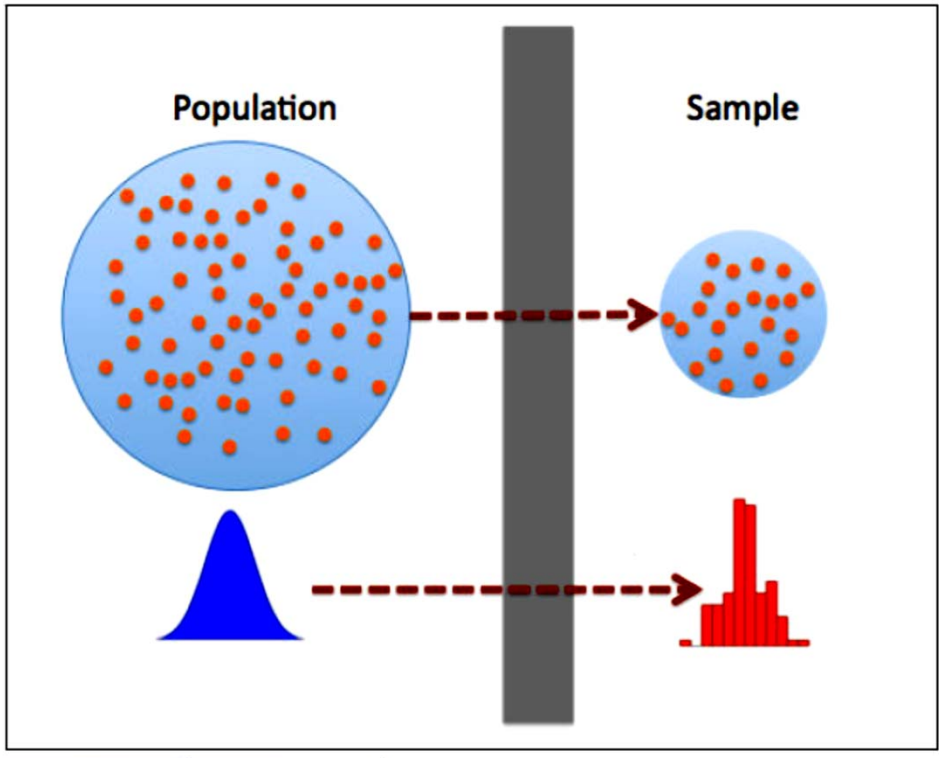


Nội dung

- Phân tích dữ liệu khám phá (Exploratory Data Analysis, EDA)
- ***Phân bố dữ liệu và phân bố lấy mẫu (Data and Sampling Distributions)***
- Kiểm định giả thiết
- Suy diễn Bayes

Data and Sampling Distributions

- Sự gia tăng của dữ liệu có chất lượng và mức độ liên quan khác nhau củng cố nhu cầu lấy mẫu như một công cụ để làm việc hiệu quả với nhiều loại dữ liệu và giảm thiểu sự sai lệch.
- Trong phân tích dữ liệu lớn, các mô hình dự đoán thường được phát triển và thử nghiệm với các mẫu.



Random sampling and sample bias

- **Random sampling**:
  - Một quá trình trong đó mỗi thành phần sẵn có của tập dữ liệu được lấy mẫu đều có cơ hội được chọn lấy mẫu như nhau ở mỗi lần lấy.
  - Việc lấy mẫu có thể được thực hiện với sự thay thế, trong đó các thành phần quan sát được đưa trở lại tập dữ liệu sau mỗi lần lấy để có thể lựa chọn lại trong tương lai.
    - có thể thực hiện mà không cần thay thế, các quan sát một khi đã được chọn, không thể sử dụng cho các lần lấy mẫu tiếp theo

Random sampling and sample bias

- Bias
  - Đề cập đến các lỗi đo lường hoặc lấy mẫu có hệ thống và được tạo ra bởi quá trình đo lường hoặc lấy mẫu.
  - Có nhiều dạng khác nhau và có thể quan sát được hoặc không nhìn thấy được.
    - Khi một kết quả dẫn đến sự sai lệch (bằng cách tham chiếu đến điểm chuẩn hoặc các giá trị thực tế), thường là *một chỉ báo cho thấy mô hình thống kê hoặc học máy đã bị xác định sai hoặc một biến quan trọng đã bị bỏ sót.*

Random sampling and sample bias

- Selection bias
  - Đề cập đến việc kinh nghiệm lựa chọn dữ liệu một cách có chọn lọc - một cách có ý thức hoặc một cách vô thức sẽ dẫn đến một kết luận sai lệch.
    - Với một giả thuyết, việc tiến hành thực nghiệm được thiết kế tốt để kiểm tra nó, thì có thể nhận được độ tin cậy cao về kết luận: ***điều này thường không xảy ra***.
    - Thông thường cần quan sát dữ liệu có sẵn và phân biệt các mẫu. Nhưng liệu mô hình này có thật hay chỉ là sản phẩm của việc "**rình mò dữ liệu**", hay săn lùng dữ liệu trên diện rộng cho đến khi điều gì đó thú vị xuất hiện?
  - Các dạng Selection bias điển hình trong thống kê
    - Lấy mẫu không ngẫu nhiên
    - Lựa chọn khoảng thời gian làm nổi bật hiệu ứng thống kê từng phần và dừng thử nghiệm khi kết quả trông "thú vị" hơn.

Sampling Distribution

- Một mẫu được trích xuất với mục tiêu đo lường thứ gì đó (với thống kê mẫu) hoặc mô hình hóa thứ gì đó (với mô hình thống kê hoặc máy học).
  - Ước tính hoặc mô hình dựa trên một mẫu có thể bị nhầm lẫn và có thể khác nhau nếu trích xuất một mẫu khác.
  - Vấn đề cần quan tâm là nó có thể khác nhau như thế nào: mối quan tâm chính là sự thay đổi lấy mẫu.
  - Nếu có nhiều dữ liệu, có thể trích xuất các mẫu bổ sung và quan sát sự phân bố của thống kê mẫu một cách trực tiếp.
  - Thông thường, cần tính toán ước lượng hoặc mô hình bằng cách sử dụng càng nhiều dữ liệu càng thuận lợi, vì vậy tùy chọn trích xuất thêm mẫu từ tập dữ liệu là không khả dụng.
  - Sự phân bố của một thống kê mẫu như giá trị trung bình có thể khả dụng hơn và có hình dạng tốt hơn so với sự phân bố của chính dữ liệu.

Sampling distributions

- Lấy mẫu từ một tập dữ liệu hữu hạn (Finite Population)
- Lấy mẫu từ một tập dữ liệu vô hạn (Infinite Population)

Sampling distributions

- Lấy mẫu từ một tập dữ liệu hữu hạn
  - Một mẫu ngẫu nhiên đơn giản cỡ n từ một tập hợp hữu hạn cỡ N là mẫu được chọn sao cho mỗi mẫu có thể có cỡ n đều có xác suất được chọn như nhau.
  - Việc thay thế từng phần tử được lấy mẫu trước khi chọn các phần tử tiếp theo được gọi là lấy mẫu có thay thế.
    - Lấy mẫu mà không cần thay thế là quy trình được sử dụng thường xuyên nhất.
    - Trong các dự án lấy mẫu lớn, các số ngẫu nhiên do máy tính tạo ra thường được sử dụng để tự động hóa quá trình chọn mẫu.

Sampling distributions

- Lấy mẫu từ một tập dữ liệu vô hạn
  - Cần phải chọn một mẫu ngẫu nhiên để đưa ra các suy luận thống kê hợp lệ về tập dữ liệu mà từ đó mẫu được lấy.
  - Mẫu ngẫu nhiên được chọn sao cho thỏa mãn các điều kiện:
    - Mỗi phần tử được chọn đến từ tập hợp quan tâm (the population of interest).
    - Mỗi phần tử được chọn độc lập.

Sampling distributions

- Dữ liệu mẫu cung cấp giá trị cho giá trị trung bình mẫu $\bar{x}$
- Trung bình của tập dữ liệu with mean μ
  - Giá trị kỳ vong của $\bar{x}$

$$E(\bar{x}) = \mu$$

  - *Khi giá trị kỳ vọng của ước lượng bằng với tham số tập dữ liệu, thì ước lượng là không chệch.*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \text{ has a normal distribution}$$

$$\text{with mean } \mu_{\bar{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$$

$$\text{and variance } \sigma^2_{\bar{X}} = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2}$$

Sampling distributions

- Standard Deviation of $\bar{x}$

$$\sigma_x = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_x = \frac{\sigma}{\sqrt{n}}$$

tập dữ liệu hữu hạn tập dữ liệu vô hạn

- Một tập hợp hữu hạn được coi là vô hạn nếu n/N <0,05.
- Khi tổng thể dữ liệu có phân phối chuẩn, phân phối lấy mẫu $\bar{x}$ là phân phối chuẩn cho bất kỳ cỡ mẫu nào.
- Phân phối lấy mẫu $\bar{x}$ có thể được xấp xỉ bằng phân phối chuẩn bất cứ khi nào mẫu có kích thước từ 30 trở lên (nếu có độ lệch lớn hoặc có sự khác biệt, có thể cần các mẫu cỡ 50).

Central Limit Theorem

- Khi tập hợp mà chúng ta đang chọn một mẫu ngẫu nhiên không có phân phối chuẩn, thì định lý giới hạn trung tâm (Central Limit Theorem) sẽ hữu ích trong việc xác định hình dạng của phân phối lấy mẫu.
- Khi chọn mẫu ngẫu nhiên có kích thước n từ một tập hợp, phân phối lấy mẫu của trung bình mẫu có thể được xấp xỉ bằng phân phối chuẩn khi kích thước mẫu trở nên lớn.
  - **Standardized z-score:** mức độ cao hơn hoặc thấp hơn trung bình mẫu so với trung bình tổng thể tập dữ liệu và được tính bằng đơn vị sai số chuẩn.

$$z = \frac{\text{sample mean} - \mu}{\text{standard error}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Central Limit Theorem

- Kích thước mẫu càng lớn thì phân phối mẫu có nghĩa càng chuẩn.
- Là trọng tâm của khái niệm suy luận thống kê vì nó cho phép chúng ta đưa ra kết luận về tập dữ liệu dựa trên dữ liệu mẫu một cách chặt chẽ mà không cần có hiểu biết về sự phân bố của tập dữ liệu cơ bản.

Sample Proportions

- Phân phối lấy mẫu của $\bar{p}$ là phân phối xác suất của tất cả các giá trị có thể có của tỷ lệ mẫu sample proportion p.
- Trung bình của tập dữ liệu with mean μ
  - Giá trị kỳ vong của $\bar{x}$

$$E(\bar{p}) = p$$

  - *p là tỷ lệ tổng thể tập dữ liệu,*

$$P = \frac{X}{N}$$

P = population proportion

X = count of successes

N = size of the population

Sample Proportions

- Standard Deviation of $\bar{p}$ (standard error of the proportion)

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

tập dữ liệu hữu hạn tập dữ liệu vô hạn

- Sự phân bố lấy mẫu của $\bar{p}$ có thể là gần đúng theo phân phối chuẩn bất cứ khi nào kích thước mẫu đủ lớn để thỏa mãn hai điều kiện:

$$np \geq 5; \;\; n(1-p) \geq 5$$

vì khi các điều kiện này được thỏa mãn, phân phối xác suất của x trong tỷ lệ mẫu $\bar{p} = \mathrm{x}/\mathrm{n}$, có thể được tính gần đúng bằng phân phối chuẩn (n là hằng số).

Sample Proportions

- Standardized sample proportion:

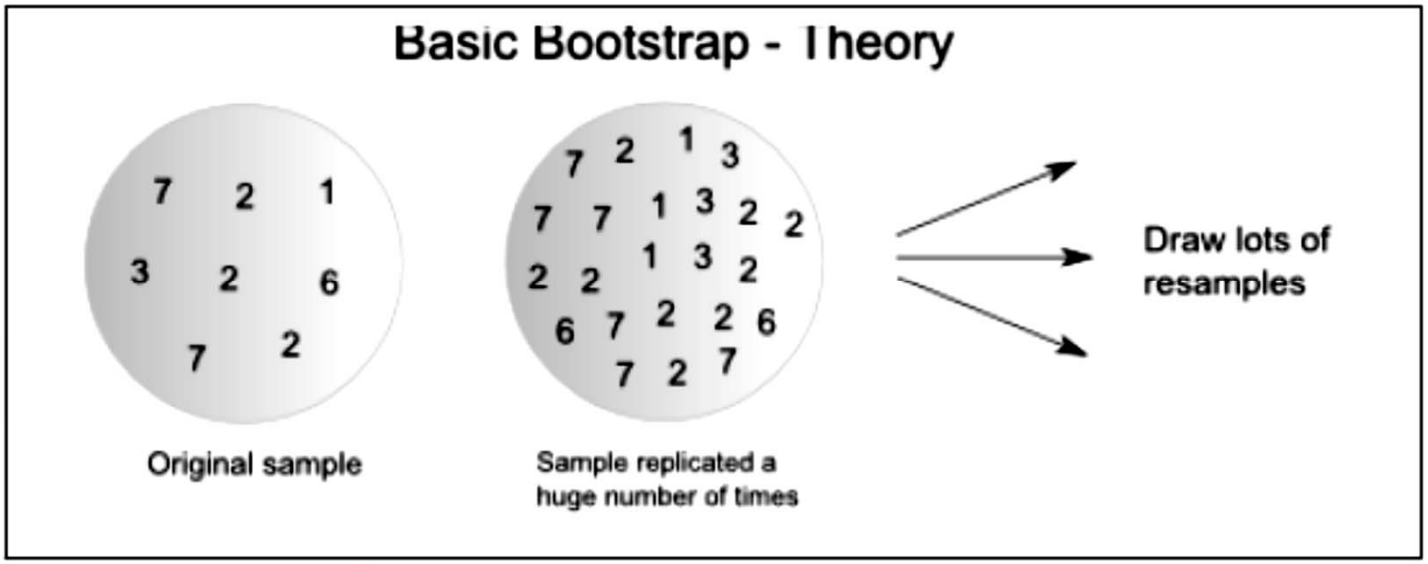
$$z = \frac{\text{sample proportion} - p}{\text{standard error}}$$

The bootstrap

- Một cách dễ dàng và hiệu quả để ước tính phân phối lấy mẫu của một thống kê hoặc của các tham số mô hình là trích xuất các mẫu bổ sung với sự thay thế từ chính mẫu đó và tính toán lại thống kê hoặc mô hình cho mỗi mẫu lấy lại, thủ tục này được gọi là ***bootstrap***
  - không nhất thiết phải liên quan đến bất kỳ giả định nào về dữ liệu, hoặc thống kê mẫu.

The bootstrap

- Bootstrap giống như sao chép mẫu ban đầu hàng nghìn hoặc hàng triệu lần để:
  - có một tập hợp giả định bao gồm tất cả hiểu biết từ mẫu ban đầu
  - có thể lấy mẫu từ tổng thể giả định này với mục đích ước tính phân phối lấy mẫu.



The bootstrap

- Trong thực tế, không nhất thiết phải lặp lại mẫu nhiều lần: chỉ đơn giản là thay thế mỗi quan sát sau mỗi lần trích xuất (lấy mẫu bằng thay thế)
  - Điều này sẽ tạo ra một tập hợp vô hạn một cách hiệu quả, trong đó xác suất của một phần tử được rút ra không thay đổi
  - Thuật toán cho việc lấy mẫu lại giá trị trung bình của **bootstrap (**mẫu có kích thước N):
    - Trích rút một giá trị mẫu, ghi lại, thay thế nó
    - Lặp lại N lần
    - Ghi lại giá trị trung bình của N giá trị được lấy mẫu lại
    - Lặp lại các bước trên với M lần (M là số lần lặp lại của bootstrap)
    - Sử dụng kết quả sau M lần lặp để:
      - a. Tính độ lệch chuẩn của chúng (đối với mẫu ước tính này có nghĩa là lỗi chuẩn)
      - b. Tạo biểu đồ histogram hoặc biểu đồ boxplot
      - c. Tìm khoảng tin cậy

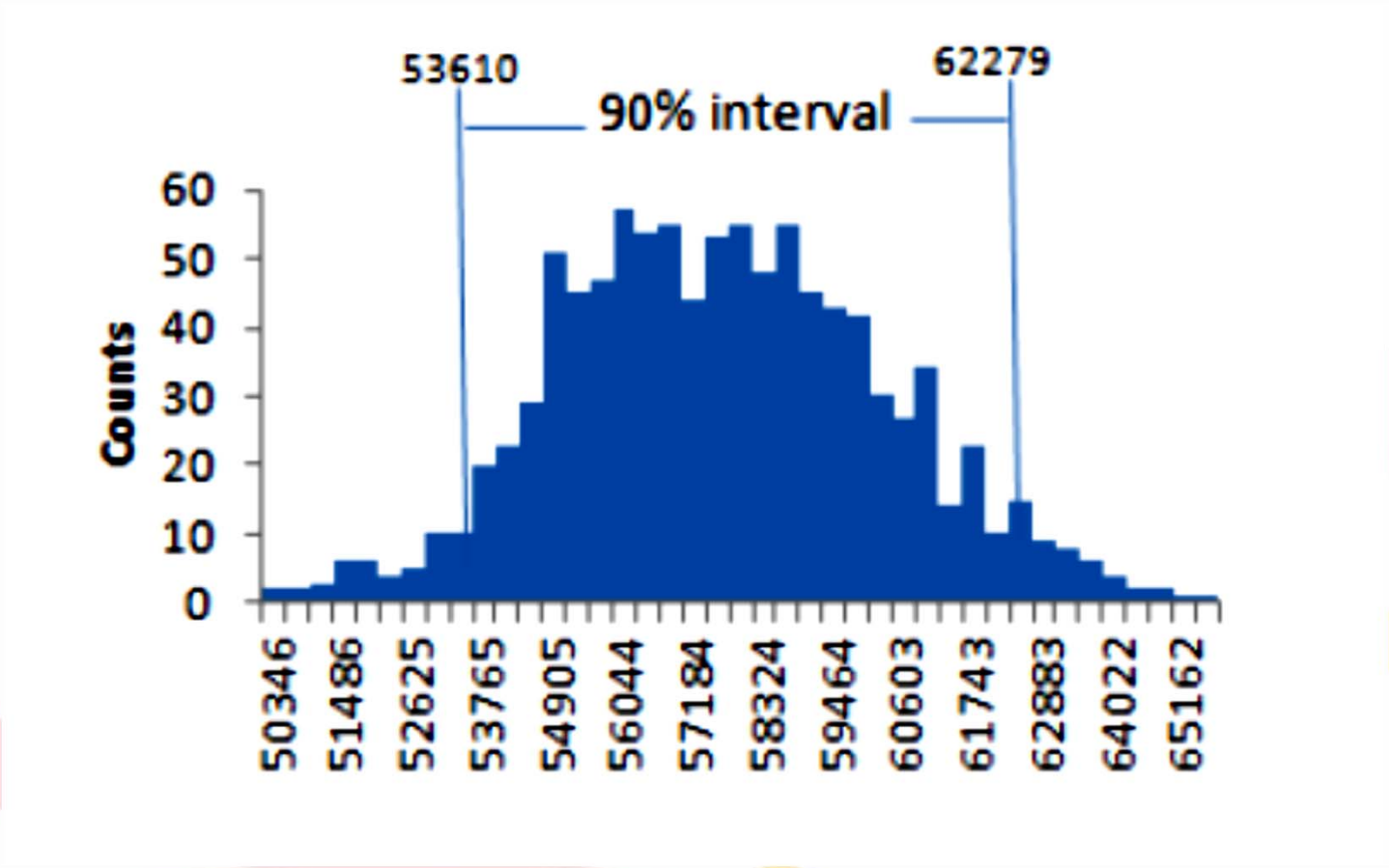
Confidence intervals

- Khoảng tin cậy (Confidence intervals) là một loại ước tính được tính toán từ thống kê của dữ liệu quan sát.
    - Khoảng tin cậy luôn đi kèm với mức độ bao phủ, được biểu thị bằng phần trăm (cao), chẳng hạn như 90% hoặc 95%.

Confidence intervals

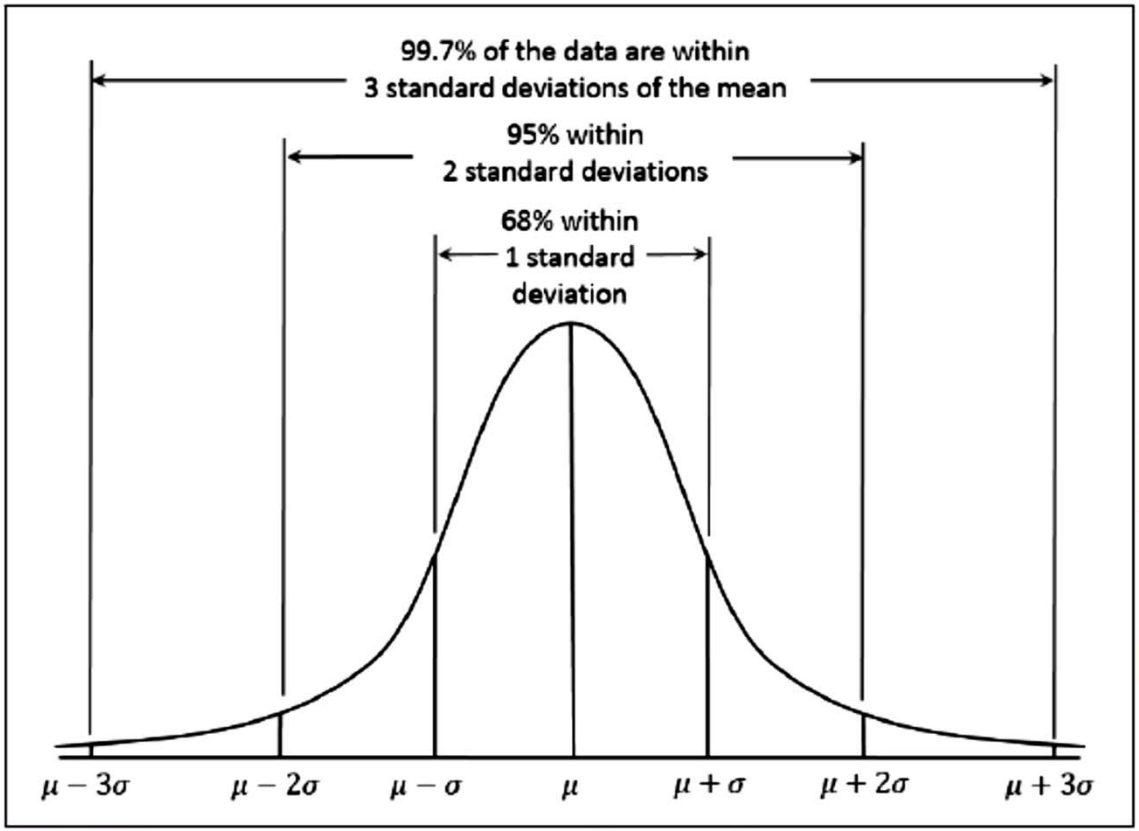
- Đưa ra một mẫu kích thước n và một thống kê mẫu được quan tâm, thuật toán cho một khoảng tin cậy của bootstrapnhư sau:
  - Trích xuất một mẫu ngẫu nhiên có kích thước n với sự thay thế từ dữ liệu (một mẫu trích xuất lại lại)
  - Ghi lại thống kê cho mẫu lấy lại
  - Lặp lại các bước 1-2 nhiều lần, gọi là B lần
  - Đối với khoảng tin cậy x%, hãy cắt [(1-x)/2]% kết quả lấy lại mẫu sau B lần từ một trong hai đầu của phân phối.
  - Các điểm cắt là điểm cuối của khoảng tin cậy x% bootstrap

Confidence intervals



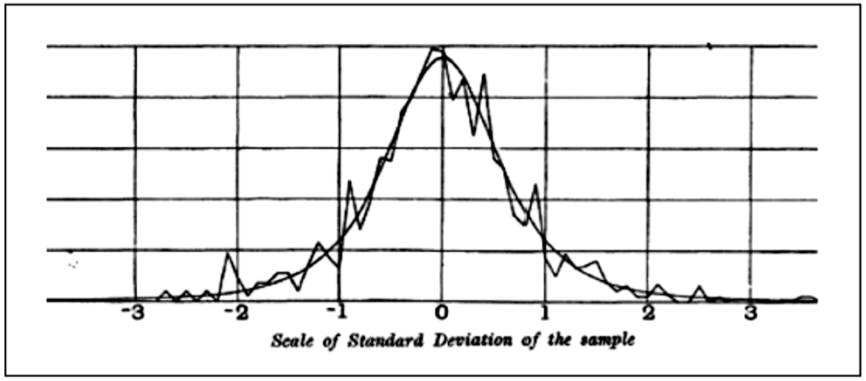
Distribution

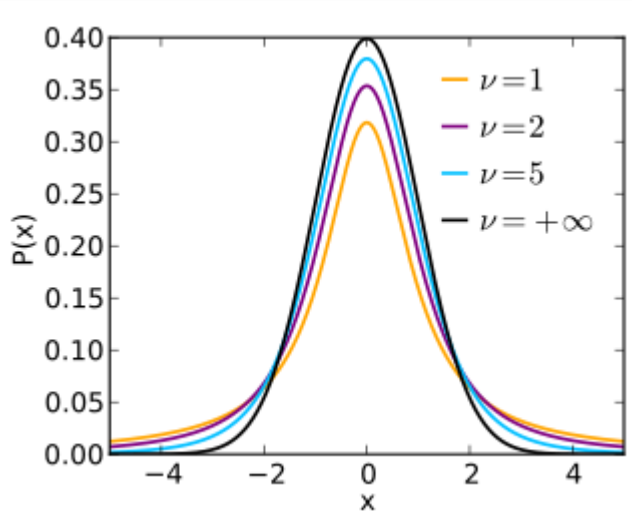
- Normal distribution



Distribution

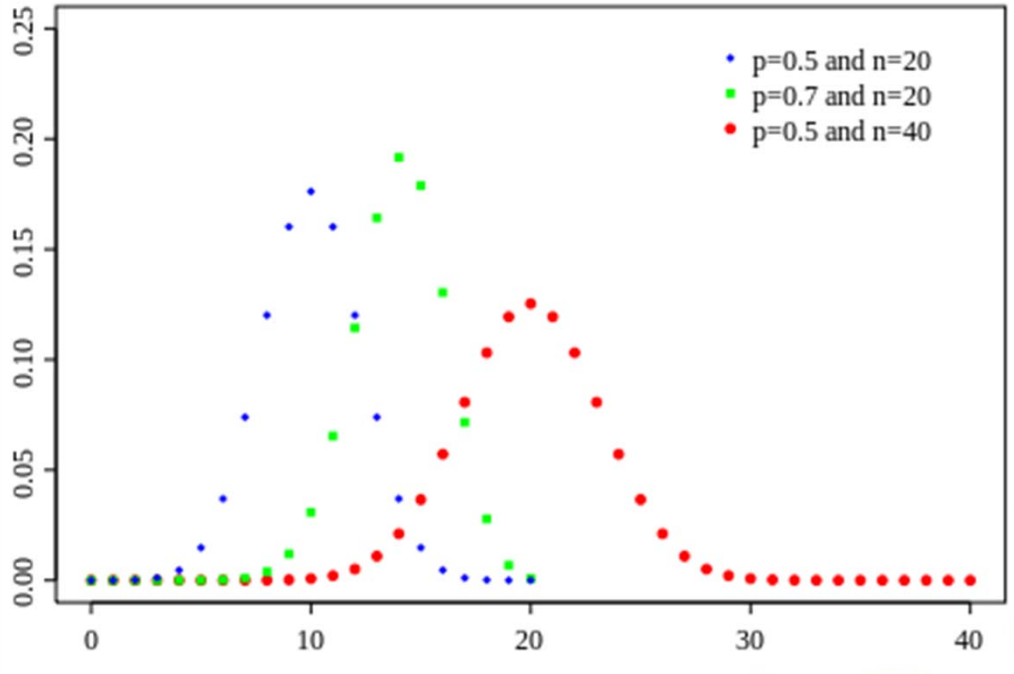
- Student's t distribution





Distribution

- Binomial distribution



Distribution

- Poisson Distributions
- Exponential distribution
- **Weibull distribution**