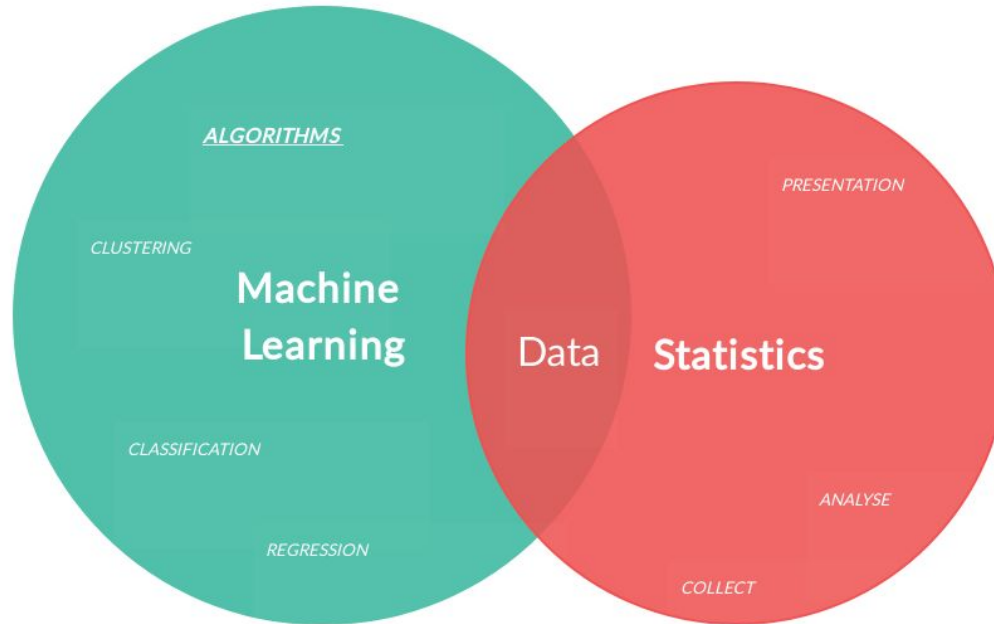


Sao họ biết mình muốn mua gì?
Sao họ biết mình muốn ăn gì?
Sao họ biết mình muốn xem gì?
Sao họ biết tự tag bạn mình vào ảnh?
Sao...
Sao...
Sao họ biết về mình lắm thế?

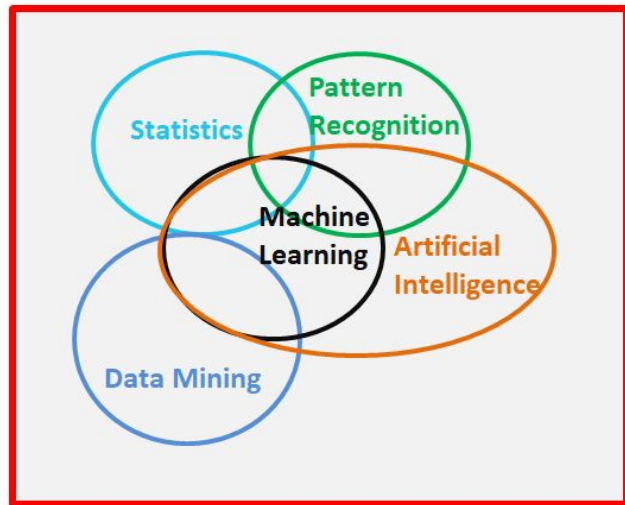


Statistical Machine Learning (SML)

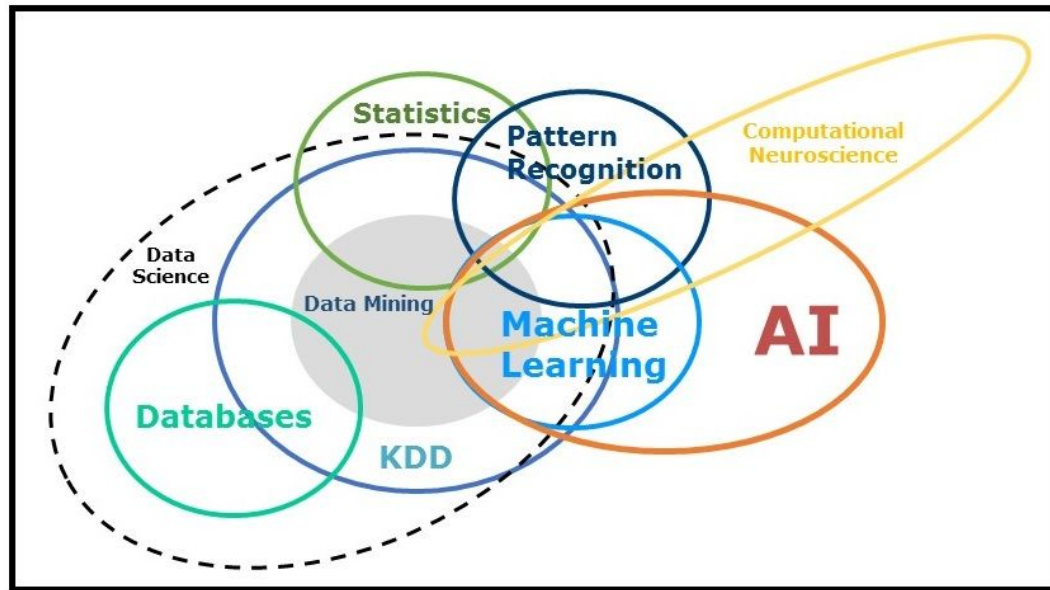


Tại sao phải học SML?

Sự giao thoa giữa các ngành

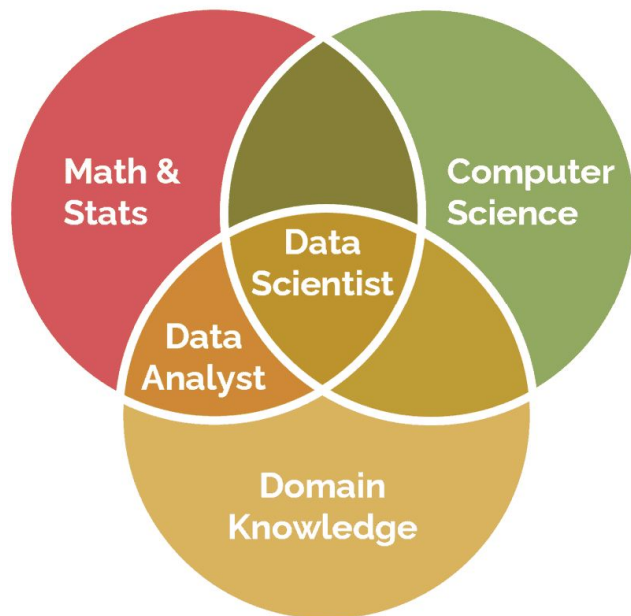


DATA SCIENCE



(KDD: Knowledge discovery in databases)

Tại sao phải học SML?



	Data Analyst	Data Scientist
Data Analyst Vs Data Scientist: Job Description	<ul style="list-style-type: none"> Conduct analyses Create visualizations, Report to stakeholders 	<ul style="list-style-type: none"> Create models Predict future trends Develop machine learning algorithms
Data Analyst Vs Data Scientist: Day-to-Day	<ul style="list-style-type: none"> Meetings, emails Clean and study data Communicate between teams 	<ul style="list-style-type: none"> Meetings, emails Clean and study data Communicate between teams Create and maintain models
Data Analyst Vs Data Scientist: Salary and Career Prospects	<ul style="list-style-type: none"> Highly sought after Average salary of \$70k IT, Healthcare, Finance, Insurance industries 	<ul style="list-style-type: none"> In-demand job Average salary of \$100k IT, Healthcare, Finance, Insurance industries
Data Analyst Vs Data Scientist: Background and Education	<ul style="list-style-type: none"> B.A. or B.S. in statistical field Some coding and database knowledge 	<ul style="list-style-type: none"> Masters or PhD Mastery in coding and database languages
Data Analyst Vs Data Scientist: Skills	<ul style="list-style-type: none"> Passion for business Communication skills Problem-solving instincts Data cleaning/analyzing skills 	<ul style="list-style-type: none"> Ability to see the bigger picture Interdisciplinary communication Model-building skills
Data Analyst Vs Data Scientist: What Comes Next?	<ul style="list-style-type: none"> Data scientist Senior specialist data analyst Data analytics consultant 	<ul style="list-style-type: none"> Individual contributor as data science specialist Data science team manager
Data Analyst Vs Data Scientist: Which is Best for You?	<ul style="list-style-type: none"> If you don't have the skillset yet, or don't have experience building models 	<ul style="list-style-type: none"> If you have a Masters or PhD in a statistical field and mastery in programming languages and databases

Các nguồn tham khảo...Internet



Statistical machine learning



[All](#)

[Images](#)

[Videos](#)

[News](#)

[More](#)

Tools

About 263,000,000 results (0.60 seconds)

Ad · <https://www.nus.edu.sg/>

Know Machine Learning Tools - Machine Learning Course

Adapt An In-Depth Introduction To **Machine Learning** Concepts, Algorithm And Techniques. Learn To Use Software Tools And Libraries In **Machine Learning** Projects. Sign Up Today! Power Your Career Now. Study At Your Own Pace. Accelerate Your Career.

Course List

Check out the various programmes we are offering today

Stackable Programmes

Study at your own pace. Take only the required modular courses

Blended Learning

Develop critical key digital skills beyond traditional classroom model

Executive Education

NUS ISS Offers A Suite Of High Quality Certified Short Courses In

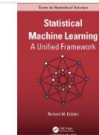
Scholarly articles for **Statistical machine learning**

Introduction to **statistical machine learning** ... - Sugiyama - Cited by 173

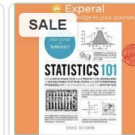
... and future applications of **statistical machine learning** ... - Rehman - Cited by 187

Statistical machine learning methods and remote ... - Holloway - Cited by 143

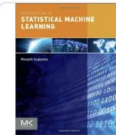
Ads · Shop Statistical machine l...



Statistical
Machine...
₫3,821,369
Fado.vn



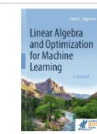
[Rẻ Hơn
Hoàn Tiền] ...
₫400,000 4...
tiki.vn



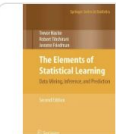
Introduction to
Statistical...
₫3,337,969
Fado.vn



An
Introduction...
₫3,076,441
Fado.vn



Linear
Algebra An...
₫160,000
Lazada Viet...

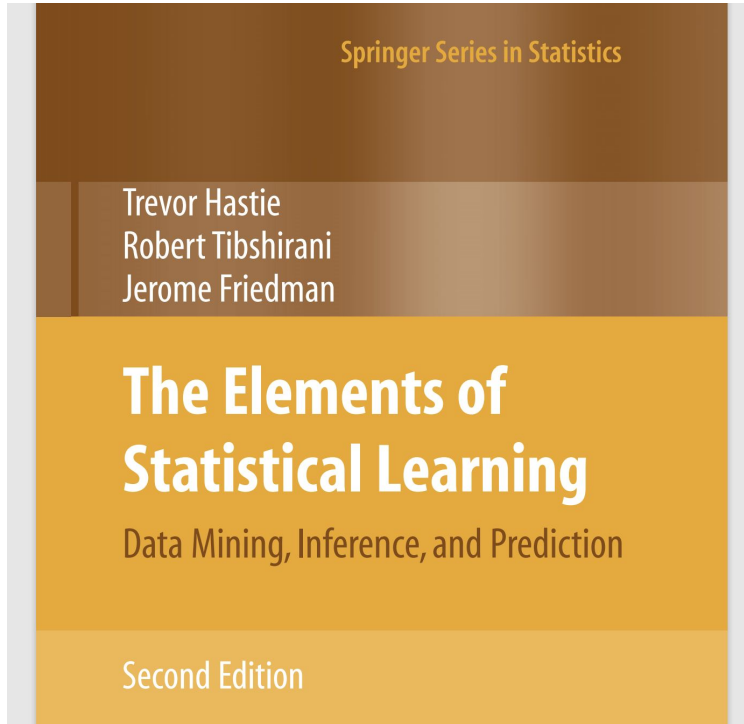


The Elements
of Statistica...
₫1,148,979
Fado.vn



Sách tham khảo trông như thế nào?

Ví dụ: <https://hastie.su.domains/Papers/ESLII.pdf>



viii Preface to the Second Edition	
Chapter	What's new
1. Introduction	
2. Overview of Supervised Learning	
3. Linear Methods for Regression	LAR algorithm and generalizations of the lasso
4. Linear Methods for Classification	Lasso path for logistic regression
5. Basis Expansions and Regularization	Additional illustrations of RKHS
6. Kernel Smoothing Methods	
7. Model Assessment and Selection	Strengths and pitfalls of cross-validation
8. Model Inference and Averaging	
9. Additive Models, Trees, and Related Methods	
10. Boosting and Additive Trees	New example from ecology; some material split off to Chapter 16.
11. Neural Networks	Bayesian neural nets and the NIPS 2003 challenge
12. Support Vector Machines and Flexible Discriminants	Path algorithm for SVM classifier
13. Prototype Methods and Nearest-Neighbors	
14. Unsupervised Learning	Spectral clustering, kernel PCA, sparse PCA, non-negative matrix factorization, archetypal analysis, nonlinear dimension reduction, Google page rank algorithm, a direct approach to ICA
15. Random Forests	New
16. Ensemble Learning	New
17. Undirected Graphical Models	New
18. High-Dimensional Problems	New

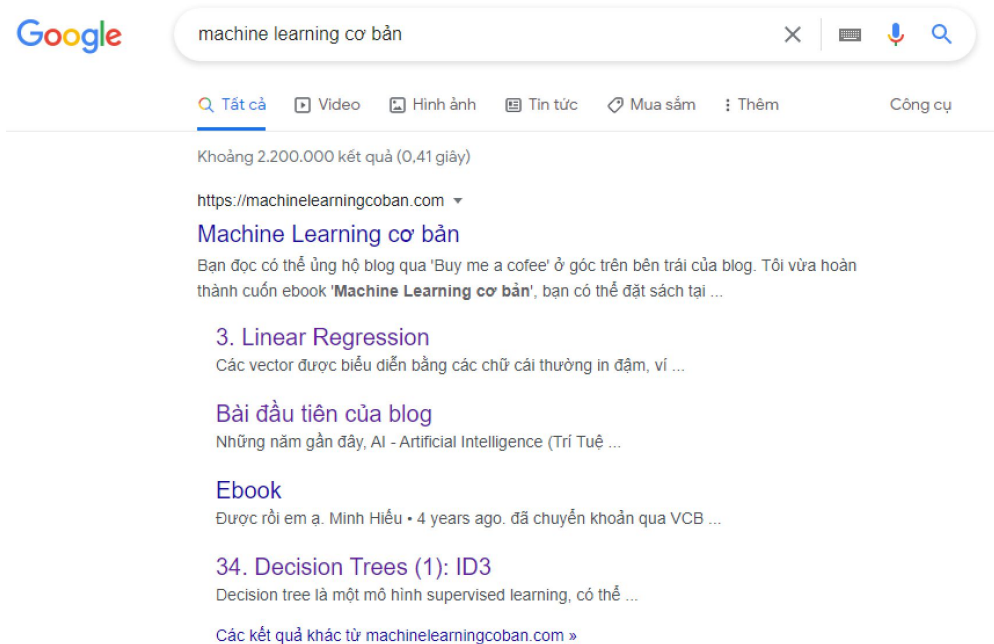
Some further notes:

- Our first edition was unfriendly to colorblind readers; in particular, we tended to favor red/green contrasts which are particularly troublesome. We have changed the color palette in this edition to a large extent, replacing the above with an orange/blue contrast.

Các nguồn tham khảo...Internet

Sách “Machine Learning Cơ Bản” (Vũ Hữu Tiệp)

Link: <https://machinelearningcoban.com/>



Google search results for "machine learning cơ bản". The search bar shows the query and the Google logo. Below the search bar, there are tabs for "Tất cả", "Video", "Hình ảnh", "Tin tức", "Mua sắm", and "Thêm", along with a "Công cụ" link. The search results show approximately 2,200,000 results in 0.41 seconds. The first result is from <https://machinelearningcoban.com> with the title "Machine Learning cơ bản". The snippet describes a book by Vũ Hữu Tiệp, available as an ebook. Other visible results include "3. Linear Regression", "Bài đầu tiên của blog", "Ebook", and "34. Decision Trees (1): ID3".

Khoảng 2.200.000 kết quả (0,41 giây)

<https://machinelearningcoban.com> ▾

Machine Learning cơ bản

Bạn đọc có thể ủng hộ blog qua 'Buy me a coffee' ở góc trên bên trái của blog. Tôi vừa hoàn thành cuốn ebook '**Machine Learning cơ bản**', bạn có thể đặt sách tại ...

3. Linear Regression

Các vector được biểu diễn bằng các chữ cái thường in đậm, ví ...

Bài đầu tiên của blog

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ ...

Ebook

Được rồi em ạ. Minh Hiếu • 4 years ago. đã chuyển khoản qua VCB ...

34. Decision Trees (1): ID3

Decision tree là một mô hình supervised learning, có thể ...

Các kết quả khác từ machinelearningcoban.com »

Còn khóa học của chúng ta?



Các bài toán phổ biến (bán lẻ, marketing, tài chính, chứng khoán): dự đoán, phân lớp, phân nhóm, gợi ý sản phẩm, cá nhân hóa, phân tích text...

Top 05 giải thuật trong phân tích dữ liệu (theo Edureka): Linear Regression, Decision Tree, Random Forest, Association Rule Mining, và K-Means Clustering

Và các thuật toán khác: PCA, LDA1, LDA2, SVD, SVM,...

Giảng viên và Trợ giảng



Quang-Khai Tran



Postdoctoral Scholar at **KISTI**
한국과학기술정보연구원



Studied Big Data Analytics at
**Korea University of Science
and Technology**

Facebook: <https://www.facebook.com/tgkhai2705/>

Email: tgkhai0527@gmail.com

[CyberLab] - Machine Learning 02

Friday & Monday: 6:00 – 10:00pm (Vietnam Time)

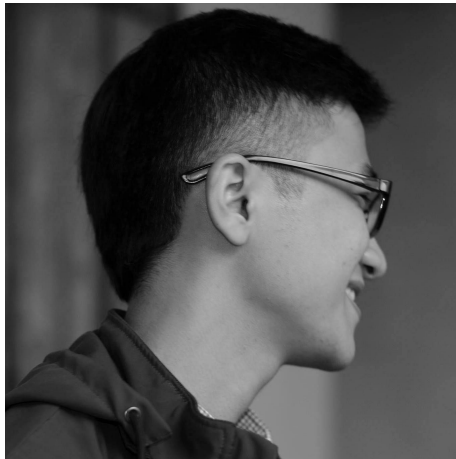
Google Meet: <https://meet.google.com/zaq-ahzh-qtf>

Giảng viên và Trợ giảng



Mai Ngọc Kiên

Thạc sĩ Khoa học Dữ liệu
ĐH UST - Học viện KISTI (Hàn Quốc)



Facebook: Kien Mai Ngoc
<https://www.facebook.com/hamsterviel.kien>

Huỳnh Quang Bảo

Thạc sĩ Khoa học Máy tính
ĐH Bách Khoa Tp. HCM



Facebook: Bao Huynh
<https://www.facebook.com/bao.huynh.1048554>

Statistics & Statistical Machine Learning

Bài 1: Descriptive Statistics



Quang-Khai Tran, Ph.D
CyberLab, 09/2022



(Ảnh: Internet)

Nội dung



1. Tổng quan về Probability & Statistics
2. Statistical Distributions
3. Ôn tập Descriptive Analytics

Tài liệu tiếng Việt:

- ❖ Khóa học Xác Suất Thống Kê, ViMentor:
<https://vimentor.com/vi/lesson/khoa-hoc-thong-ke>
- ❖ Tài liệu Thống Kê Ứng dụng:
https://maths.uel.edu.vn/Resources/Docs/SubDomain/maths/TaiLieuHocTap/ToanUngDung/thng_k_ng_dng.html
- ❖ Suy luận thống kê:
<https://nghiencuugiaoduc.com.vn/suy-luan-thong-ke/>

Tài liệu tiếng Anh:

- ❖ Google, Youtube: “probability and statistics”



Phần 1. Tổng Quan

1.1. Giới thiệu Tổng Quan

1.2. Một số thuật ngữ cơ bản

- Biến ngẫu nhiên
- Hàm phân phối XS
- Tổng thể & Mẫu
- Giá trị kỳ vọng

1.3. Hai định lý/luật quan trọng:

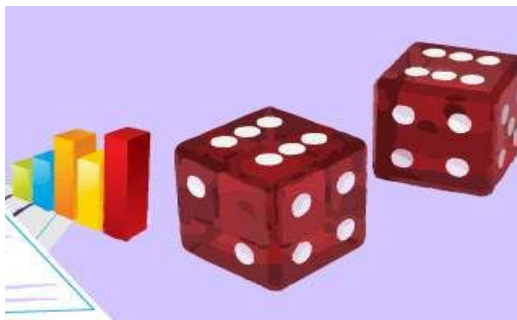
- Law of large number
- Central Limit Theorem

Xác Suất Thống Kê là gì?

- ❖ Về bản chất, đây là hai ngành khác nhau trong Toán học
- ❖ Nhưng có quan hệ mật thiết

Xác suất (Probability)

Là ngành khoa học nghiên cứu về khả năng xảy ra của các kết quả (outcome) của các hiện tượng



Thống kê (Statistics)

Là ngành khoa học nghiên cứu về việc thu thập, phân tích, giải thích, trình bày, và tổ chức dữ liệu



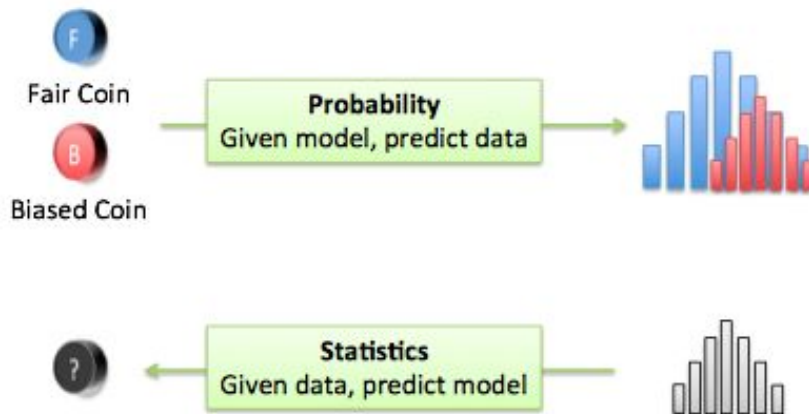
Phân biệt Xác Suất vs. Thống Kê

Xác suất (Probability)

ám chỉ 'khả năng' hoặc 'cơ hội' xảy ra của một sự việc

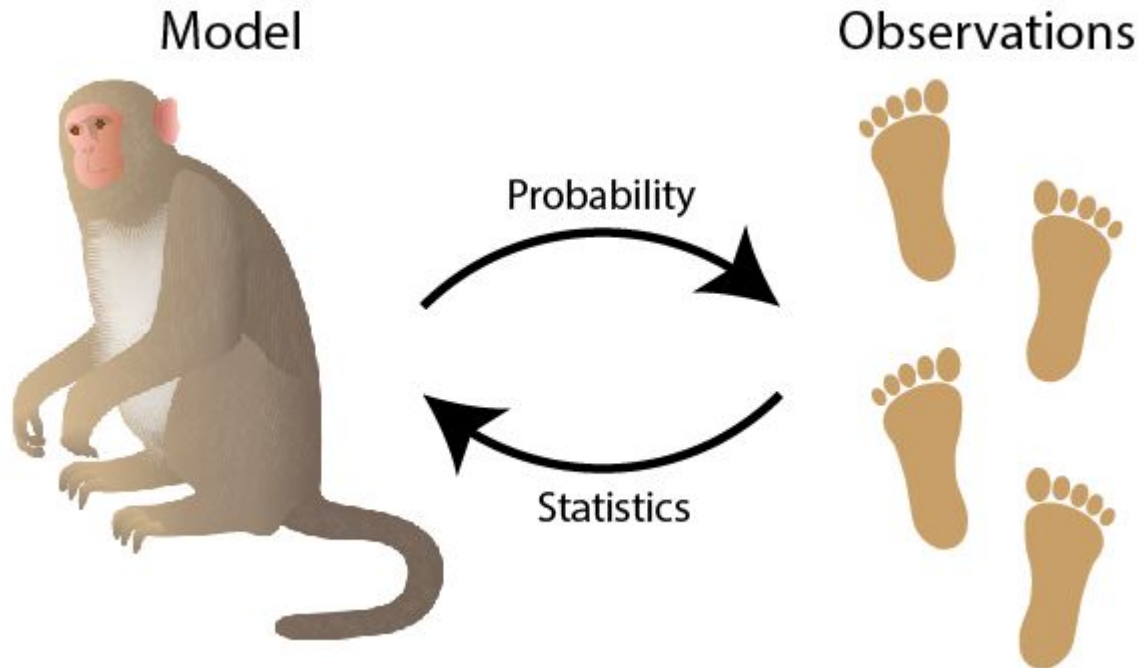
Thống kê (Statistics)

là cách chúng ta hiểu về nhiều loại data bằng các kỹ thuật khác nhau, nhằm biểu diễn data phức tạp thành các dạng dễ hiểu



1 Tổng quan về Probability & Statistics

Phân biệt Xác Suất vs. Thống Kê



Ứng dụng:

- ❖ Nền tảng quan trọng của các mô hình học máy và phân tích dữ liệu
- ❖ Hiểu về các phân phối xác suất giúp data scientist
 - Nắm bắt được dữ liệu một cách toàn diện hơn,
 - Thực hiện được các phân tích phức tạp hơn,
 - Chọn được mô hình phù hợp hơn
- ❖ Ứng dụng nhiều trong mọi mặt đời sống:
Kinh tế, Khoa học Kỹ thuật, Y học, Xã hội học...

Bài viết "Áp dụng cách suy nghĩ xác suất vào trong cuộc sống" (Huyền Chip)

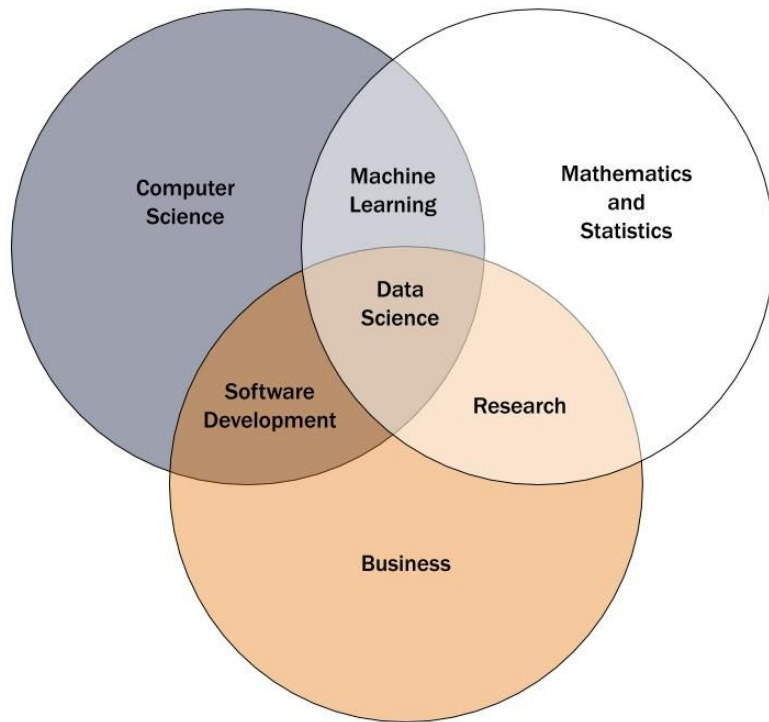
Link: <http://www.cscb.vimaru.edu.vn/ap-dung-cach-suy-nghi-xac-suat-vao-trong-cuoc-song>

Vai trò của XSTK trong KHDL

Josh Wills, a former head of data engineering at Slack



Vai trò của XSTK

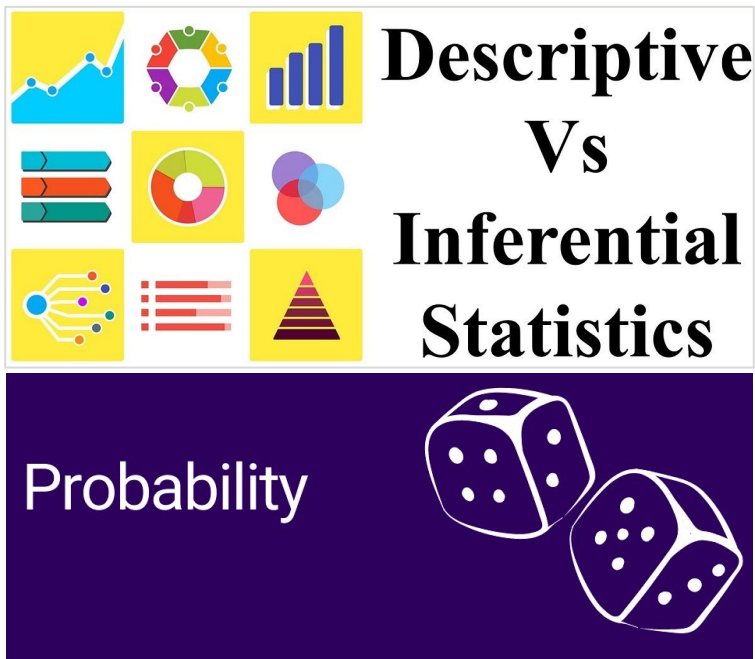


Vai trò của XSTK



Ba nhánh chính của Xác Suất Thống Kê

- ❖ Descriptive Statistics
(thống kê mô tả)
- ❖ Inferential Statistics
(thống kê suy luận)
- ❖ Probability (xác suất)



❖ Hai loại thống kê

Descriptive Statistics

(là khoa học về việc) thu thập, sắp xếp, tóm tắt và hiển thị dữ liệu

Inferential Statistics

(là khoa học về việc) sử dụng descriptive statistics để ước lượng (estimate) các tham số của quần thể

❖ Hai trường phái Frequentist vs. Bayesian

Frequentist

- Xác suất là một con số khách quan, một thực tế (fact), hoàn toàn độc lập với niềm tin của người phân tích
- Các phương pháp frequentist thường yêu cầu số lượng quan sát lớn

Bayesian

- Xác suất là một ý kiến chủ quan (opinion), một niềm tin (belief), phụ thuộc vào niềm tin của người phân tích
- Có thể áp dụng với số lượng quan sát lớn hoặc nhỏ

Tham khảo: <https://linhnghiem.org/2019/11/03/cuoc-chien-frequentist-vs-bayesian/>

❖ Hai trường phái Frequentist vs. Bayesian

Frequentist

[repeat repeat repeat]



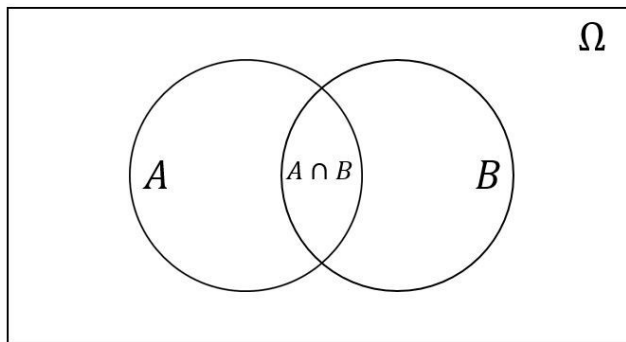
Bayesian

[observe, guess, experiment]



Tham khảo: <https://aicurious.io/posts/xstk-khai-niem/>

- ❖ Không gian mẫu là tập hợp Ω , các sự kiện là tập con của Ω
- ❖ Với mọi sự kiện A : $0 \leq P(A) \leq 1$
- ❖ $P(\Omega) = 1$
- ❖ Nếu A và B độc lập: $P(A \cup B) = P(A) + P(B)$
- ❖ Với A và B bất kỳ: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



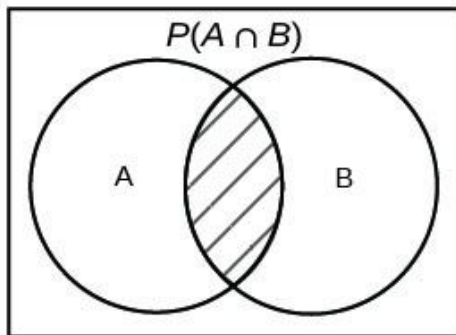
Xác suất có điều kiện

❖ Với A và B bất kỳ:

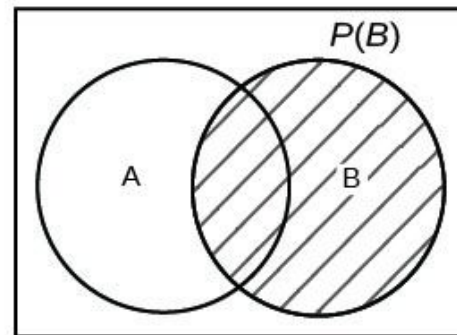
$$P(A | B) = P(A \cap B) / P(B)$$

$$P(A \cap B) = P(A) \cdot P(B | A)$$

$$P(A|B) =$$



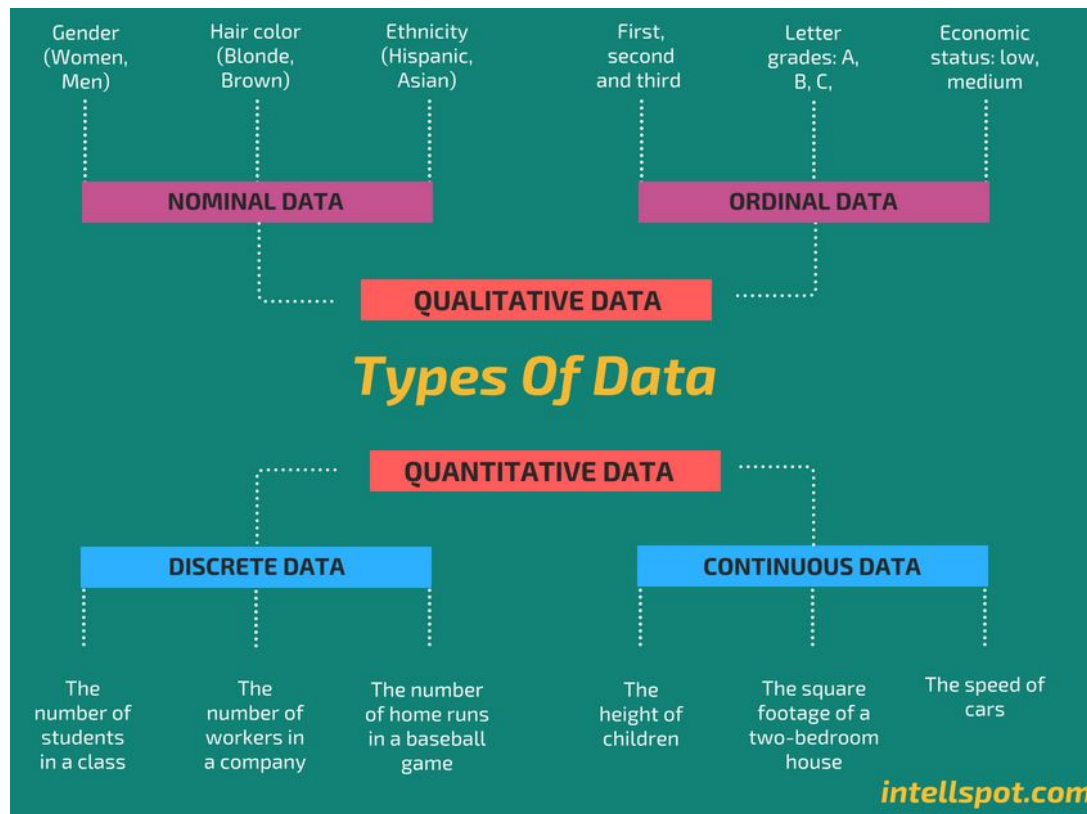
\div



1.2 Một số thuật ngữ cơ bản

❖ Phân loại dữ liệu trong thống kê:

- Qualitative
- Quantitative



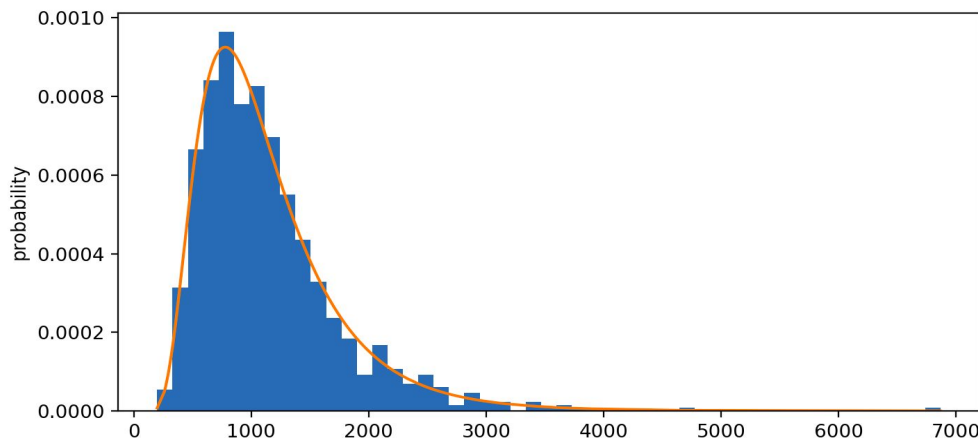
1.2 Một số thuật ngữ cơ bản

Các hàm phân phối xác suất: PMF, PDF, CDF

Hàm phân phối xác suất:

- Là quy luật cho biết cách gán mỗi xác suất cho mỗi khoảng giá trị của tập số thực
- Sao cho các tiên đề xác suất được thỏa mãn

(*Wikipedia: [Link](#)*)

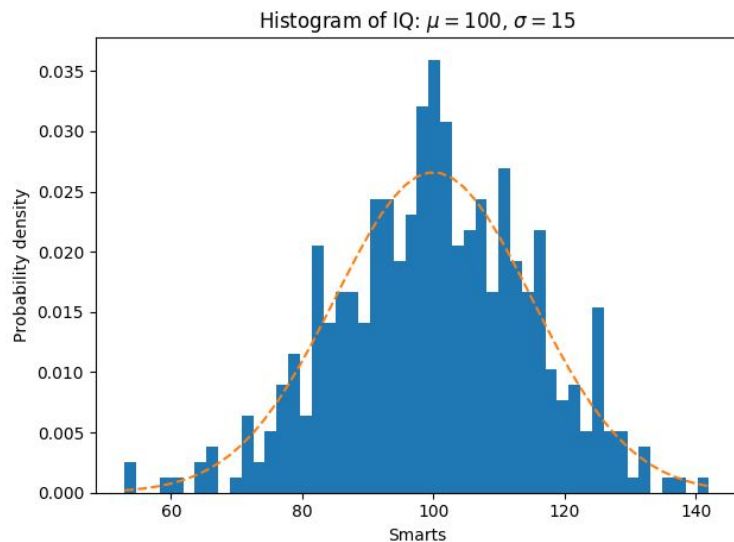


1.2 Một số thuật ngữ cơ bản

Các hàm phân phối xác suất: PMF, PDF, CDF

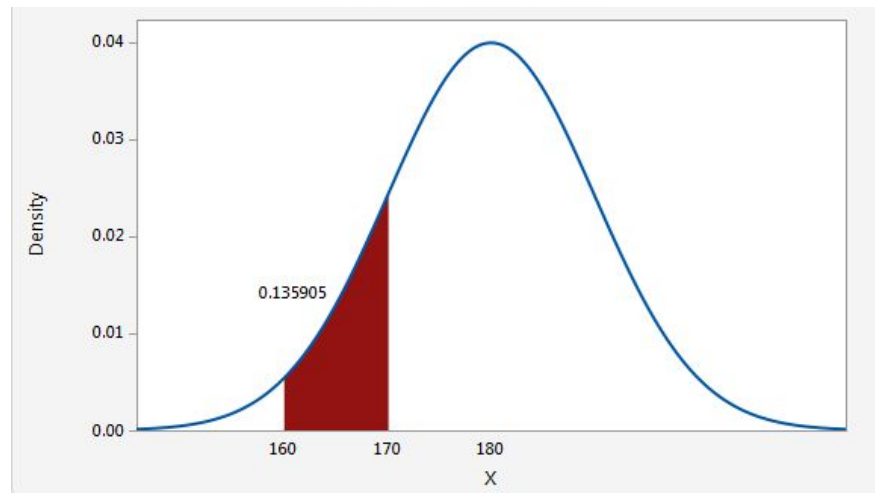
Với biến rời rạc (PMF)

$$f(x) = \text{Prob}(X=x)$$



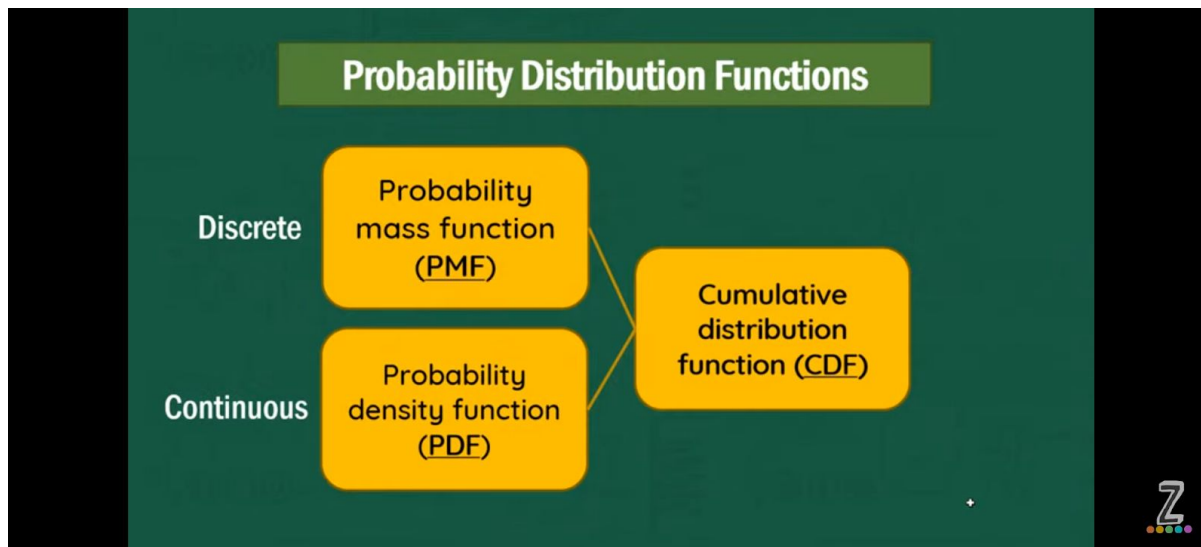
Với biến liên tục (PDF)

Nếu X là một biến ngẫu nhiên, phân phối xác suất tương ứng gán cho đoạn $[a, b]$ một xác suất $f(X) = P[a \leq X \leq b]$



1.2 Một số thuật ngữ cơ bản

Các hàm phân phối xác suất: PMF, PDF, CDF

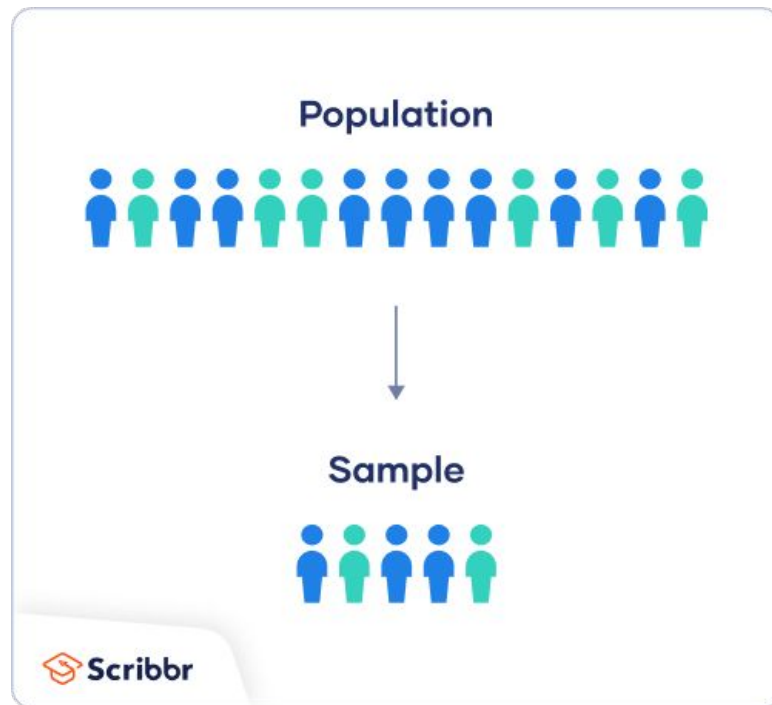
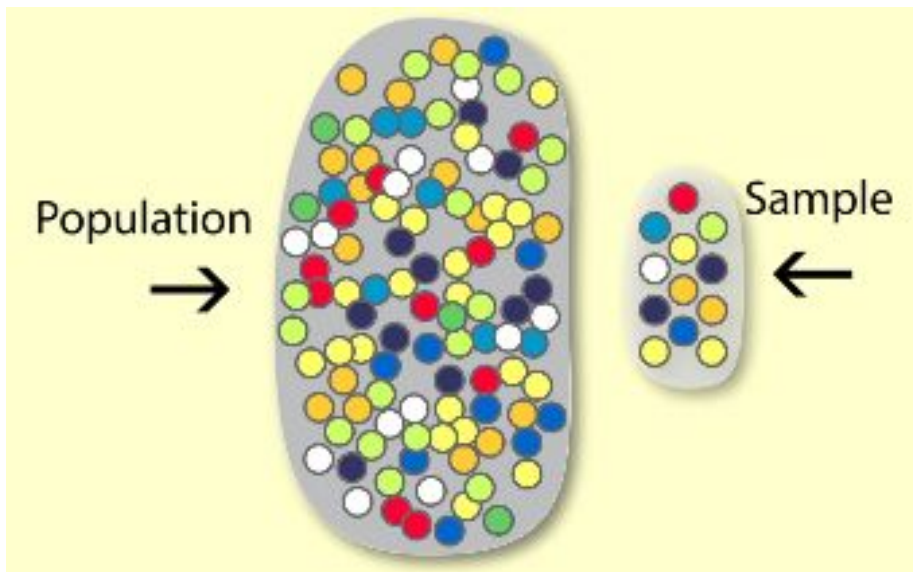


Các thuật ngữ về Tổng thể & Mẫu

- ❖ Population vs. Sample (quần thể vs. mẫu)
 - Population: mọi thứ (hoặc mọi người) được nghiên cứu/
 - Sample: một phần của population
- ❖ Parameter vs. Statistic (tham số vs. thống kê)
 - Parameter: là thuộc tính/đặc tính của population
 - Statistic: là thuộc tính/đặc tính của sample
- ❖ Variable (biến): là thuộc tính/đặc tính mà ta quan tâm trên mỗi item (thành phần) của một sample
- ❖ Data (dữ liệu): các giá trị thực sự của các variables

1.2 Một số thuật ngữ cơ bản

❖ Population vs. Sample



1.2 Một số thuật ngữ cơ bản

Ví dụ:

Population	Toàn bộ học viên của CyberLab từ trước đến nay
Parameter	Tuổi trung bình của tất cả các học viên
Sample	31 học viên
Statistic	Tuổi trung bình của 31 học viên này
Variable	(Thuộc tính) Tuổi của một học viên nào đó
Data	Giá trị trên thực tế của tuổi các học viên (chẳng hạn một bạn nào đó 25 tuổi, một bạn khác 27 tuổi)

Khái niệm Frequency (tần số): một giá trị xuất hiện bao nhiêu lần?

1.2 Một số thuật ngữ cơ bản

Giá trị kỳ vọng (expected value) của một biến ngẫu nhiên

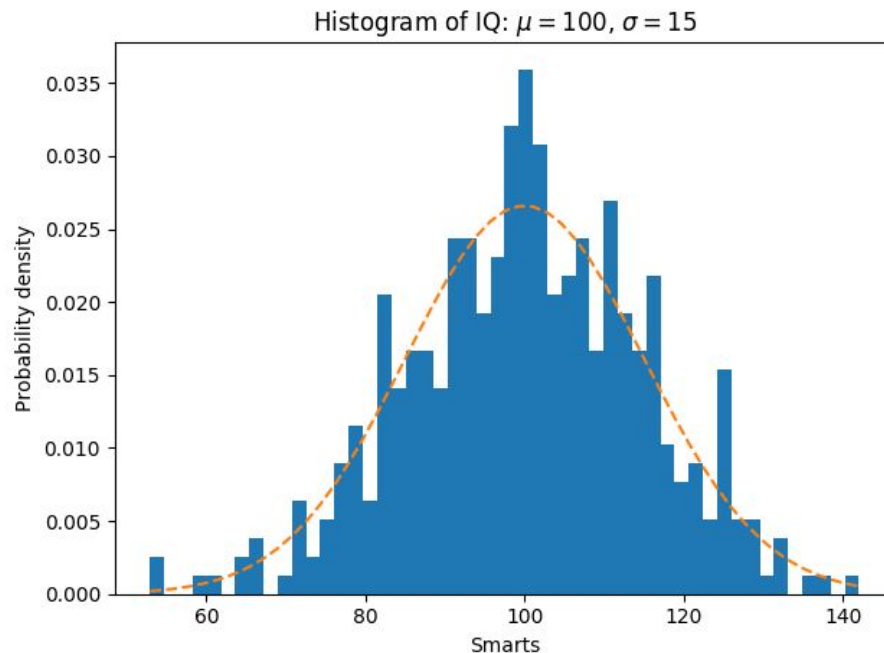
- ❖ Thường được hiểu là thuật ngữ về xác suất (probability)
- ❖ Tên gọi khác: giá trị mong đợi, hoặc trung bình
- ❖ **Định nghĩa:** là trung bình có trọng số của tất cả các giá trị cụ thể của biến
- ❖ **Cách tính:** tổng các tích giữa xác suất xảy ra của mỗi giá trị với giá trị đó

$$E[X] = \sum_i x_i f(x_i)$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

1.2 Một số thuật ngữ cơ bản

Ví dụ: Giá trị kỳ vọng của chỉ số IQ



1.2 Một số thuật ngữ cơ bản

Ví dụ: Giá trị kỳ vọng của một phân bố chuẩn tắc (khi $\mu = 0$, $\sigma = 1$, x là Z)

- Probability density function

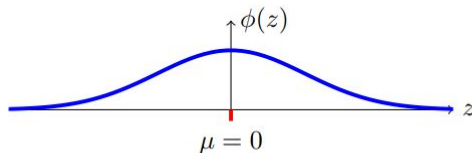
$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ Expected value

σ Standard deviation



$$E(Z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Big|_{-\infty}^{\infty} = \boxed{0}.$$



1.2 Một số thuật ngữ cơ bản

- ❖ Variance (Phương sai), Standard Deviation (Độ lệch chuẩn)
- ❖ Coefficient of Variation (hệ số biến thiên)

đo độ phân tán của dữ liệu so với giá trị trung bình

⇒ giúp so sánh được các trường hợp/lựa chọn



Coefficient of Variation Formula = $\frac{\text{Standard Deviation}}{\text{Mean}}$



Quỹ	Lợi nhuận trung bình năm	Độ lệch chuẩn	%CV
A	5.47%	14.68%	2.68
B	6.88%	21.31%	3.09
C	7.16%	19.46%	2.72

⇒ Trường hợp A là tốt nhất (vì ít rủi ro hơn). Nguồn:
<https://vietnambiz.vn/he-so-bien-thien-coefficient-of-variation-cv-la-gi-nhung-dac-diem-can-luu-y-20191121233238319.htm>

Bằng trực giác:

- Giá trị trung bình của nhiều giá trị (đã đo được) cho ta một ước lượng tốt hơn chỉ có một giá trị
- Lý do là vì các yếu tố gây lỗi/nhiều ngẫu nhiên sẽ trung hòa (cancel out) lẫn nhau

Central Limit Theorem (CLT)

states that when sample size tends to infinity, the sample mean will be normally distributed.

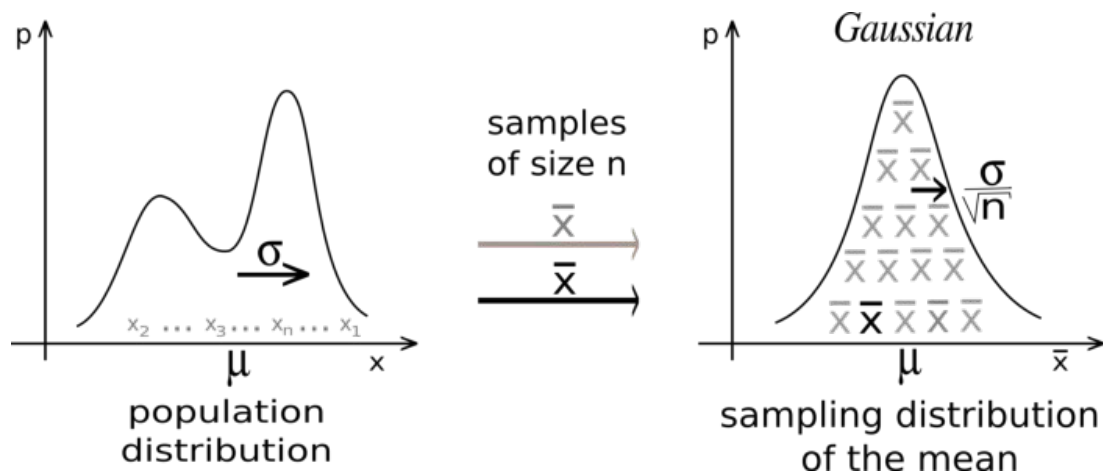
(Định lý Giới hạn Trung tâm phát biểu rằng khi kích thước của mẫu tiến tới vô cùng lớn, giá trị trung bình của mẫu sẽ tuân theo phân bố chuẩn)

Law of Large Number (LoLN)

states that when sample size tends to infinity, the sample mean equals to population mean.

(Định luật Số Lớn phát biểu rằng khi kích thước mẫu tiến tới vô cùng lớn, giá trị trung bình của mẫu sẽ bằng trung bình của quần thể)

1.3 Hai định lý/luật quan trọng



1.3 Hai định lý/luật quan trọng

Từ đó rút ra:

- ❖ LoLN: trung bình của nhiều samples độc lập khác nhau sẽ tiến gần đến điểm trung bình của toàn thể.
- ❖ CLT: tổng hoặc trung bình của nhiều phiên bản độc lập (independent copies) của một biến ngẫu nhiên là xấp xỉ phân bố chuẩn.
- ❖ Điều tương tự cũng đúng với mean và std của một biến thông thường.

Trên thực tế: $n \geq 30$

Theo giáo trình của MIT ([Link](#))

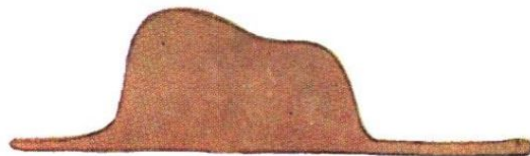


Phần 2. Phân Phối Thống Kê

- 2.1. Phân phối TK là gì?
- 2.2. Các phân phối XS cơ bản
- 2.3. Demo

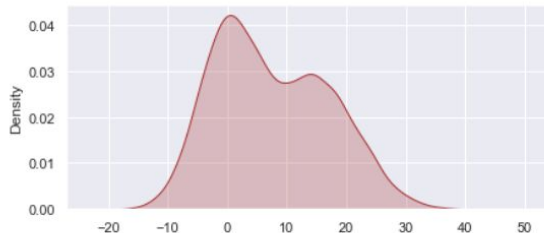
(Các) Phân phối thống kê:

là cách diễn đạt cấu trúc của một tập dữ liệu hoặc một quần thể bằng một trong các (hàm) phân phối xác suất



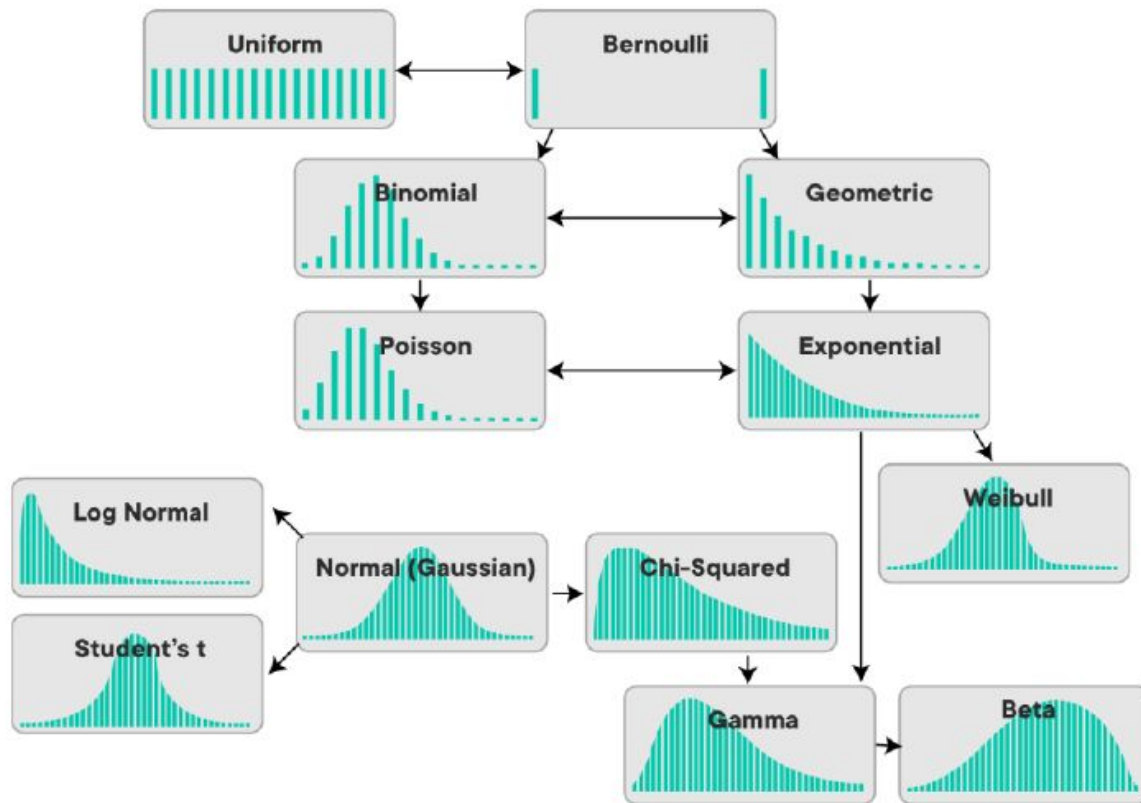
Source:

<https://towardsdatascience.com/6-useful-probability-distributions-with-applications-to-data-science-problems-2c0bee7cef28>



Phân phối xác suất:

- Là các hàm toán học (PDF, PMF, CDF)
- Cho ra xác suất của các outcomes khác nhau của một biến ngẫu nhiên



Biến ngẫu nhiên rời rạc và Biến ngẫu nhiên liên tục

❖ **Random variable:**

là một mô tả bằng số của số lượng hoặc đối tượng trong một đo đạc/thử nghiệm thống kê

❖ **Discrete variable:**

chỉ nhận các giá trị bằng số có thể đếm được của các giá trong một khoảng nào đó

❖ **Continuous variable:**

có thể nhận bất cứ giá trị bằng số nào trong một khoảng

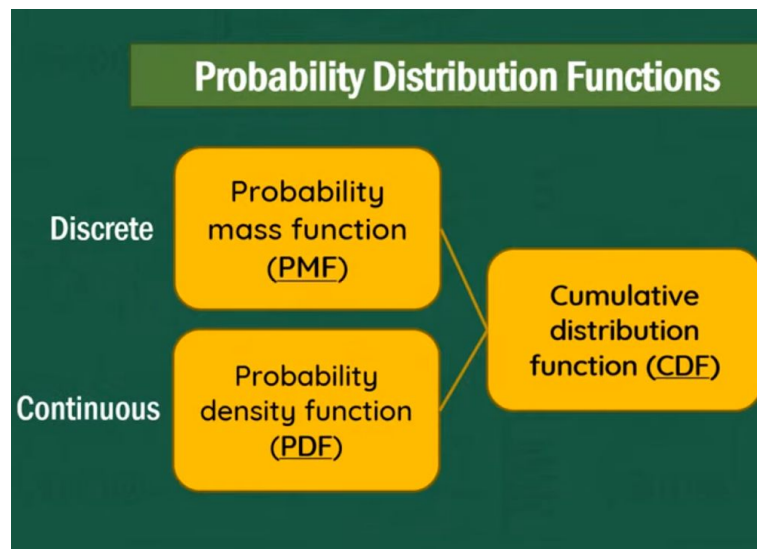
Phân phối rời rạc và Phân phối liên tục

❖ Phân phối rời rạc:

- Thuộc các biến rời rạc
- Hàm Probability Mass Function (hàm khối/lượng xác suất)
- Ví dụ: Bernoulli, Binomial, Geometric, Poisson

❖ Phân phối liên tục:

- Thuộc các biến liên tục
- Hàm Probability Density Function (hàm mật độ xác suất)
- Ví dụ: Uniform, Normal, Exponential



2.2 Các phân phối XS cơ bản

Các hàm tạo phân bố:

❖ Numpy:

<https://numpy.org/doc/stable/reference/random/generator.html#distributions>

❖ Scipy:

<https://docs.scipy.org/doc/scipy/reference/stats.html#probability-distributions>

2.2 Các phân phối XS cơ bản

Tham khảo: 7 phân phối thống kê cơ bản "phải biết"

- ❖ Tiếng Việt:
<https://ichi.pro/vi/bay-phan-phoi-thong-ke-p-hai-biet-va-mo-phong-cua-chung-cho-khoa-hoc-du-lieu-114470991175218>
- ❖ Bản gốc (tiếng Anh):
<https://towardsdatascience.com/seven-must-know-statistical-distributions-and-their-simulations-for-data-science-681c5ac41e32>



7 phân phối thống kê cơ bản "phải biết"

❖ Các phân phối rời rạc:

- (Bernoulli & Binomial & Geometric Distributions)
- Phân phối Poisson (Poisson Distribution)

❖ Các phân phối liên tục:

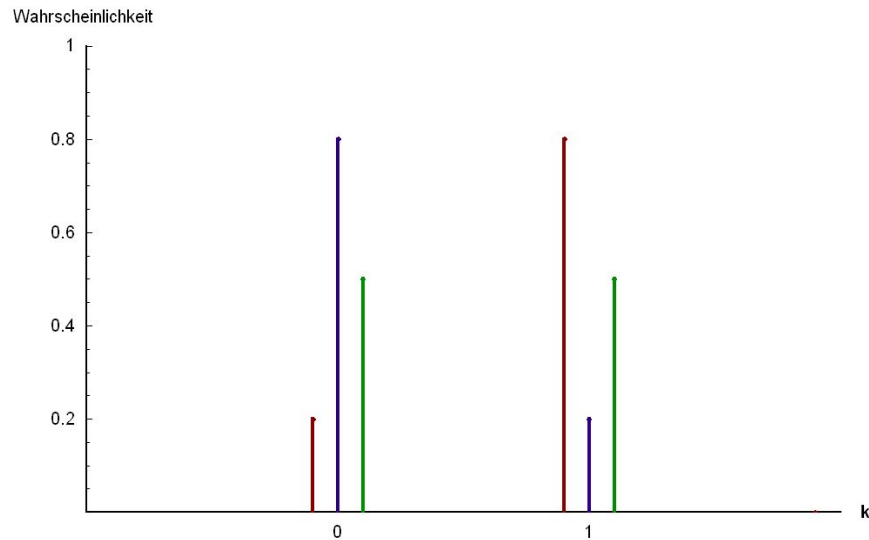
- Phân phối đều và phân phối chuẩn (Uniform & Normal Distributions)
- Phân phối lũy thừa (Exponential Distribution)

Các phân phối khác

- ❖ Phân phối Beta
- ❖ Phân phối Gamma
- ❖ Phân phối Pareto
- ❖ Phân phối Chi-square

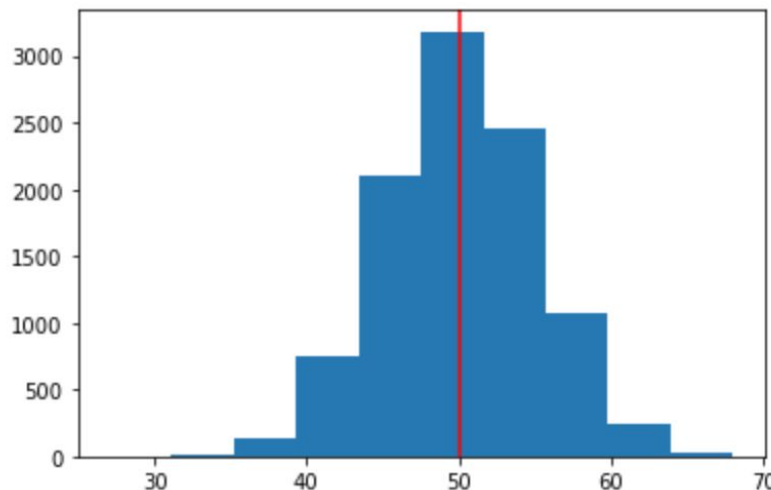
Phân phối Bernoulli

- Chỉ có 2 outcomes: thường là 1 và 0 (hoặc True/False)
- Chỉ có 1 lần thử



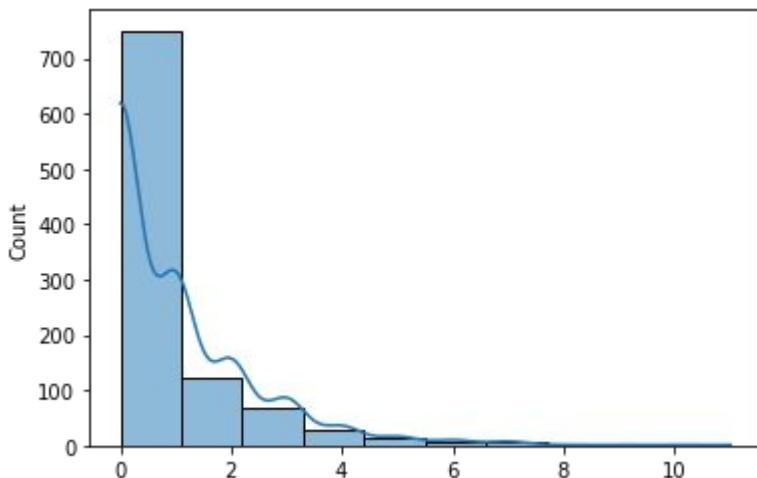
Phân phối Binomial (nhị thức)

- Tương tự phân phối Bernoulli nhưng có nhiều lần thử
- Đếm số lần thành công
- Ví dụ: số khách hàng mua hàng trong mỗi 100 khách



Phân phối Geometric (hình học)

- Tương tự phân phối Bernoulli nhưng có nhiều lần thử
- Đếm số số lượt thử cần thiết cho lần thành công đầu tiên
- Ví dụ: số users xem hàng trên website trước khi có một người thực hiện mua



$$f(x) = \begin{cases} p(1-p)^x, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

$$E(x) = \frac{1-p}{p}$$

$$Var(x) = \frac{1-p}{p^2}$$

Phân phối Geometric (hình học)

Ví dụ:

- Giả sử một cửa hàng quan sát thấy có 20% khách hàng sẽ mua hàng
- Giả sử lúc mới mở cửa, họ thấy có 4 khách hàng vào nhưng không mua. Vậy xác suất khách hàng thứ 5 mua hàng là bao nhiêu?

$$P(X = x) = p \times (1 - p)^{x - 1}$$

$$P(X = 5) = 0.2 \times (1 - 0.2)^{5-1} = 0.2 \times 0.8^4 = 0.08192 \approx 8.2\%$$

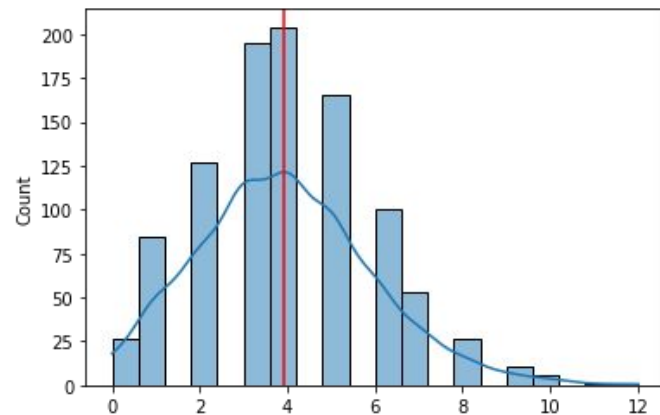


Phân phối Poisson

- Là phân phối rời rạc
- Biểu diễn xác suất của số lần xuất hiện của một sự kiện trong một khoảng cố định thời gian hoặc không gian
- Ví dụ:
 - Số khách hàng vào cửa hàng trong 1 phút
 - Số xe đi qua một ngã tư trong một giờ

$$f(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

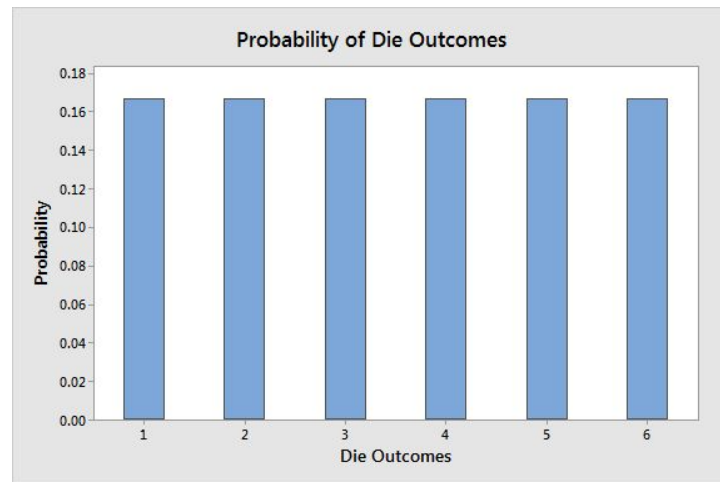
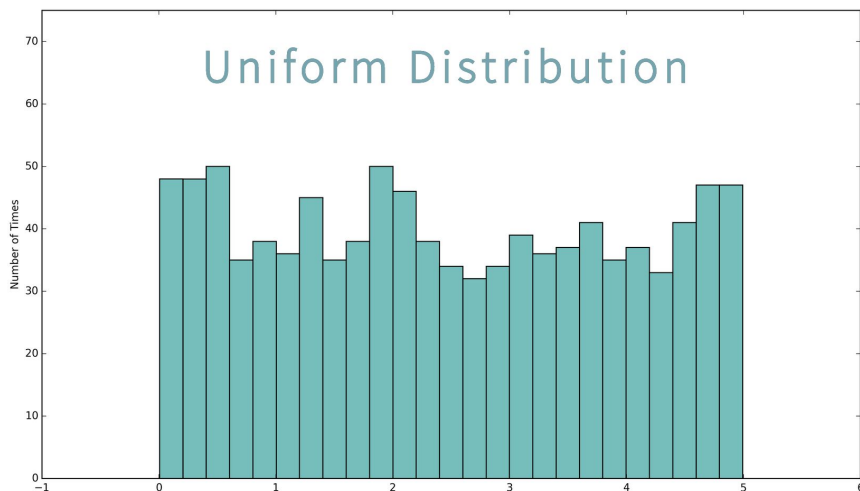
$$E(x) = Var(x) = \lambda$$



Phân phối Uniform (đều)

là phân phối cho các giá trị liên tục có xác suất bằng nhau

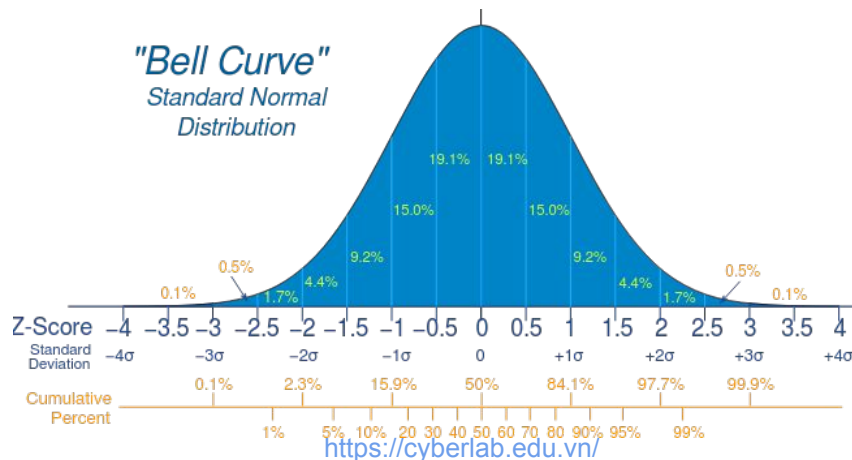
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & otherwise \end{cases}$$



Normal Distribution (Phân phối Chuẩn)

- Là phân phối liên tục phổ biến nhất
- Đặc tính: mean = mode = median = μ
- Hàm PDF:

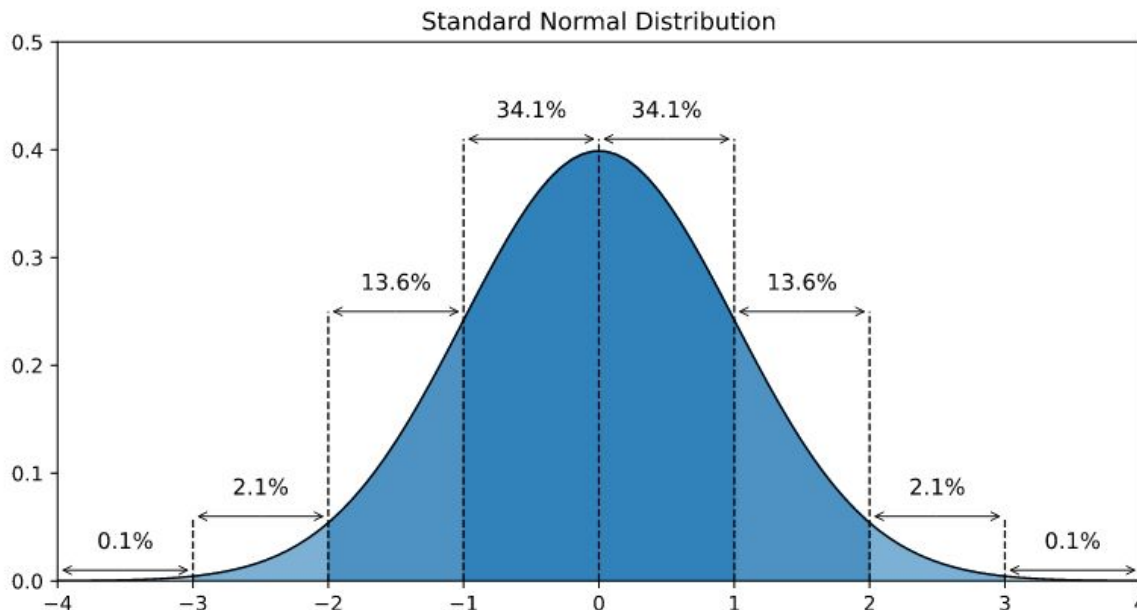
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



2.2 Các phân phối XS cơ bản

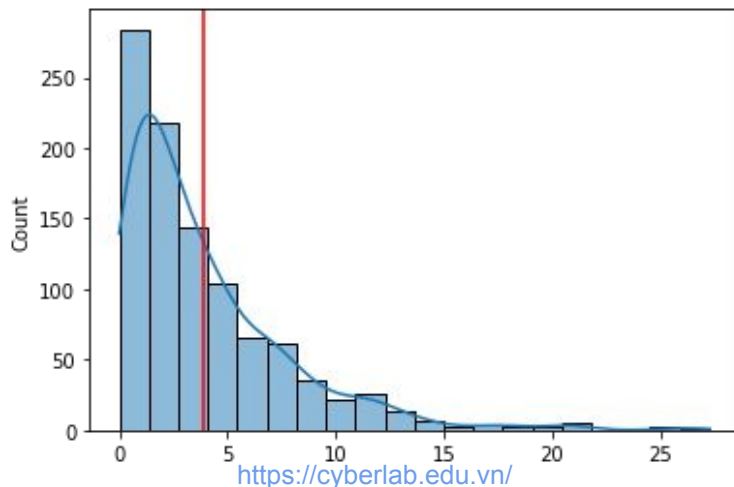
Standard Normal Distribution (Phân phối Chuẩn Tắc)

Khi $\mu = 0$ và $\sigma = 1$



Exponential Distribution (Phân phối Lũy thừa)

- Là phân phối liên tục của khoảng thời gian giữa các sự kiện Poisson
- Ví dụ:
 - Thời gian cách nhau mà 2 khách hàng vào cửa hàng
 - Thời gian giữa 2 cuộc gọi đến tổng đài CSKH



2.2 Các phân phối XS cơ bản

Các phân phối khác

Beta	Mô tả xác suất của xác suất	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$
Gamma	Tương tự exponential nhưng mô tả phân phối của thời gian đến sự kiện thứ k	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \quad \left \quad \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Pareto	Phân bố xác suất theo quy luật quyền lực; sử dụng để mô tả xã hội, kiểm soát chất lượng, khoa học, địa vật lý, tính toán ...	$\frac{a b^2}{a - 2}$
Chi-square	Là một trường hợp đặc biệt của phân phối gamma, dùng để mô tả phân phối tổng bình phương của k biến ngẫu nhiên chuẩn độc lập chuẩn (i.i.d)	$\frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$



Phần 3.

Ôn tập: Phân Tích Mô Tả

3.1. Phân tích Mô tả đơn biến

3.2. Phân tích Mô tả đa biến

Các loại phân tích mô tả từ cổ điển đến nâng cao:

(Tham khảo: <https://uc-r.github.io/descriptive>)

1. Phân tích kiểu cổ điển (classical)

- Phân tích số liệu đơn biến (descriptive statistics)
- Phân tích số liệu đa biến (descriptive statistics)
- Phân tích số liệu với thống kê suy luận (inferential statistics)

2. Khai thác dữ liệu text (text mining)

- Unstructured information extracting
- Sentiment analysis

3. Giải thuật học không giám sát (unsupervised learning)

- Principal Component Analysis
- Trend analysis
- Cluster analysis (k-Means, kNNs)

3.1 Phân tích mô tả đơn biến

1. Measures of frequency: Number of Occurrences, Percentage
2. Measures of central tendency: Mean, Median, Mode
3. Measures of spread (dispersion/variability):
Quartiles, Variance & Standard Deviation
4. Measures of position: Percentiles & Quantiles, Standard Scores
5. Measures of shape: Skewness/Kurtosis, Normal Distribution

3.2 Phân tích mô tả đa biến

Các loại phân tích mô tả số liệu đa biến (multivariate analysis) cơ bản:

- ❖ Covariance (hiệp phương sai)
- ❖ Correlation & Coefficient (sự tương quan và hệ số tương quan)

THANK YOU!

