**ĐẠI HỌC KINH TẾ QUỐC DÂN**
**KHOA TOÁN KINH TẾ**
**Bộ môn Toán kinh tế**

# Bài giảng
# LÝ THUYẾT XÁC SUẤT
# &
# THỐNG KÊ TOÁN

**www.mfe.neu.edu.vn**

**2021**

# Thông tin học phần

- Tiếng Anh: ***Probability and Mathematical Statistics***
- Số tín chỉ: 3 Thời lượng: 45 tiết
- Thông tin chi tiết về Giảng dạy và học tập học phần: www.mfe.edu.vn

⇨ Cần biết

⇨ "***Hướng dẫn giảng dạy học tập học phần Lý thuyết xác suất và Thống kê toán***" gồm có:

- Đề cương chi tiết
- Slide bài giảng
- Bảng số và công thức cơ bản (được dùng khi thi)
- Hướng dẫn thực hành Excel + Số liệu
- Bài tập và câu hỏi trắc nghiệm tham khảo
- Nội dung giảng dạy học tập cụ thể

# Đánh giá

- Giảng viên đánh giá: Điểm (1), (2), (3)
  - (1) Chuyên cần: 10%
  - (2) Bài kiểm tra 1: 20%, bài 1 – bài 5
  - (3) Bài kiểm tra 2: 20%, bài 7 – bài 9, có Excel
- Thi cuối kì trắc nghiệm trên máy tính, 40 câu, 60 phút
  - (4) Thi cuối kì: 50%
- Nghỉ trên 20% số giờ giảng trên lớp không được thi

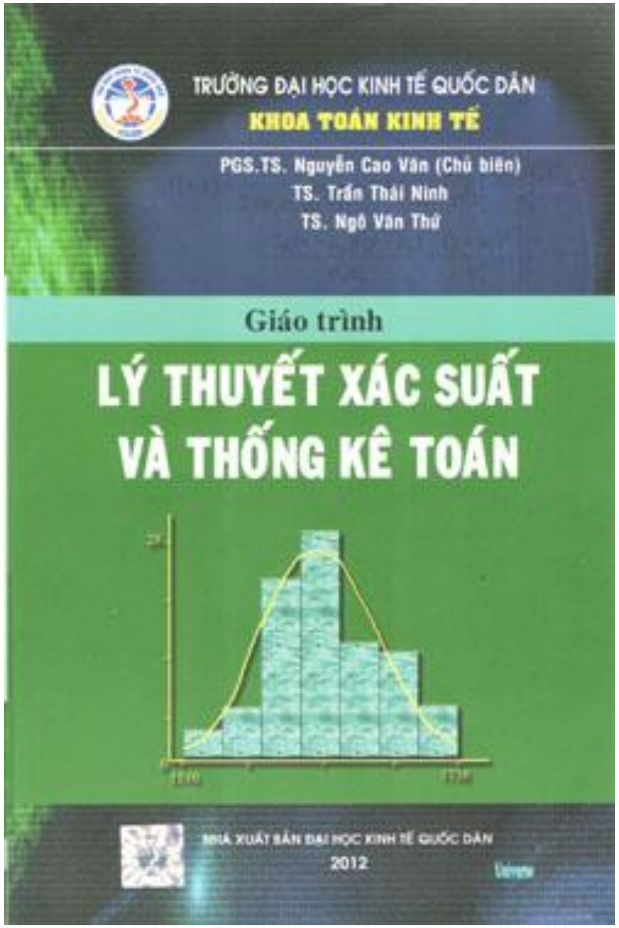
# Tài liệu


[1] Nguyễn Cao Văn, Trần Thái Ninh, Ngô Văn Thứ (2018), ***Giáo trình Lý thuyết xác suất và Thống kê toán***, NXB ĐHKTQD.

[2] Bùi Dương Hải (2016), ***Tài liệu hướng dẫn thực hành Excel***, Lưu hành hội bộ.

[3] Paul Newbold, William L. Carlson, Betty Thorne (2017), ***Statistics for Business and Economics***, 8th edition, Pearson.


- Các tài liệu tham khảo khác
- Website: [www.mfe.neu.edu.vn](http://www.mfe.neu.edu.vn)



# Nhà khoa học

- **Thế kỉ 16**: Galilei O Galile (*Italia*)
- **Thế kỉ 17**: Blaise Pascal, Piere de Fermat (*Pháp*), Christian Huygens (*Hà Lan*), Jakob Bernoulli (*Thụy Sĩ*)
- **Thế kỉ 18**: Nicolaus Bernoulli (*Thụy Sĩ*), Thomas Bayes (*Anh*), Pierre Simon Laplace (*Pháp*)
- **Thế kỉ 19**: Carl Friedrich Gauss (*Đức*), Simeon Denis Poisson (*Pháp*), Pafuni Chebyshev (*Nga*), Francis Galton, Karl Pearson (*Anh*)
- **Thế kỉ 20**: Charles Spearman, Royal Aylmer Fisher (*Anh*), Andrei Kolmogorov (*Nga*)

# NỘI DUNG

- **Bài 1: Mở đầu**
- **Bài 2: Biến cố và Xác suất**
- **Bài 3: Biến ngẫu nhiên rời rạc và phân phối xác suất**
- **Bài 4: Biến ngẫu nhiên liên tục và phân phối xác suất**
- **Bài 5: Biến ngẫu nhiên hai chiều**
- **Bài 6: Luật số lớn**
- **Bài 7: Mẫu ngẫu nhiên**
- **Bài 8: Ước lượng tham số**
- **Bài 9: Kiểm định giả thuyết**

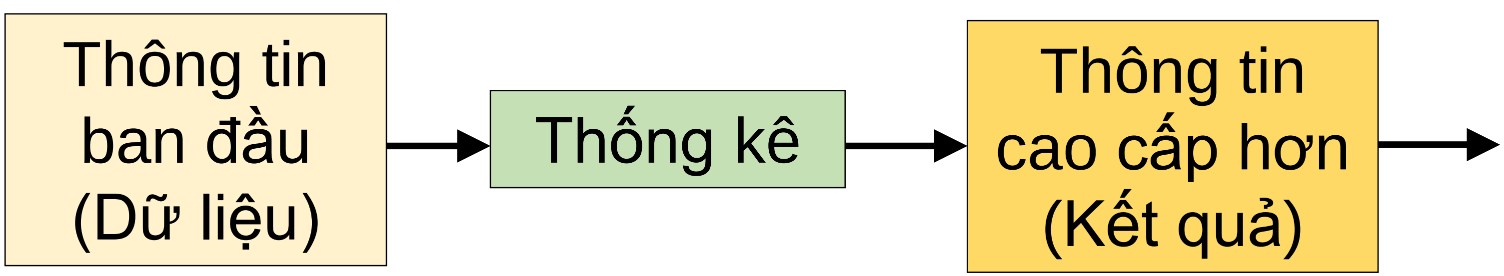
# BÀI 1 – MỞ ĐẦU

- **1.1. Các khái niệm cơ bản**
  - Thống kê – Xác suất
  - Tổng thể, mẫu, quan sát, biến
- **1.2. Bảng biểu – Đồ thị**
  - Bảng tần số, tần suất
  - Đồ thị tròn, cột, phân phối giá trị
- **1.3. Thống kê mô tả**
  - Nhóm xu thế trung tâm
  - Nhóm đo độ phân tán
  - Nhóm hình dáng phân phối, mối quan hệ

# 1.1. CÁC KHÁI NIỆM CƠ BẢN

THỐNG KÊ (*STATISTICS*)

- Thu thập dữ liệu
- Xử lý dữ liệu
- Trình bày, biểu diễn dữ liệu
- Phân tích dữ liệu để có được thông tin ở mức cao hơn
- Suy diễn về thông tin



# Thống kê và Xác suất

- Thống kê có hai nhánh chính
- **Thống kê mô tả (*Descriptive Statistics*)**: sắp xếp, tổng hợp, trình bày dữ liệu theo những cách hợp lý, thuận tiện nhất.
- **Thống kê suy diễn (*Inferential Statistics*)**: dự đoán, kiểm chứng, phân tích dữ liệu để có các kết luận tổng quát.
- Kết nối là **Lý thuyết xác suất (*Probability Theory*)**
- Thống kê mô tả là thông tin cơ bản cho thống kê suy diễn

# Tổng thể và Mẫu

- **Tổng thể (*Population*)**: tất cả các phần tử cần quan tâm
  - Kích thước tổng thể: $N$, có thể vô hạn
  - Giá trị tính từ tổng thể: **Tham số (*parameter*)**
- **Mẫu (*Sample*)**: tập con rút ra từ tổng thể
  - Kích thước mẫu: $n$, hữu hạn
  - Giá trị tính từ mẫu: ***Thống kê* (*statistic*)**

# Cấu trúc dữ liệu truyền thống

- Gồm: Quan sát – Biến – Giá trị
- (*Observation* – *Variable* – *Value*)

| | | Biến | | | | | |
|---|---|---|---|---|---|---|---|
| | TT | Họ tên | Giới | Tuổi | Điểm T.Anh | Điểm Toán | … |
| Quan sát | *1* | Nguyễn A | M | 19 | A | 8 | … |
| | *2* | Trần T. | F | 20 | C | 9 | … |
| | *3* | Lê …. | M | 20 | B | 7 | |
| | … | … | … | … | … | …. | …. |

# Phân loại biến

- **Định tính (*Qualitative*)** và **Định lượng (*Quantitative*)**
- **Biến định tính**: Định danh và Thứ bậc
  - **Định danh** (*Nominal*): VD: Tên, địa chỉ, ngành học
  - **Thứ bậc** (*Ordinal*): VD: Thứ hạng, cỡ giày,...
  - Riêng: **Nhị phân** (*Binary*): Đúng / Sai, Nam / Nữ...
- **Biến Định lượng** có đơn vị đo lường
  - **Rời rạc** (*Discrete*): VD: tuổi, số buổi học,...
  - **Liên tục** (*Continuous*): VD: thời gian, cân nặng
  - Biến định lượng có thể sử dụng để xác định biến định tính

# Phân loại biến

| Định tính | | Định lượng | |
|---|---|---|---|
| Định danh | Thứ bậc | Rời rạc | Liên tục |
| Liệt kê<br>Nhóm, gộp | Liệt kê, nhóm<br>Sắp xếp thứ tự, so sánh | Liệt kê, nhóm, gộp<br>Sắp xếp thứ tự, so sánh<br>Tính toán, +, -, × , ÷ ,… | |

Mã hóa bởi con số

Sử dụng để xếp hạng

# Thang đo Likert

- Sử dụng trong các bảng hỏi đánh giá, nhận xét
- Thang Likert 5 bậc, 7 bậc

| Rất không đồng ý | Không đồng ý | Bình thường | Đồng ý | Rất đồng ý |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| Hoàn toàn không đồng ý | Rất không đồng ý | Không đồng ý | Bình thường / Phân vân | Đồng ý | Rất đồng ý | Hoàn toàn đồng ý |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

# 1.2. BẢNG BIỂU & ĐỒ THỊ

- Bảng tần số, tần suất, tần suất tích lũy
- Đồ thị tròn (*pie chart*), cột (*column chart*, *bar chart*)
- Đồ thị phân phối giá trị (*histogram*)
- Đồ thị rải điểm (*scatter plot*)

[1] Chương 6, trang 312 – 323.

[2] Mục 4,5, trang 27 – 45.

[3] Chapter 1, pp. 21 – 59.

# Ví dụ 1.1. Dữ liệu VHLSS 2012

- Dữ liệu hộ gia đình ở Hà Nội, **n = 420** quan sát

| STT | Khu vực | Số người | Thu nhập (triệu VND) | TNBQ/người so với toàn quốc |
|---|---|---|---|---|
| 1 | Thành thị | 3 | 130,8 | Cao |
| 2 | Thành thị | 5 | 133,1 | TB cao |
| 3 | Nông thôn | 4 | 104,3 | TB thấp |
| … | | | | |
| 420 | Nông thôn | 7 | 25,7 | Thấp |

# Tần số & Tần suất (tỷ lệ)

- Bảng tần số (*frequency table*) của biến Khu vực

| Khu vực | Thành thị | Nông thôn |
|---|---|---|
| Tần số | 183 | 237 |

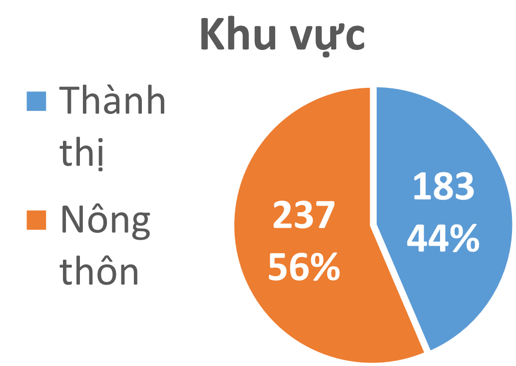
- Bảng tần suất (*relative frequency*) hay bảng tỉ lệ (*proportion*)

$$\text{Tần suất} = \frac{\text{Tần số}}{\text{Tổng số phần tử}}$$

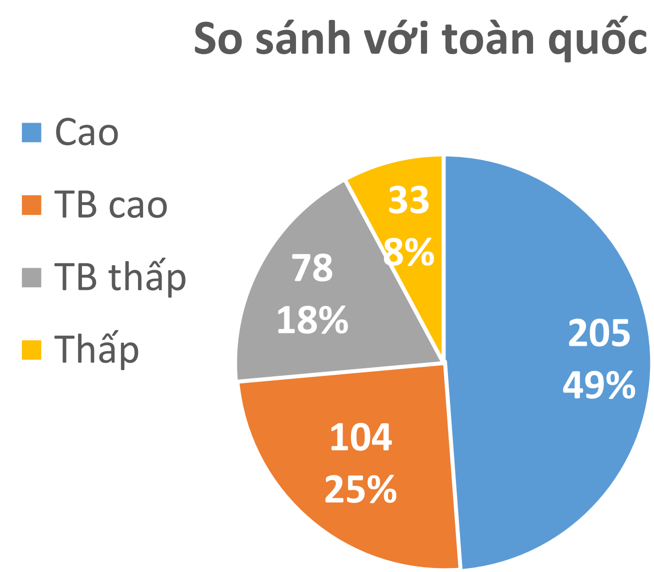
| Khu vực | Thành thị | Nông thôn |
|---|---|---|
| Tần suất<br>Tỉ lệ phần trăm | 0,436<br>43,6 % | 0,564<br>56,4 % |

# Đồ thị tròn (*pie chart*)

| Khu vực | Thành thị | Nông thôn |
|---|---|---|
| Tần số | 183 | 237 |
| Tỷ lệ | 43,6% | 56,4% |

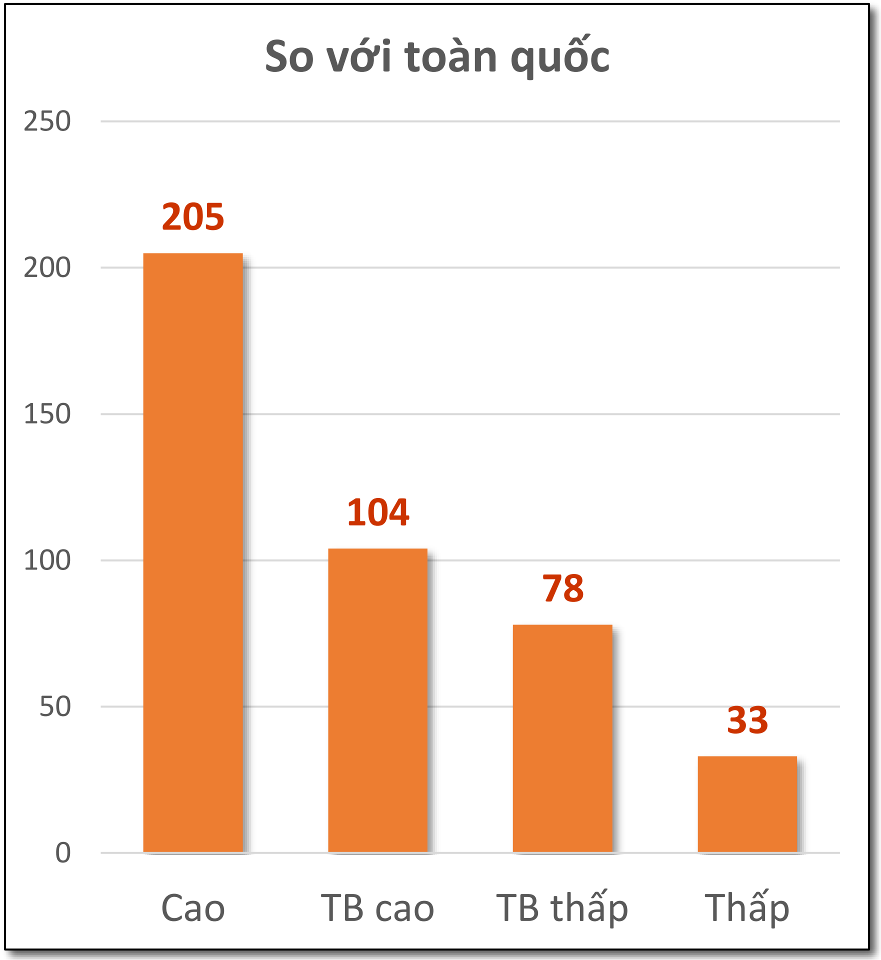



| TNBQ/người so với toàn quốc | Cao | TB cao | TB thấp | Thấp |
|---|---|---|---|---|
| Tần số | 205 | 104 | 78 | 33 |
| % | 49% | 25% | 19% | 8% |



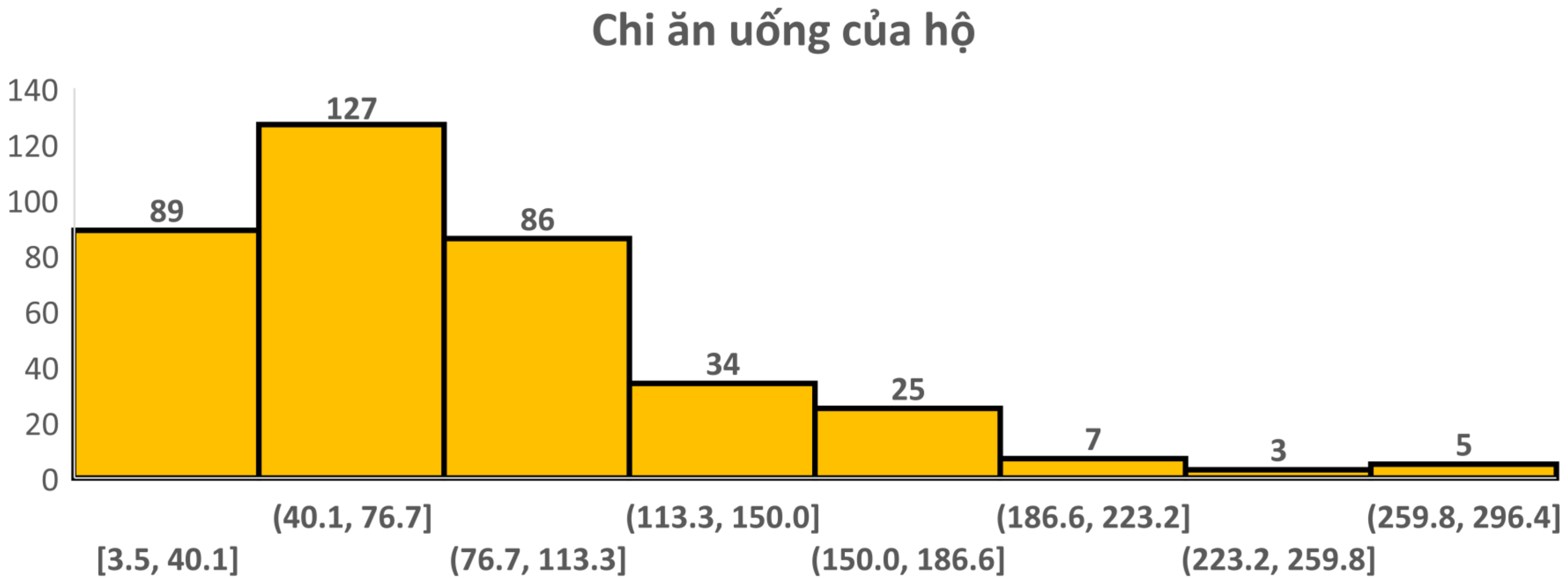
# Đồ thị cột biến định tính (*column chart*)

| TNBQ/người so với toàn quốc | Tần số | Tỉ lệ |
|---|---|---|
| Cao | 205 | 49% |
| TB cao | 104 | 25% |
| TB thấp | 78 | 19% |
| Thấp | 33 | 8% |
| ***Tổng*** | ***420*** | ***100%*** |



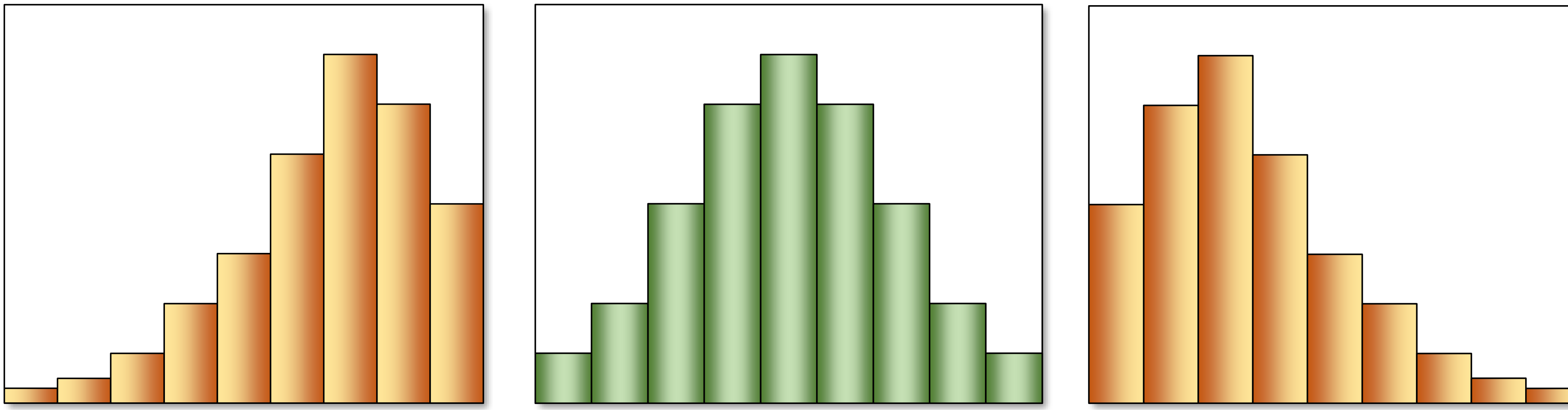
# Đồ thị cột biến định lượng (*histogram*)

| Chi ăn uống (trđ.) | 3,5 - 40,1 | 40,1 - 76,7 | 76,7 - 113,3 | 113,3 - 150,0 | 150,0 - 186,6 | 186,6 - 223,2 | 223,2 - 259,8 | 259,8 - 296,4 | Tổng |
|---|---|---|---|---|---|---|---|---|---|
| **Tần số** | 89 | 127 | 86 | 34 | 25 | 7 | 3 | 5 | ***376*** |
| **Tỉ lệ** | 23,67 | 33,78 | 22,87 | 9,04 | 6,65 | 1,86 | 0,80 | 1,33 | ***100*** |



# Hình dạng của phân phối giá trị

- Phân phối đối xứng và bất đối xứng



Lệch trái (lệch âm)
*Negatively skewed*
*Left skewed*

Đối xứng,
dạng chuông
(*Symmertrical*)

Lệch phải (lệch dương)
*Positively skewed*
*Right skewed*

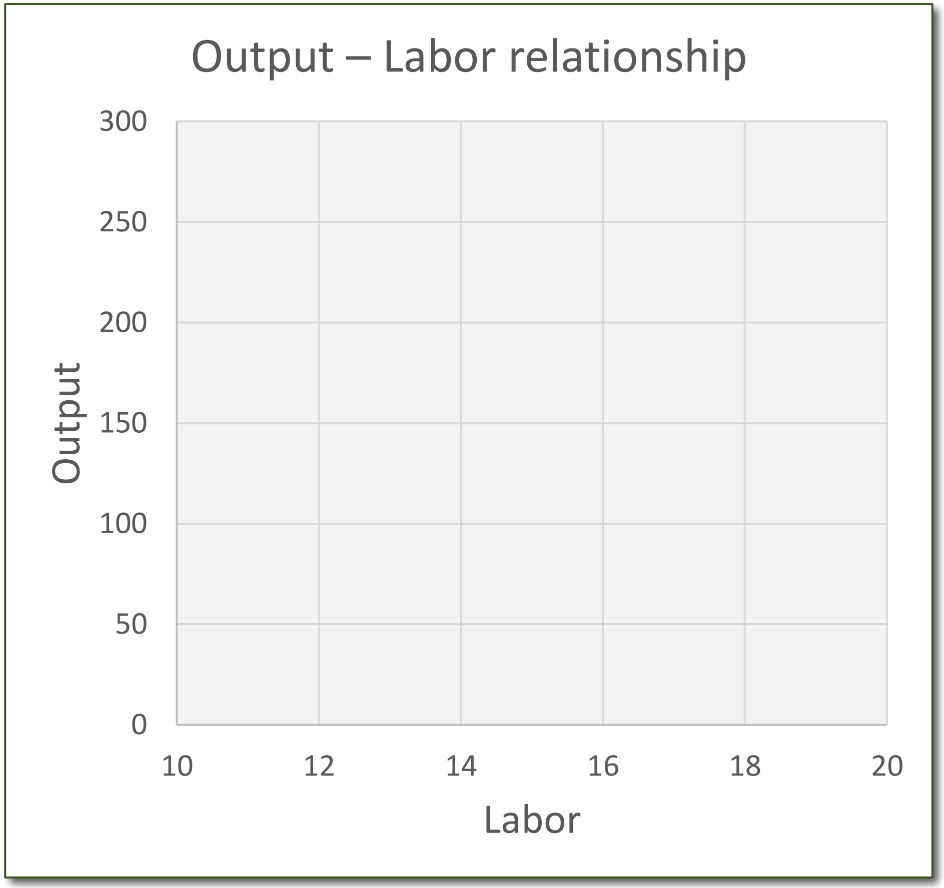
Bất đối xứng / phân phối lệch

# Đồ thị rải điểm (*scatter plot*)

- Sử dụng với số liệu hai chiều

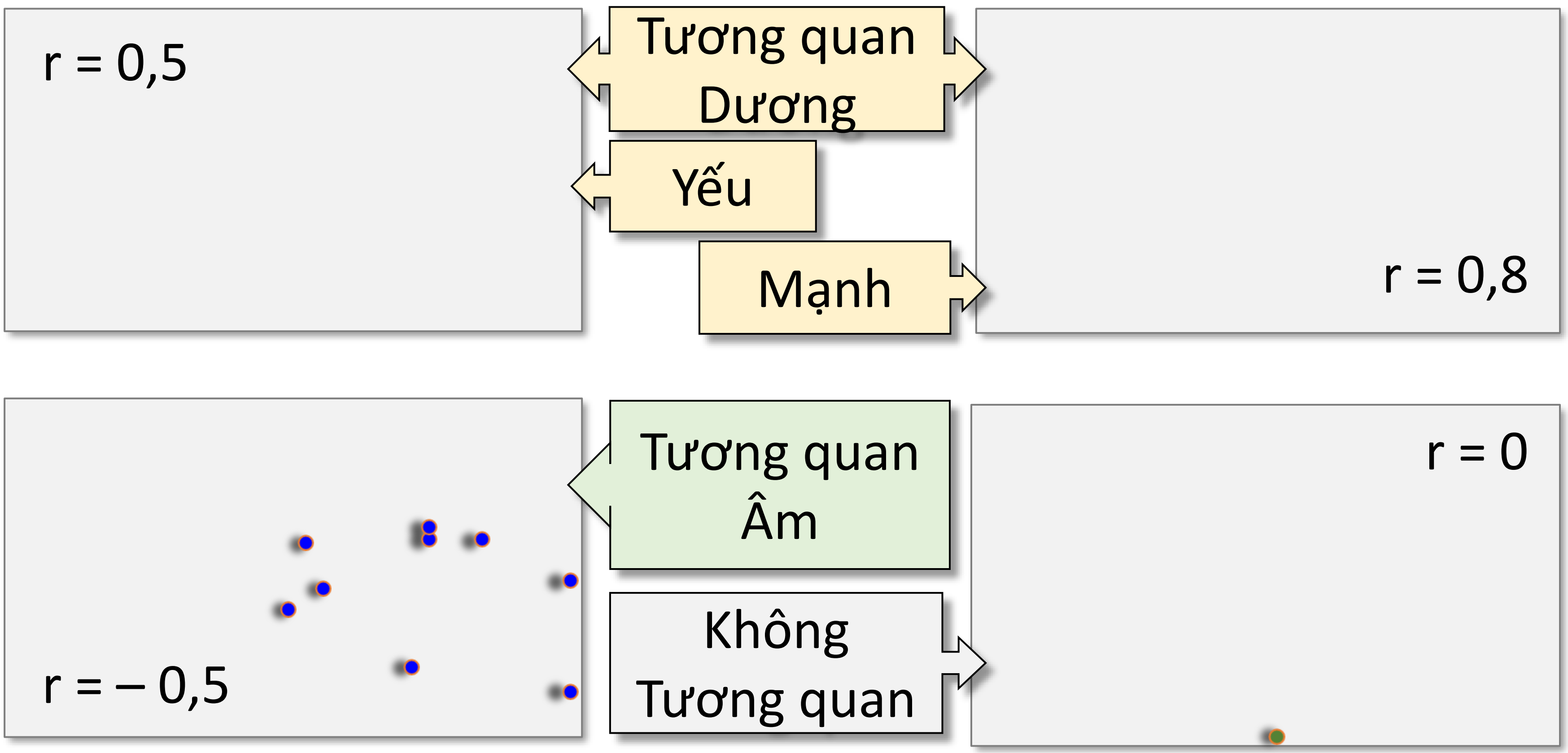
| Labor | Output |
|---|---|
| 11 | 80 |
| 11 | 130 |
| 12 | 150 |
| 13 | 110 |
| 13 | 150 |
| 13 | 200 |
| 15 | 170 |
| 14 | 180 |

| Labor | Output |
|---|---|
| 15 | 250 |
| 16 | 220 |
| 17 | 210 |
| 18 | 240 |
| 18 | 200 |
| 17 | 260 |
| 19 | 240 |
| 19 | 280 |



# Tương quan giữa hai biến định lượng

- Dạng cơ bản của đồ thị điểm hai biến định lượng



# 1.3. THỐNG KÊ MÔ TẢ BẰNG SỐ

- Xu thế trung tâm: Trung bình, trung vị, mốt, ...
- Các vị trí: Tứ phân vị, các phân vị, ...
- Đo độ phân tán: Phương sai, độ lệch chuẩn, hệ số biến thiên, khoảng biến thiên, khoảng tứ phân vị, ...
- Hình dáng phân phối: Hệ số bất đối xứng, hệ số nhọn
- Đo độ liên hệ: Hiệp phương sai, hệ số tương quan, ...
- [1] Chương 6, trang 312 – 323.
- [2] Mục 3, trang 14– 26.
- [3] Chapter 1, pp. 59 – 92.

# Nhóm xu thế trung tâm

- Dùng một giá trị đại diện cho bộ số liệu
- Ba thống kê xu thế trung tâm thường dùng:
  - Trung bình (*mean*)
  - Trung vị (*median*)
  - Mốt (*mode*)

# Trung bình (*mean*)

- Công thức tính

| Tổng thể | Mẫu |
|---|---|
| Dữ liệu: $\{x_1, x_2, \dots, x_N\}$ | Dữ liệu: $\{x_1, x_2, \dots, x_n\}$ |
| $\boldsymbol{\mu} = \dfrac{\boldsymbol{x_1 + x_2 + \cdots + x_N}}{\boldsymbol{N}}$ | $\overline{\boldsymbol{x}} = \dfrac{\boldsymbol{x_1 + x_2 + \cdots + x_n}}{\boldsymbol{n}}$ |

- Dùng cho biến định lượng
- Có cùng đơn vị với biến được mô tả trong bộ dữ liệu

# Trung bình có trọng số

- Nếu số liệu dạng có trọng số (*weighted data*)
- Giá trị $x_i$ có trọng số là $w_i$
- Trung bình có trọng số (*weighted mean*):

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{w_1 x_1 + w_2 x_2 + \cdots + w_k x_k}}{\boldsymbol{w_1 + w_2 + \cdots + w_k}} = \frac{\sum \boldsymbol{w_i x_i}}{\sum \boldsymbol{w_i}}$$

- **Ví dụ 1.2.**

| Giá bán (triệu / kg) | 10 | 12 | 14 | 16 |
|---|---|---|---|---|
| Số lượng (kg) | 3 | 10 | 5 | 2 |

# Trung vị (*median*) và Mốt (*mode*)

- **Trung vị:** là giá trị trung tâm của bộ dữ liệu khi dữ liệu được sắp xếp từ nhỏ nhất đến lớn nhất
  - Khi $n$ lẻ, trung vị nhận giá trị tại quan sát thứ $\frac{n+1}{2}$
  - Khi $n$ chẵn, trung vị nhận giá trị là trung bình cộng của 2 giá trị đứng giữa.
  - Ví dụ với dữ liệu: a) 5, 2, 7, 4, 2 $\rightarrow$ 2, 2, 4, 5, 7 $\rightarrow$ **4**

    b) 3, 8, 12, 14 $\rightarrow$ 3, 8, 12, 14 $\rightarrow$ **10**
- **Mốt:** Là giá trị xuất hiện ít nhất 2 lần và xuất hiện thường xuyên nhất trong bộ dữ liệu
  - Ví dụ tìm Mốt của số lần mua hàng: 5, 2, 7, 4, 2

$\rightarrow$ Mốt = **2**

# Các vị trí

- **Tứ phân vị (Quartile):** $Q_1, Q_2, Q_3$ chia bộ dữ liệu thành 4 phần với số lượng phần tử bằng nhau
  Tứ phân vị thứ hai chính là trung vị
- **Bách phân vị (Percentile):** $P_1, P_2, \ldots, P_{99}$ chia bộ dữ liệu thành 100 phần với số lượng phần tử bằng nhau
- Mở rộng có các phân vị (quantile) mức $\beta$

# Nhóm thống kê đo độ phân tán

Đo độ phân tán hoặc mức độ “đồng đều” (*homogeneity*) của biến định lượng

- Khoảng biến thiên (*Range*)
- Phương sai (*Variance*)
- Độ lệch chuẩn (*Standard Deviation*)
- Hệ số biến thiên (*Coefficient of Variation*)
- Khoảng tứ phân vị (*Interquartile Range*)

# Khoảng biến thiên

- Thống kê đơn giản nhất đo độ biến động của bộ dữ liệu là khoảng biến thiên:
- Range = Giá trị lớn nhất – Giá trị nhỏ nhất
- **Ví dụ 1.3.**

  Xác định khoảng biến thiên cho thu nhập của 5 lao động:

  12, 10, 8, 20, 21

# Phương sai, độ lệch chuẩn, hệ số biến thiên

| | Tổng thể | Mẫu |
|---|---|---|
| Số liệu | $\{x_1, x_2, \dots, x_N\}$ | $\{x_1, x_2, \dots, x_n\}$ |
| Trung bình | $\boldsymbol{\mu = \frac{\sum x_i}{N}}$ | $\boldsymbol{\bar{x} = \frac{\sum x_i}{n}}$ |
| Tổng bình phương | $SS = \sum (x_i - \mu)^2$ | $SS = \sum (x_i - \bar{x})^2$ |
| **Phương sai** | $\boldsymbol{\sigma^2} = \frac{SS}{\boldsymbol{N}}$ | $\boldsymbol{s^2} = \frac{SS}{\boldsymbol{n-1}}$ |
| **Độ lệch chuẩn** | $\boldsymbol{\sigma} = \sqrt{\sigma^2}$ | $\boldsymbol{s} = \sqrt{s^2}$ |
| **Hệ số biến thiên** | $\boldsymbol{CV} = \frac{\sigma}{\mu} \times 100\%$ | $\boldsymbol{CV} = \frac{s}{\bar{x}} \times 100\%$ |

# Phương sai, độ lệch chuẩn, hệ số biến thiên

- Phương sai và độ lệch chuẩn đo mức độ phân tán, biến động, dao động, đồng đều, ổn định của dữ liệu
- Phương sai lớn hơn $\rightarrow$ phân tán hơn, biến động hơn
  Phương sai nhỏ hơn $\rightarrow$ đồng đều hơn, ổn định hơn
- Phương sai có đơn vị là bình phương đơn vị của dữ liệu
- Độ lệch chuẩn có đơn vị là đơn vị của dữ liệu
- Hệ số biến thiên có đơn vị là phần trăm (%).
- Hệ số biến thiên đo độ biến động tương đối, để so sánh độ phân tán của các biến có đơn vị đo khác nhau.

# Phương sai, độ lệch chuẩn, hệ số biến thiên

- **Ví dụ 1.4.** Tính phương sai, độ lệch chuẩn, hệ số biến thiên của:

| | **Tổng thể** | **Mẫu** |
|---|---|---|
| Số liệu | 10, 12, 15, 10, 12, 13, 18, 19, 14, 15 | 10, 12, 14, 15, 19 |
| Trung bình | $\boldsymbol{\mu} =$ | $\overline{\boldsymbol{x}} =$ |
| Tổng bình phương | $SS =$ | $SS =$ |
| Phương sai | $\boldsymbol{\sigma^2} =$ | $\boldsymbol{s^2} =$ |
| Độ lệch chuẩn | $\sigma =$ | $s =$ |
| Hệ số biến thiên | $CV =$ | $CV =$ |

# Trung bình, phương sai và độ lệch chuẩn

Hàm tính toán trên EXCEL

| | **Tổng thể** | **Mẫu** |
|---|---|---|
| Trung bình | = Average (Data) | = Average (Data) |
| Trung vị | = Median (Data) | = Median (Data) |
| Mode | = Mode (Data) | = Mode (Data) |
| Phương sai | = Var.P (Data) | = Var.S (Data) |
| Độ lệch chuẩn | = Stdev.P (Data) | = Stdev.S (Data) |

# Sử dụng máy tính Casio fx 570VN plus

- Bật tần số:

  Shift → Setup → ↓ → 4 → 1 (turn on frequency)
- Nhập dữ liệu

  Mode → 3 → 1 → (nhập dữ liệu cho $X$ và tần số)

  Xác nhận dữ liệu với =, kết thúc nhập với AC
- Tính thống kê cho $X$

  Shift → 1 → 4 → chọn thống kê cần tính

| | |
|---|---|
| 1 $n$ = Số quan sát | 2 $\bar{x}$ = Trung bình |
| 3 $\sigma x$ = Độ lệch chuẩn tổng thể | 4 $sx$ = Độ lệch chuẩn mẫu |

# Khoảng tứ phân vị

- Khoảng tứ phân vị: (*interquartile range* - $IQR$)

$$IQR = Q_3 - Q_1$$

- $IQR$ dùng để đánh giá độ phân tán của bộ dữ liệu
- Sử dụng khoảng $(Q_1 - 1{,}5 \times IQR\,; Q_3 + 1{,}5 \times IQR)$ là tiêu chuẩn để xác định giá trị ngoại lai (*outlier*)

# Phân vị và tứ phân vị

Hàm tính toán trên EXCEL

| | Chứa min và max | Không chứa min và max |
|---|---|---|
| Phân vị $\beta$ | = Percentile.exc (Data, $\beta$) | = Percentile.inc (Data, $\beta$) |
| Min | = Min (Data) | |
| Q1 | = Quartile.exc (Data,1) | = Quartile.inc (Data,1) |
| Q2 | = Quartile.exc (Data,2) | = Quartile.inc (Data,2) |
| Q3 | = Quartile.exc (Data,3) | = Quartile.inc (Data,3) |
| Max | = Max (Data) | |

# Giá trị chuẩn hóa

- Giá trị chuẩn hóa (*Z-score*)

$$z_i = \frac{x_i - \text{Trung bình}}{\text{Độ lệch chuẩn}}$$

- *Z-score* có trung bình bằng 0, phương sai bằng 1
- **Ví dụ 1.5.** So sánh điểm số của 1 sinh viên với hai môn học qua giá trị chuẩn hóa

| | Vi mô | Vĩ mô |
|---|---|---|
| Điểm | 8 | 9 |
| Trung bình cả lớp | 7 | 7,5 |
| Độ lệch chuẩn cả lớp | 0.5 | 1,5 |

# Thống kê mô tả hình dáng phân phối

| Hệ số bất đối xứng mẫu | Hệ số nhọn mẫu |
|---|---|
| Skewness | Kurtosis |
| $a_3 = \dfrac{\sum(x_i - \bar{x})^3/n}{s^3}$ | $a_4 = \dfrac{\sum(x_i - \bar{x})^4/n}{s^4}$ |

- Skewness > 0: lệch phải, đuôi phải
- Skewness < 0: lệch trái, đuôi trái
- Hàm Excel: **Skew(data)** và **Kurtoris(data)**
- Excel hiệu chỉnh

$$a_3^{Excel} = \frac{n\sum(x_i-\bar{x})^3}{(n-1)(n-2)s^3}; a_4^{Excel} = \frac{n(n+1)\sum(x_i-\bar{x})^4}{(n-1)(n-2)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

# Thống kê mô tả mức độ liên hệ

- Hai biến $X$ và $Y$ là số liệu theo cặp $(x_i, y_i)$

| | Tổng thể | Mẫu |
|---|---|---|
| Hiệp phương sai Covariance Cov = | $\frac{\sum(x_i-\mu_X)(y_i-\mu_Y)}{N}$ | $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$ $=\frac{n}{n-1}\left(\overline{x\cdot y}-\bar{x}\cdot\bar{y}\right)$ |
| Hệ số tương quan Correlation Cor = | $\rho_{X,Y}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ | $r_{X,Y}=\frac{Cov(X,Y)}{s_Xs_Y}$ |

# Thống kê mô tả mức độ liên hệ

- Hệ số tương quan nằm trong đoạn $[-1,1]$
- Hệ số tương quan càng gần 0: càng yếu, lỏng
- Hệ số tương quan càng gần $\pm 1$: càng mạnh, chặt
- Hàm tính toán trên EXCEL

| | **Tổng thể** | **Mẫu** |
|---|---|---|
| Hiệp phương sai Covariance | = Covariance.P (Data) hoặc<br>= Covar (Data) | = Covariance.S (Data) |
| Hệ số tương quan Correlation | = Correl (Data) | |

# Hiệp phương sai - hệ số tương quan

- **Ví dụ 1.6.** Tính hiệp phương sai và hệ số tương quan của mẫu sau:

| | | | | | |
|---|---|---|---|---|---|
| $x_i$ | 2 | 5 | 8 | 7 | 9 |
| $y_i$ | 4 | 5 | 7 | 8 | 10 |

| Stt | $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i \cdot y_i$ |
|---|---|---|---|---|---|
| 1 | **2** | **4** | 4 | 16 | 8 |
| 2 | **5** | **5** | 25 | 25 | 25 |
| 3 | **8** | **7** | 64 | 49 | 56 |
| 4 | **7** | **8** | 49 | 64 | 56 |
| 5 | **9** | **10** | 81 | 100 | 90 |
| $\sum$ | 31 | 34 | 223 | 254 | 235 |

$Cov(X,Y) =$

$s_X =$

$s_Y =$

$r_{X,Y} =$

# BÀI 2 – BIẾN CỐ VÀ XÁC SUẤT

- **2.1. Phép thử và biến cố**
- **2.2. Xác suất của biến cố**
- **2.3. Các công thức xác suất**

[1] Chương 1, trang 5-77

[3] Chapter 3, pp. 93-145

# 2.1. PHÉP THỬ VÀ BIẾN CỐ

- Thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó có thể xảy ra hay không gọi là một **phép thử (experiment)**
- Hiện tượng có thể xảy ra trong kết quả phép thử → **kết cục (outcomes)**
- Kết cục không thể chia nhỏ thành các thành phần → **kết cục sơ cấp (basic outcomes)**
- Tập hợp tất cả các kết cục sơ cấp → **không gian mẫu (sample space), kí hiệu là S**
- Tập con của không gian mẫu → **biến cố (event)**

# Phép thử và biến cố

| Phép thử | Kết cục sơ cấp | Biến cố |
|---|---|---|
| Tung 1 đồng xu | Sấp, ngửa | 'sấp', 'ngửa' |
| Tung xúc xắc | 1, 2, 3, 4, 5, 6 chấm | 'nhỏ hơn 4';<br>'lẻ chấm' |
| Làm 3 câu hỏi | Đúng cả ba câu<br>Chỉ đúng câu thứ i<br>Chỉ sai câu thứ i<br>Sai cả ba câu<br>(i = 1, 2, 3) | 'đúng một câu';<br>'có làm sai';<br>… |

# Phép thử và biến cố

- **Phân loại biến cố**:
  - Biến cố chắc chắn (*certain*): $\Omega$
  - Biến cố không thể có (*impossible*): $\emptyset$
  - Biến cố ngẫu nhiên (*random*): $A, B, \ldots$ hay $A_1, A_2, \ldots$
- **Ví dụ:** Tung 1 con xúc xắc 1 lần
  - Biến cố chắc chắn: $\Omega =$ "xuất hiện số chấm < 7"
  - Biến cố không thể có: $\emptyset =$ "xuất hiện 7 chấm"
  - Biến cố ngẫu nhiên: $A_1 =$ "xuất hiện mặt 1 chấm"

    $A_2 =$ "xuất hiện mặt 2 chấm"

    $\mathrm{B} =$ "xuất hiện lẻ chấm"

# 2.2. XÁC SUẤT CỦA BIẾN CỐ

- Khái niệm xác suất
- Định nghĩa cổ điển về xác suất
- Định nghĩa thống kê về xác suất
- Nguyên lý xác suất lớn và  nguyên lý xác suất nhỏ

[1] Chương 1, trang 5-77

[3] Chapter 3, pp. 93-145

# Xác suất của biến cố

- **Xác suất (*probability*)** của một biến cố là một con số đặc trưng khả năng khách quan xuất hiện biến cố đó khi thực hiện một phép thử.
- $\boldsymbol{P(A)}$: Xác suất xảy ra biến cố $A$
- Các tính chất của xác suất:
  - $0 \leq P(\text{Biến cố}) \leq 1$
  - $\Omega$ là biến cố chắc chắn, $P(\Omega) = 1$
  - $\emptyset$ là biến cố không thể, $P(\emptyset) = 0$
  - $A$ là biến cố ngẫu nhiên, $0 < P(A) < 1$

# Xác suất có điều kiện

- Xác suất của biến cố $A$ với điều kiện biến cố $B$ đã xảy ra gọi là **xác suất có điều kiện** của $A$, hay xác suất của $A$ trong điều kiện $B$
- Ký hiệu: $\boldsymbol{P(A|B)}$

# Định nghĩa cổ điển về xác suất

- **Định nghĩa cổ điển về xác suất**
  - Giả thiết: các kết cục sơ cấp có khả năng xảy ra bằng nhau
  - Số kết cục sơ cấp thuận lợi cho biến cố $A$: $N_A$
  - Số kết cục sơ cấp của phép thử: $N$
  - Xác suất xảy ra biến cố $A$:

$$P(A) = \frac{N_A}{N}$$

# Tính xác suất theo định nghĩa cổ điển

- Để tính xác suất biến cố $A$ theo định nghĩa cổ điển cần xác định $N_A$ và $N$: có thể sử dụng các phương pháp sau
  - Suy luận trực tiếp (đếm số kết cục sơ cấp)
  - Sơ đồ
    - Cây (*tree*)
    - Bảng (*table*)
    - Venn (*subset*)
  - Tính theo đại số tổ hợp

# Đếm số kết cục sơ cấp

- **Ví dụ 2.1.** Tung một đồng xu cân đối đồng chất. Tính xác suất xuất hiện mặt sấp?

- **Ví dụ 2.2.** Tung một đồng xu hai lần liên tiếp. Tính xác suất

  a) Hai lần xuất hiện mặt sấp

  b) Xuất hiện 1 lần mặt sấp, 1 lần mặt ngửa

  c) Hai lần xuất hiện mặt ngửa

# Dùng sơ đồ cây

- **Ví dụ 2.3.** Giả sử xác suất sinh con trai – gái như nhau. Một gia đình có 3 người con. Tính xác suất:
- a) Hộ gia đình có 2 con gái
- b) Hộ gia đình có ít nhất 1 con gái
- c) Hộ gia đình có 2 con gái, biết con đầu lòng là gái
- d) Hộ gia đình có 2 con gái, biết con đầu lòng là trai

# Dùng sơ đồ Venn

- **Ví dụ 2.4.** Trong các khách hàng biết đến sản phẩm của công ty thì 30% biết từ internet, 40% biết thông qua người thân giới thiệu và 12% biết từ cả hai nguồn thông tin trên. Chọn ngẫu nhiên 1 khách hàng, tính xác suất:
- a) Người đó biết sản phẩm từ nguồn thông tin khác 2 nguồn trên
- b) Người đó biết sản phẩm từ ít nhất 1 trong 2 nguồn thông tin trên
- c) Người đó chỉ biết sản phẩm từ đúng 1 nguồn thông tin trong 2 nguồn trên.

# Tính theo đại số tổ hợp

- **Ví dụ 2.5**. Trên quầy có 6 sản phẩm từ công ty A, 4 sản phẩm từ công ty B.
- a) Lấy ngẫu nhiên cùng lúc 2 sản phẩm. Tính xác suất 2 sản phẩm đều của công ty A
- b) Lấy ngẫu nhiên cùng lúc 4 sản phẩm. Tính xác suất trong đó có 2 của công ty A và 2 của công ty B
- c) Lấy ngẫu nhiên cùng lúc 2 sản phẩm. Tính xác suất hai sản phẩm đó từ cùng một công ty

# Định nghĩa thống kê về xác suất

- **Định nghĩa thống kê về xác suất**
  - Thực hiện $n$ phép thử
  - Số lần xuất hiện biến cố $A$: $n_A$
  - Tỉ lệ xuất hiện biến cố $A$: $\frac{n_A}{n}$
  - Khi số phép thử $n$ đủ lớn: $P(A) \approx \frac{n_A}{n}$
- **Ví dụ 2.6.**
  - Trong 100.000 trẻ mới sinh, có 51.000 bé trai. Xác suất "sinh bé trai" là khoảng 0,51.

# Nguyên lý xác suất

- **Nguyên lý xác suất lớn**: Nếu biến cố ngẫu nhiên có xác suất rất lớn thì thực tế có thể cho rằng trong một phép thử biến cố đó sẽ xảy ra.
- **Nguyên lý xác suất nhỏ**: Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử, biến cố đó sẽ không xảy ra.

# 2.3. CÁC CÔNG THỨC XÁC SUẤT

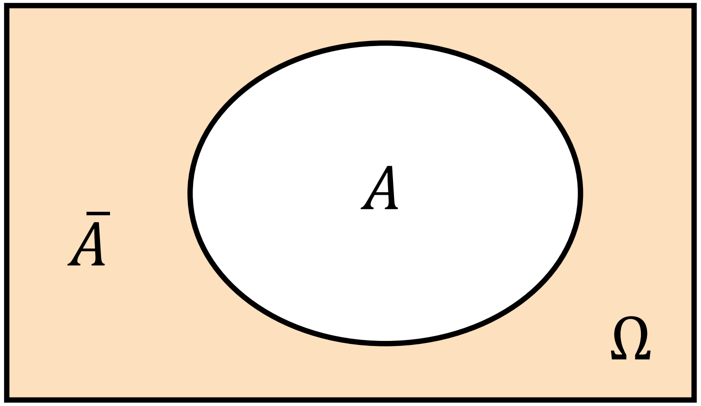
- Xác suất của biến cố tích
- Xác suất của biến cố tổng
- Công thức Bernoulli
- Công thức xác suất đầy đủ - Bayes

[1] Chương 1, trang 5-77

[3] Chapter 3, pp. 93-145

# Biến cố đối lập

- Hai **biến cố đối lập (*complement*)** với nhau khi trong kết quả của phép thử sẽ xảy ra đúng một biến cố trong hai biến cố đó

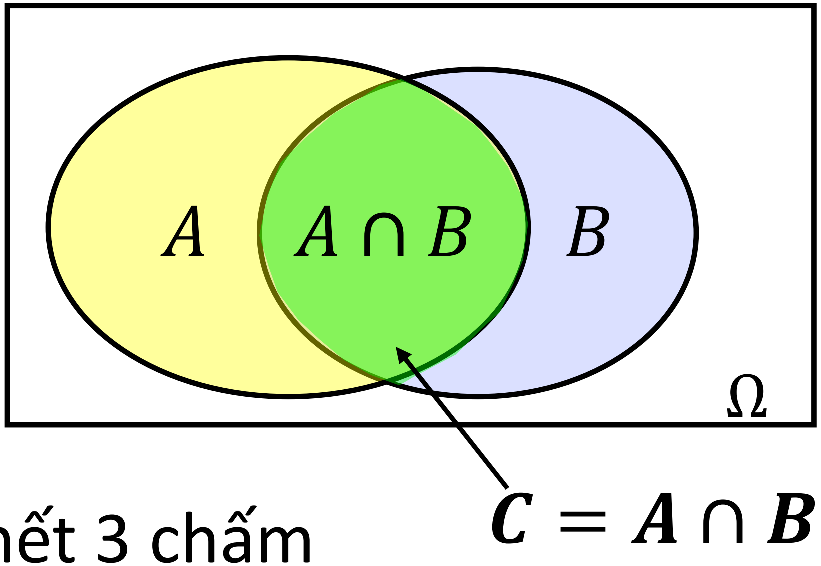



- Kí hiệu biến cố đối lập của biến cố $A$ là $\bar{A}$

- Xác suất của biến cố đối lập: $P(\bar{A}) = 1 - P(A)$

# Biến cố tích

- Biến cố $C$ là **tích** (***intersection***) của hai biến cố $A$ và $B$ nếu $C$ xảy ra khi và chỉ khi cả hai biến cố $A$ và $B$ đồng thời xảy ra.
- Ký hiệu: $C = A \cap B$




- **Ví dụ 2.7.** Tung xúc xắc
  - $A$: mặt lẻ chấm
  - $B$: mặt có số chấm chia hết 3 chấm
  - $C = A \cap B =$ "mặt có 3 chấm"

# Biến cố tích

- Biến cố $A$ là **tích của $\boldsymbol{n}$ biến cố** $A_1, A_2, \dots, A_n$ nếu $A$ xảy ra khi và chỉ khi $n$ biến cố trên đồng thời xảy ra.
- Ký hiệu $A = \bigcap_{i=1}^{n} A_i$
- **Ví dụ 2.8.** Lấy lần lượt 5 quả bóng từ trong hộp gồm 6 bóng trắng và 4 bóng đen.

  $A_i =$ "Lấy được bóng trắng ở lần lấy thứ $i$",

  $$i = 1,2, \dots 5.$$

  $A =$ "Lấy được 5 bóng trắng"

  $$A = (A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5)$$

# Hai biến cố độc lập

- Hai biến cố là **độc lập (*independent*)** nếu việc xảy ra hay không xảy ra của biến cố này không làm thay đổi xác suất xảy ra của biến cố kia và ngược lại.
- Nếu $A$ và $B$ là hai biến cố độc lập thì

$$P(B|A) = P(B|\bar{A}) = P(B)$$

$$\text{và } P(A|B) = P(A|\bar{B}) = P(A)$$

- Kiểm tra tính độc lập của 2 biến cố trong ví dụ 2.7.

# Độc lập từng đôi và độc lập toàn phần

- Các biến cố $A_1, A_2, \dots, A_n$ gọi là **độc lập từng đôi** với nhau nếu mỗi cặp hai trong *n* biến cố đó độc lập nhau.
- Các biến cố $A_1, A_2, \dots, A_n$ gọi là **độc lập toàn phần** nếu mỗi biến cố độc lập với mọi tổ hợp bất kỳ của các biến cố còn lại.
- Độc lập toàn phần $\Rightarrow$ độc lập từng đôi
- Trong ví dụ 2.8, nếu lấy lần lượt có hoàn lại thì các biến cố là độc lập toàn phần

# Xác suất của biến cố tích

- **Công thức nhân xác suất**

$$\boldsymbol{P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)}$$

- $A$ và $B$ là hai biến cố độc lập

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

- Công thức tính xác suất có điều kiện

Nếu $\mathrm{P(B)} > 0$ thì $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

Nếu $\mathrm{P(B)} = 0$ thì $P(A \mid B)$ không xác định

- Nếu $A_1, A_2, \dots A_n$ là các biến cố độc lập toàn phần thì

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n)$$

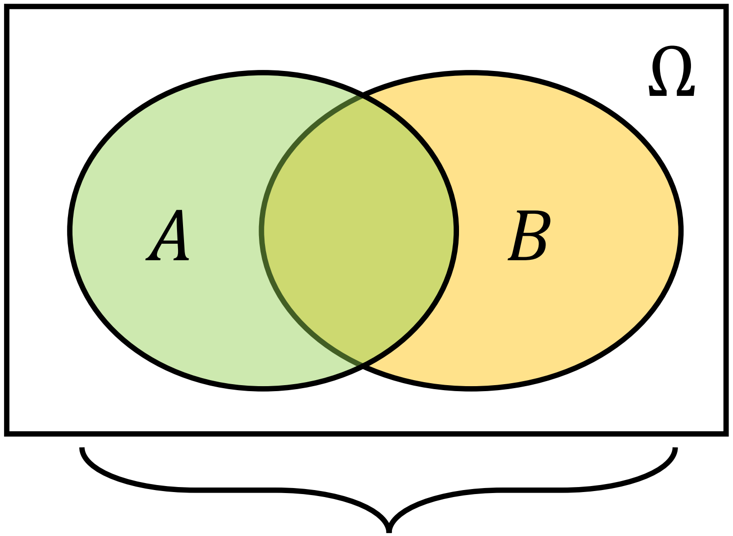
# Xác suất của biến cố tích

- **Ví dụ 2.9.** Xác suất một sinh viên thi đạt hai môn A và B lần lượt là 0,6 và 0,8. Nếu sinh viên đó thi đạt môn A thì xác suất để anh ta thi đạt môn B là 0,9.
  - a) Tính xác suất sinh viên thi đạt cả hai môn.
  - b) Các biến cố "Thi đạt môn A" và "Thi đạt môn B" có phải là các biến cố độc lập không?
  - c) Hãy tính xác suất sinh viên đó thi đạt môn A, biết rằng anh ta đã thi đạt môn B.

- **Ví dụ 2.10.** Hộp gồm 6 bóng trắng và 4 bóng đen. Lấy lần lượt có hoàn lại 5 quả bóng. Tính xác suất để lấy được 5 quả bóng màu trắng.

# Biến cố tổng

- Biến cố $D$ được gọi là **tổng (*union*)** của hai biến cố $A$ và $B$ nếu $D$ xảy ra khi và chỉ khi có ít nhất một trong hai biến cố $A$ và $B$ xảy ra.




$$\boldsymbol{D = A \cup B}$$

- Ký hiệu: $D = A \cup B$

- $A \cup B$ là tổng của các biến cố:
  $A \cap B, \bar{A} \cap B$ và $A \cap \bar{B}$
  Biểu diễn: $A \cup B = (A \cap B) \cup (\bar{A} \cap B) \cup (A \cap \bar{B})$

# Biến cố tổng

- **Ví dụ 2.11.** Lấy lần lượt 2 quả bóng từ hộp 6 bóng trắng và 4 bóng đen

  $A =$ "Lần 1 lấy được bóng trắng",

  $B =$ "Lần 2 lấy được bóng trắng"

  $A \cup B =$ "Có ít nhất 1 lần lấy được bóng trắng"

- Biến cố $A$ được gọi là **tổng của $n$ biến cố** $A_1, A_2, \dots, A_n$ nếu $A$ xảy ra khi có ít nhất một trong $n$ biến cố ấy xảy ra.
- Ký hiệu: $A = \bigcup_{i=1}^{n} A_i$

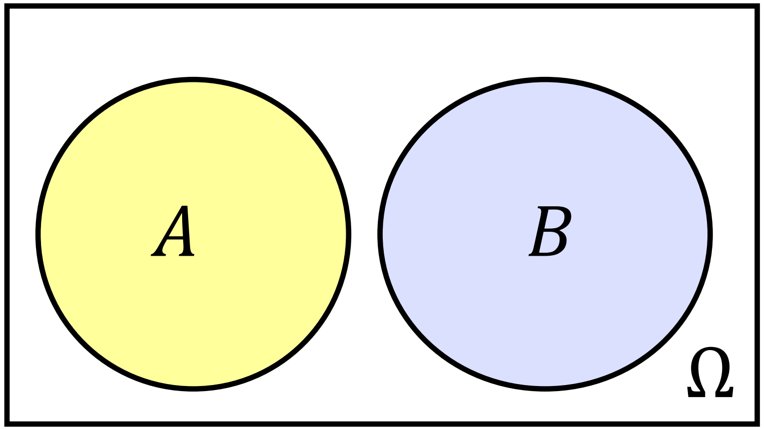
# Hai biến cố xung khắc

- Hai biến cố $A$ và $B$ gọi là **xung khắc** (***mutually exclusive***) với nhau nếu chúng không đồng thời xảy ra trong một phép thử.
- $A \cap B = \emptyset$
- Xét $A$ và $\bar{A}$ ?
- **Ví dụ 2.12.** Lấy ngẫu nhiên hai quả bóng từ trong hộp.

  $A =$ "Hai trắng", $B =$ "Hai đen"

  $C =$ "Có ít nhất một quả trắng"

  Nhận xét về tính xung khắc của các biến cố $A, B$ và $C$.



# Nhóm xung khắc từng đôi và nhóm đầy đủ

- Nhóm *n* biến cố $A_1, A_2, \dots, A_n$ được gọi là **xung khắc từng đôi** nếu bất kỳ hai biến cố nào trong nhóm này cũng xung khắc với nhau.

- Các biến cố $A_1, A_2, \dots, A_n$ được gọi là một **nhóm đầy đủ** (***universal set, partitions***) các biến cố nếu trong kết quả phép thử sẽ xảy ra một và chỉ một trong các biến cố đó.

- Nhóm đầy đủ ⇒ Nhóm xung khắc từng đôi?

# Ví dụ

- **Ví dụ 2.13.** Tung một đồng xu 2 lần

  Hãy tìm các nhóm đầy đủ.

$A_1 =$ "2 lần sấp" $A_2 =$ "2 lần ngửa"

$A_3 =$ "1 lần sấp, 1 lần ngửa" $A_4 =$ "có ít nhất 1 lần sấp"

# Xác suất của biến cố tổng

- **Công thức cộng xác suất**

$$\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

- $A$ và $B$ là hai biến cố xung khắc

$$\Rightarrow P(A \cup B) = P(A) + P(B)$$

- Nếu nhóm $A_1, A_2, \dots, A_n$ là nhóm các biến cố xung khắc từng đôi thì

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

- Nếu nhóm $A_1, A_2, \dots, A_n$ là một nhóm đầy đủ thì

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n) = 1$$

# Xác suất của biến cố tổng

- **Ví dụ 2.14.** Xác suất một sinh viên thi đạt hai môn A và B lần lượt là 0,6 và 0,8. Nếu sinh viên đó thi đạt môn A thì xác suất anh ta để thi đạt môn B là 0,9.
  - a) Đạt ít nhất một môn
  - b) Không đạt ít nhất một môn

# Xác suất của biến cố tổng

- Từ ví dụ 2.14, lập bảng xác suất đồng thời

| Bảng xác suất đồng thời | | *Môn B* Đạt $B$ | *Môn B* Không đạt $\overline{B}$ | *Tổng* |
|---|---|---|---|---|
| *Môn A* | Đạt $A$ | **0,54** | **0,06** | **0,6** |
| | Không đạt $\overline{A}$ | **0,26** | **0,14** | **0,4** |
| | *Tổng* | **0.8** | **0,2** | **1** |

- a) $P(A+B) =$
- b) $P(\bar{A}+\bar{B}) =$

# Công thức Bernoulli

**Ví dụ 2.15.** Một người đi bán hàng ở 3 nơi độc lập, xác suất bán được ở mỗi nơi đều bằng 0,8. Tính xác suất người đó:

a) Bán được ở đúng 1 nơi

b) Bán được ở đúng 2 nơi

c) Bán được ở ít nhất 1 nơi

# Công thức Bernoulli

- Thực hiện $n$ phép thử độc lập
- Mỗi phép thử, chỉ quan tâm $A$ (*success*) và $\bar{A}$ (*failure*)
- Mỗi phép thử, $P(A) = p$ không đổi, và $P(\bar{A}) = 1 - p$

$\Longrightarrow$ Lược đồ Bernoulli

- Xác suất để trong $n$ phép thử đó, biến cố $A$ xảy ra đúng $x$ lần được ký hiệu $P(x|n, p)$
- **Công thức Bernoulli**

$$P(x|n, p) = C_n^x p^x (1 - p)^{n-x}$$

$$x = 0, 1, 2, \dots, n$$

# Công thức xác suất đầy đủ - Bayes

**Ví dụ 2.16.** Hai hộp giống nhau:

Hộp loại I chứa 6 chính phẩm và 4 phế phẩm;

Hộp loại II chứa 8 chính phẩm và 2 phế phẩm.

Chọn ngẫu nhiên một hộp và từ đó chọn 1 sản phẩm.

a) Tính xác suất để đó là chính phẩm.

b) Chọn được chính phẩm, XS hộp được chọn là hộp loại I? là hộp loại II? Quyết định mua hàng (mua hay không mua) khi thử được chính phẩm?

c) Chọn được phế phẩm, XS hộp được chọn là hộp loại I? là hộp loại II? Quyết định mua hàng (mua hay không mua) khi thử được phế phẩm? So sánh với (b)

# Công thức xác suất đầy đủ - Bayes

- $H_1, H_2, \dots, H_n$ : nhóm đầy đủ các biến cố.
- $A$: xảy ra đồng thời với một trong các biến cố $H_1, H_2, \dots, H_n$
- **Công thức xác suất đầy đủ:**

$$\boldsymbol{P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i)}$$

- **Công thức Bayes**

$$\boldsymbol{P(H_i|A) = \frac{P(H_i \cdot A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}}$$

# BÀI 3 – BIẾN NGẪU NHIÊN RỜI RẠC VÀ PHÂN PHỐI XÁC SUẤT

- **3.1. Khái niệm biến ngẫu nhiên**
- **3.2. Bảng phân phối xác suất**
- **3.3. Hàm phân phối xác suất**
- **3.4. Các tham số đặc trưng**
- **3.5. Phân phối Không-Một**
- **3.6. Phân phối Nhị thức**
- **3.7. Phân phối Poisson**

[1] Chương 2, trang 79 – 146; Chương 3, trang 147 – 166

[2] Mục 2, trang 5 – 9; [3] Chapter 4, pp. 146 -196

# 3.1. KHÁI NIỆM BIẾN NGẪU NHIÊN

- **Ví dụ 3.1.**
  - Tung 1 con xúc xắc. Gọi $X$ là số chấm xuất hiện

    $X$ có thể nhận $1, 2, 3, 4, 5, 6$
  - Gọi $Y$ là số khách vào siêu thị trong một ngày

    $Y$ nhận các giá trị $0, 1, 2, \dots, n, \dots$
  - Gọi $Z$ là lợi suất khi đầu tư cổ phiếu

    $Z$ nhận các giá trị trong khoảng $(-\infty; +\infty)$

# Khái niệm biến ngẫu nhiên

- **Biến ngẫu nhiên** (*random variable*) là biến số mà trong kết quả phép thử nó chỉ nhận 1 và chỉ 1 trong các giá trị có thể có của nó
- Kí hiệu: $X, Y, Z, X_1, X_2, \dots$
- Giá trị có thể có của $X$: $X = (x_1, x_2, \dots, x_n)$
- Quan tâm tới các biến cố:

  $(X = x_i); (X > a); (X < b); (a < X < b)$
- Mục đích: tính xác suất và hỗ trợ ra quyết định

# Phân loại biến ngẫu nhiên

- **Biến ngẫu nhiên rời rạc (*discrete*)**: Các giá trị có thể có là một tập hợp hữu hạn hoặc đếm được.

  Biến $X$ và $Y$ trong ví dụ 3.1

- **Biến ngẫu nhiên liên tục (*continuous*)**: Các giá trị có thể có lấp đầy một khoảng trên trục số.

  Biến $Z$ trong ví dụ 3.1

# 3.2. BẢNG PHÂN PHỐI XÁC SUẤT

- Phân phối xác suất của biến ngẫu nhiên là sự tương ứng giữa các **giá trị** có thể có và các **xác suất** tương ứng.
- Các phương pháp mô tả phân phối xác suất:
  - Bảng phân phối xác suất (*Probability distribution table*)
  - Hàm phân phối xác suất (*Probability distribution function/Cumulative distribution function - CDF*)
  - Hàm mật độ xác suất (*Probability density function –PDF*)

# Bảng phân phối xác suất

Giả sử biến ngẫu nhiên rời rạc $X$ nhận các giá trị $x_1, x_2, \dots, x_n$ với xác suất tương ứng $p_1, p_2, \dots, p_n$

$$p_i = P(X = x_i)$$

| $X$ | $x_1$ | $x_2$ | … | $x_n$ |
|---|---|---|---|---|
| $P$ | $p_1$ | $p_2$ | … | $p_n$ |

(1) $0 \le p_i \le 1$

(2) $\sum_{i=1}^{n} p_i = 1$

(3) $x_0 \notin \{x_1, x_2, \dots, x_n\} \to P(X = x_0) = 0$

(4) $P(a \le X \le b) = \sum_{a \le x_i \le b} p_i$

# Bảng phân phối xác suất

**Ví dụ 3.2.** Hộp có 6 sản phẩm tốt và 4 sản phẩm xấu. Lấy đồng thời 2 sản phẩm. Gọi $X$ số sản phẩm tốt lấy được.

Lập bảng phân phối xác suất của $X$

Gọi $X$ là số sản phẩm tốt lấy được; $X = 0, 1, 2$

| $X$ | 0 | 1 | 2 |
|---|---|---|---|
| $P$ | ? | ? | ? |

# Bảng phân phối xác suất

**Ví dụ 3.3.** Số hợp đồng kí được trong một ngày ($X$) của một công ty cung cấp dịch vụ có bảng phân phối xác suất

| $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $P$ | 0,1 | 0,4 | 0,25 | 0,2 | 0,05 |

a) Tính xác suất một ngày bất kì kí được trên 3 hợp đồng.

b) Tính xác suất để trong 5 ngày bất kì có 3 ngày kí được dưới 2 hợp đồng.

# 3.3. HÀM PHÂN PHỐI XÁC SUẤT

Hàm phân phối xác suất của biến ngẫu nhiên, kí hiệu $F(x)$, là xác suất để biến ngẫu nhiên $X$ nhận giá trị nhỏ hơn $x$, với $x$ là số thực bất kì

$$F(x) = P(X < x)$$

- **Ý nghĩa:** Hàm phân bố xác suất $F(x)$ phản ánh mức độ tập trung xác suất ở bên trái số thực $x$.

# Tính chất hàm phân phối xác suất

- *Tính chất 1*: $0 \le F(x) \le 1$
- *Tính chất 2*: $F(x)$ là hàm không giảm

$$P(a \le X < b) = F(b) - F(a)$$

- *Tính chất 3*: $F(x) = 0 \; nếu \; x \le x_{min}$

$$F(x) = 1 \; nếu \; x > x_{max}$$

# 3.4. CÁC THAM SỐ ĐẶC TRƯNG

- Kì vọng (*Expected value*)
- Kì vọng của biến ngẫu nhiên $X$, kí hiệu $E(X)$

$$\boldsymbol{E(X) = \sum_{i=1}^{n} x_i p_i}$$

- **Ví dụ 3.4.** Tính kì vọng của số hợp đồng kí được trong ngày của một công ty dịch vụ có bảng phân phối xác suất như sau:

| $X$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $P$ | 0,1 | 0,4 | 0,25 | 0,2 | 0,05 |

# Kì vọng

- Kì vọng là trung bình của biến ngẫu nhiên tính theo xác suất
- $E(X)$ có đơn vị trùng với đơn vị của $X$
- **Tính chất:**

| |
|---|
| $E(C) = C$ với $C$ là hằng số |
| $E(C + X) = C + E(X)$ |
| $E(C.X) = C.E(X)$ |
| $E(X \pm Y) = E(X) \pm E(Y)$ |
| $E(X.Y) = E(X).E(Y)$ nếu $X, Y$ độc lập |

# Phương sai

- **Ví dụ 3.5.** Tỷ suất sinh lời (%) khi đầu tư vào hai loại cổ phiếu

| **Cổ phiếu A** | **6** | **7** | **8** |
|---|---|---|---|
| Xác suất | 0,4 | 0,2 | 0,4 |

| **Cổ phiếu B** | **5** | **7** | **9** |
|---|---|---|---|
| Xác suất | 0,4 | 0,2 | 0,4 |

- Nếu muốn tỷ suất sinh lời kì vọng cao hơn thì chọn cổ phiếu nào?
- Nếu muốn rủi ro thấp hơn thì chọn cổ phiếu nào?

# Phương sai

- Phương sai (variance) của $X$ kí hiệu là $V(X)$

$$\boldsymbol{V(X) = E[X - E(X)]^2}$$

Tương đương: $\boldsymbol{V(X) = E(X^2) - [E(X)]^2}$

Trong đó: $E(X^2) = \sum x_i^2 p_i$

- Chú ý: $V(X) \geq 0$
- Tính chất

| |
|---|
| $V(C) = 0$ với $C$ là hằng số |
| $V(X + C) = V(X)$ |
| $V(C \cdot X) = C^2 \cdot V(X)$ |
| $V(X \pm Y) = V(X) + V(Y)$ với $X$ và $Y$ độc lập |

# Độ lệch chuẩn

- Độ lệch chuẩn của biến ngẫu nhiên $X$, kí hiệu $\sigma_X$, là căn bậc hai của phương sai:

$$\sigma_X = \sqrt{V(X)}$$

- **Ý nghĩa:** Phương sai và độ lệch chuẩn phản ánh độ phân tán, biến động, dao động, rủi ro, bất ổn / ổn định, đồng đều của biến ngẫu nhiên
- Đơn vị của phương sai là bình phương đơn vị của $X$
- Đơn vị của độ lệch chuẩn là đơn vị của $X$

- Tính Phương sai của các biến ngẫu nhiên trong ví dụ 3.5 và kết luận.

# Trung vị - Mốt

- **Trung vị** (***median***) của BNN rời rạc $X$ ký hiệu là $m_d$ là giá trị nằm ở chính giữa phân phối xác suất.

  Nếu $m_d = x_i$ thì $F(x_i) \leq 0{,}5 < F(x_{i+1})$

- **Mốt** (***mode***) của BNN rời rạc $X$, ký hiệu $m_0$ là giá trị của X có xác suất lớn nhất.

  BNN có thể không có mốt, có 1 mốt, hoặc nhiều mốt

# 3.5. PHÂN PHỐI KHÔNG – MỘT

- Biến ngẫu nhiên rời rạc $X$ chỉ nhận 2 giá trị 0 và 1 với xác suất
$$P(X=x)=p^x(1-p)^{1-x} \text{ với } x = 0;1$$
gọi là phân phối Không-Một, hay phân phối Bernoulii
- Bảng phân phối xác suất

| $X$ | $0$ | $1$ |
|---|---|---|
| $P$ | $1-p$ | $p$ |

- Kí hiệu: $X \sim A(p)$
- **Tham số đặc trưng**:
$$E(X) = p; \; V(X) = p(1-p)$$

# 3.6. PHÂN PHỐI NHỊ THỨC

- Biến ngẫu nhiên rời rạc $X$ nhận một trong các giá trị có thể có là $0, 1, 2, \ldots$ với các xác suất tương ứng là

$$P(X = x) = C_n^x p^x (1-p)^{n-x} \quad ; x = 0,1, \ldots, n$$

gọi là phân phối Nhị thức (*Binomial*) với tham số $n$ và $p$.

- Kí hiệu: $X \sim B(n; p)$

- Nếu bài toán thỏa mãn lược đồ Bernoulli (chương 2) và $X$ là số lần $A$ xảy ra trong $n$ phép thử thì $X \sim B(n; p)$.

# Phân phối Nhị thức

**Các tham số đặc trưng**:

- Kì vọng $E(X) = np;$
- Phương sai $V(X) = np(1-p).$
- Mốt $m_0$ (giá trị có khả năng xảy ra lớn nhất)

$$np + p - 1 \le m_0 \le np + p$$

# Phân phối Nhị thức

**Ví dụ 3.6.** Một người đi chào hàng với 10 khách hàng độc lập. Xác suất bán được hàng cho mỗi khách hàng đều bằng 0,2.

a) Tính xác suất bán được hàng cho đúng 3 khách.

b) Tính xác suất bán được hàng cho ít nhất 2 khách.

c) Tính kì vọng, phương sai, mốt của số khách mua hàng.

# 3.7. PHÂN PHỐI POISSON

- Biến ngẫu nhiên rời rạc $X$ nhận một trong các giá trị có thể có là $0, 1, 2, \ldots$ với các xác suất tương ứng là

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ với } x = 0,1,2,\ldots$$

gọi là phân phối Poisson với tham số $\lambda$.

- Kí hiệu: $X \sim P(\lambda)$
- Nếu $X \sim B(n; p)$ với $n$ khá lớn, $p$ khá nhỏ và $np \approx npq$ thì coi như $X \sim P(\lambda)$ với $\lambda = np$.
- **Các tham số đặc trưng**

$$E(X) = \lambda; \;\; V(X) = \lambda; \;\; \lambda - 1 \le m_0 \le \lambda$$

# Phân phối Poisson

**Ví dụ 3.7.** Vận chuyển $5000$ chai rượu. Dựa theo định nghĩa thống kê về xác suất, xác suất 1 chai bị vỡ trong quá trình vận chuyển là $\frac{1}{1000}$.

a) Tính xác suất trong quá trình vận chuyển có đúng 2 chai bị vỡ.

b) Giả sử bị vỡ 2 chai, vậy tổn thất của 2 chai rượu đó có khả năng cao nhất sẽ tính cho ai? (Chủ kinh doanh rượu, người vận chuyển hay người tiêu dùng cuối cùng). Nếu bạn là chủ kinh doanh, bạn sẽ quyết định như thế nào để loại bỏ tổn thất ngoài ý muốn trên.

# Sử dụng các hàm trên EXCEL

| Phân phối | Lệnh | Kết quả |
|---|---|---|
| Nhị thức | = Binom.dist($x, n, p, 1$) | $P(X \leq x)$ |
| | = Binom.dist($x, n, p, 0$) | $P(X = x)$ |
| Poisson | = Poisson.dist($x, \lambda, 1$) | $P(X \leq x)$ |
| | = Poisson.dist($x, \lambda, 0$) | $P(X = x)$ |

# BÀI 4 – BIẾN NGẪU NHIÊN LIÊN TỤC VÀ PHÂN PHỐI XÁC SUẤT

- **4.1. Biến ngẫu nhiên liên tục**
- **4.2. Hàm phân phối xác suất**
- **4.3. Hàm mật độ xác suất**
- **4.4. Các tham số đặc trưng**
- **4.5. Phân phối Đều**
- **4.6. Phân phối Chuẩn**
- **4.7. Phân phối khác**

[1] Chương 2, trang 79 – 146; Chương 3, tr 167 – 208.

[2] Mục 2, trang 5 - 9.

[3] Chapter 5, pp. 197 – 244.

# 4.1. BIẾN NGẪU NHIÊN LIÊN TỤC

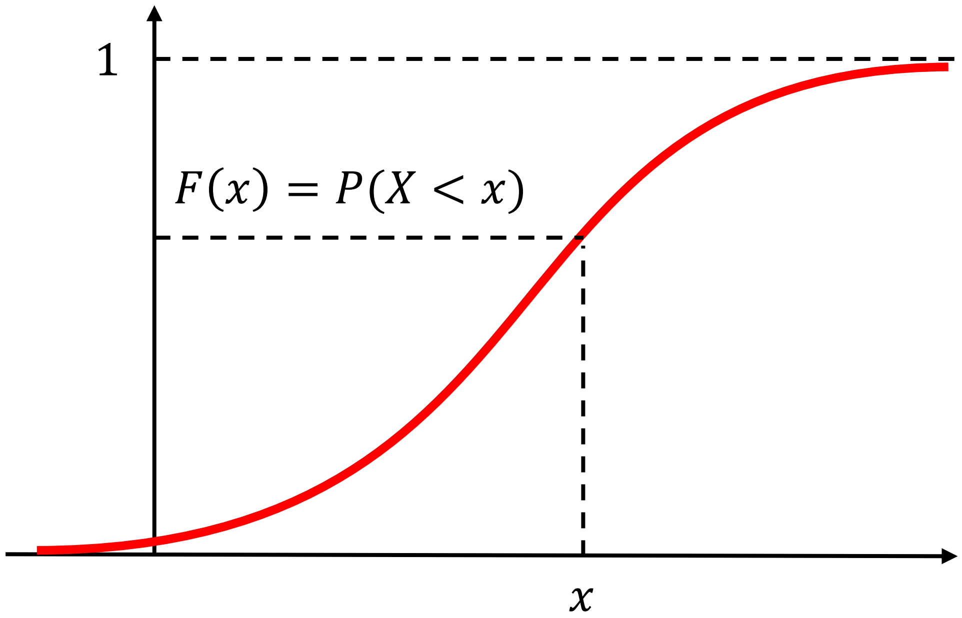
- Biến ngẫu nhiên liên tục (*Continuous Random Variable*) là biến ngẫu nhiên có thể nhận mọi giá trị trong một khoảng $(a; b), a, b \in \mathbb{R}$.
- **Ví dụ 4.1.**

- Thời gian đi từ nhà đến trường của sinh viên
- Lợi nhuận của nhà đầu tư cổ phiếu sau một năm
- Cân nặng của trẻ sơ sinh ở Việt Nam

# 4.2. HÀM PHÂN PHỐI XÁC SUẤT

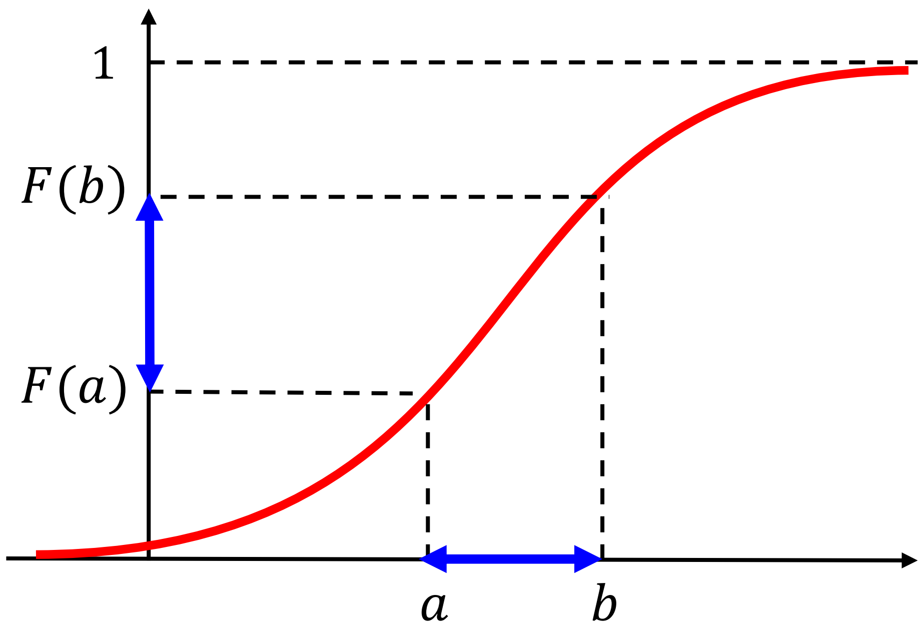
Hàm phân phối xác suất (hàm phân phối tích lũy - *Cumulative Distribution Function*) của BNN $X$ là:

$$F(x) = P(X < x), x \in \mathbb{R}$$



# Tính chất

- $P(a < X < b) = F(b) - F(a)$

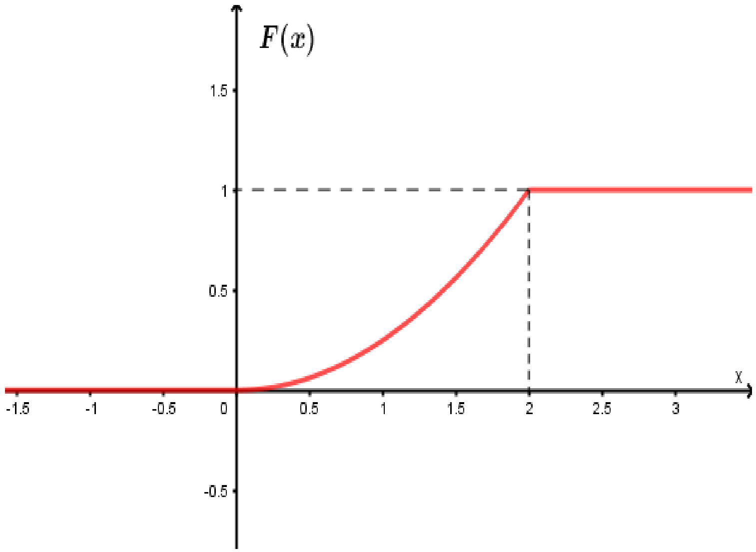



- Nếu $X$ là biến ngẫu nhiên liên tục thì: $P(X = x_0) = 0$
$\Rightarrow P(a < X < b) = P(a \le X \le b) = F(b) - F(a)$

# Ví dụ

**Ví dụ 4.2.** Tuổi thọ của một loại sản phẩm là biến ngẫu nhiên liên tục (đơn vị:năm) có hàm phân phối xác suất:

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ \frac{1}{4}x^2 & \text{nếu } 0 \le x \le 2 \\ 1 & \text{nếu } x > 2 \end{cases}$$




Lấy ngẫu nhiên một sản phẩm.

Tính xác suất để:

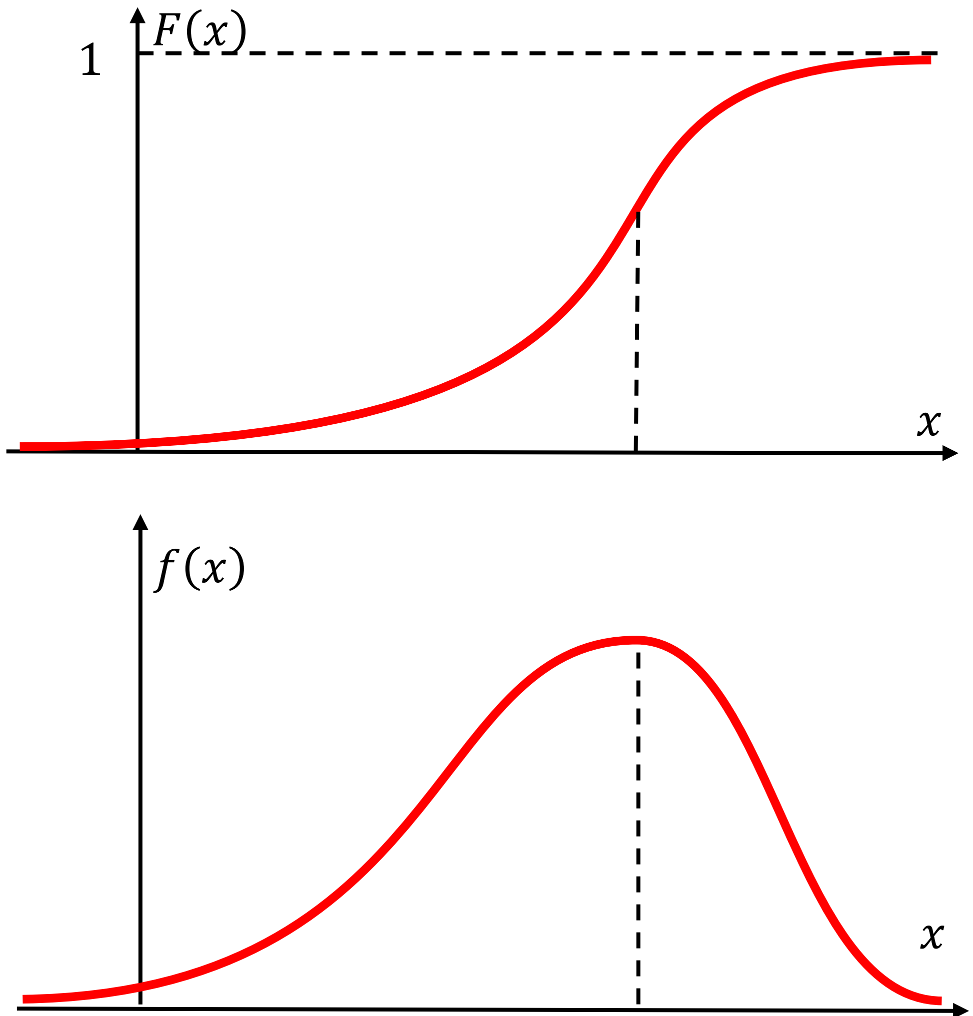
a) Sản phẩm có tuổi thọ nhỏ hơn 1 năm.

b) Sản phẩm có tuổi thọ từ 1 đến 2 năm.

# 4.3. HÀM MẬT ĐỘ XÁC SUẤT

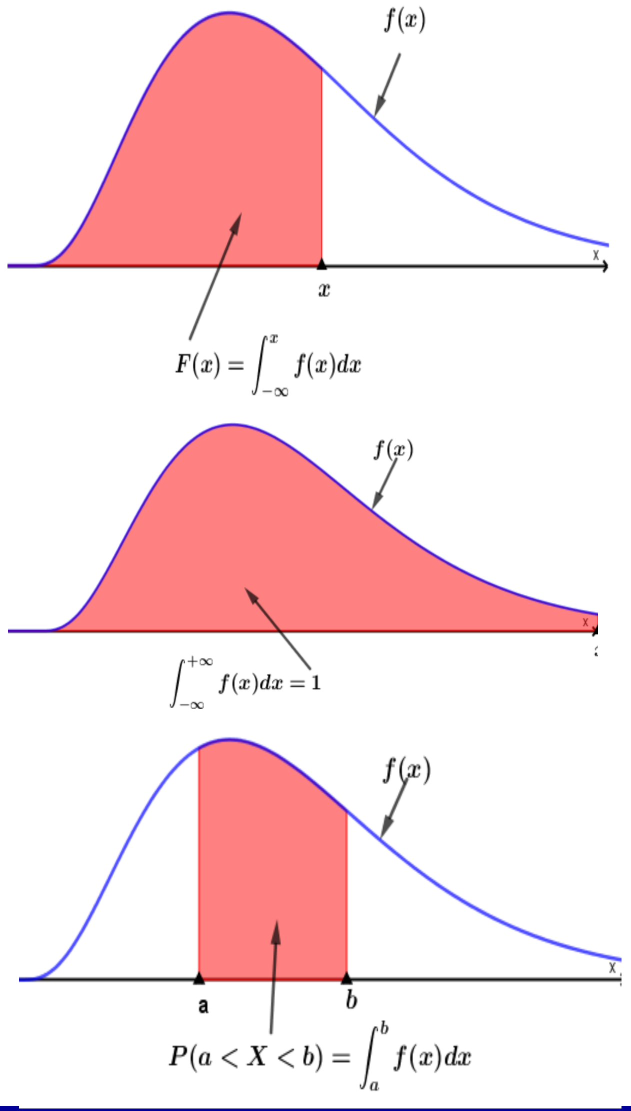
$X$ là biến ngẫu nhiên liên tục, hàm mật độ xác suất (*Probability Density Function*) của $X$, ký hiệu $f(x)$, là:

$$f(x) = F'(x), x \in \mathbb{R}$$



# Tính chất

- $f(x) \geq 0$ với $\forall x$
- $F(x) = \int_{-\infty}^{x} f(t)dt$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(a < X < b) = \int_{a}^{b} f(x)dx$



# Ví dụ

**Ví dụ 4.3.** Thời gian để công nhân hoàn thành một sản phẩm là biến ngẫu nhiên liên tục $X$ (đơn vị: phút) có hàm mật độ xác suất:

$$f(x) = \begin{cases} \dfrac{1}{16} x \text{ nếu } x \in [2;6] \\ 0 \quad \text{nếu } x \notin [2;6] \end{cases}$$

a) Tìm tỷ lệ sản phẩm có thời gian hoàn thành từ 3 đến 5 phút.

b) Tính xác suất để trong 5 sản phẩm bất kì có đúng 2 sản phẩm có thời gian hoàn thành nhiều hơn 4 phút.

# 4.4. CÁC THAM SỐ ĐẶC TRƯNG

- Kỳ vọng:

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

- Phương sai:

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2.f(x)dx$$

- Độ lệch chuẩn:

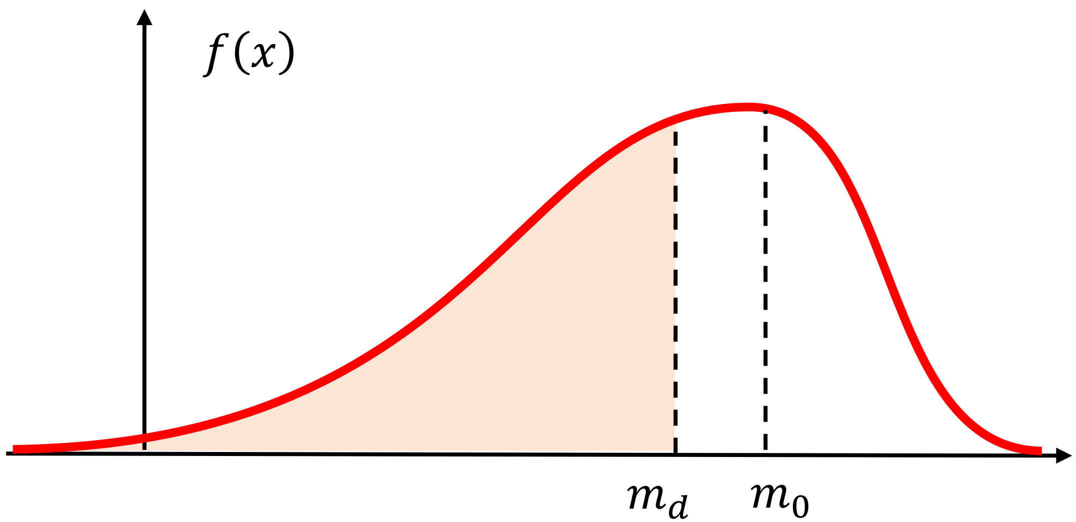
$$\sigma_X = \sqrt{V(X)}$$

# Trung vị và Mốt

- **Trung vị (median)** của BNN liên tục X, kí hiệu $m_d$, là giá trị chia phân phối xác suất của biến ngẫu nhiên thành hai phần bằng nhau.

$$\int_{-\infty}^{m_d} f(x)\,dx = 0{,}5$$

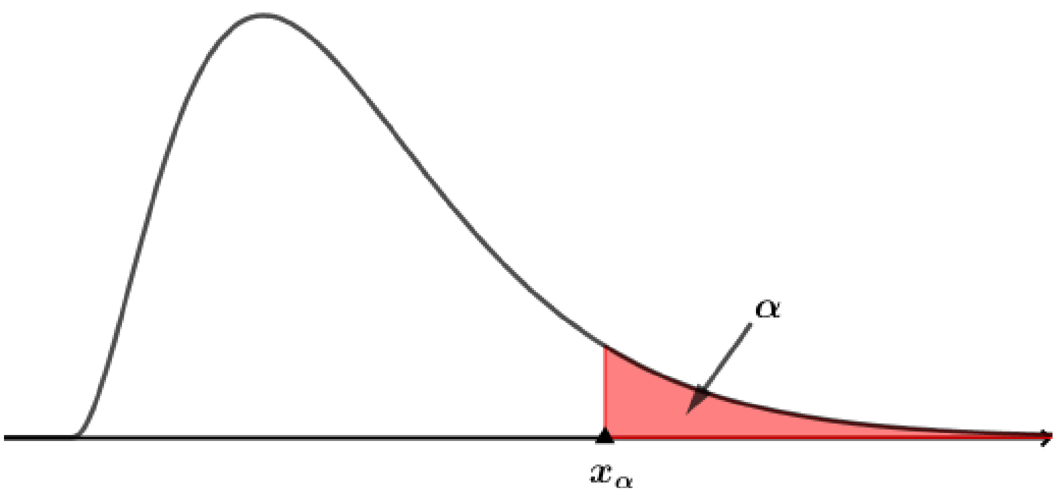
- **Mốt (mode)** của BNN liên tục X, kí hiệu $m_o$, là giá trị mà tại đó hàm mật độ xác suất $f(x)$ đạt giá trị cực đại.



# Giá trị tới hạn

Giá trị tới hạn mức $\alpha$ của biến ngẫu nhiên $X$, ký hiệu $x_\alpha$, là giá trị của $X$ thỏa mãn:

$$P(X > x_\alpha) = \alpha$$

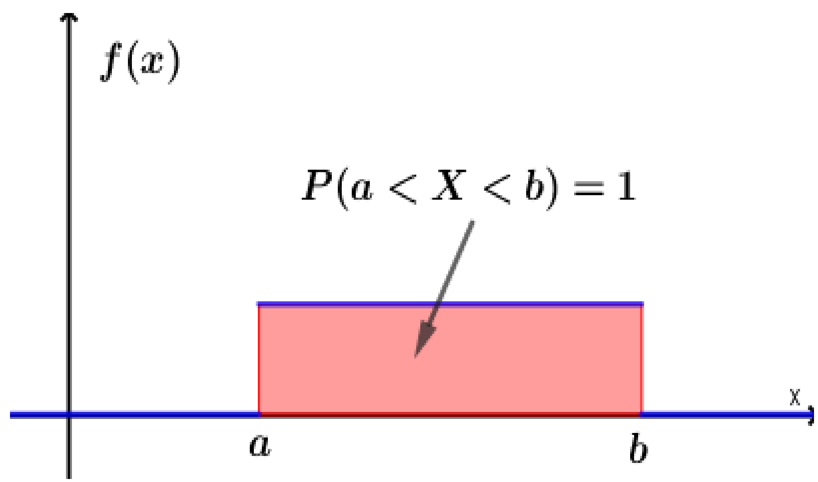


# 4.5. PHÂN PHỐI ĐỀU

- Biến ngẫu nhiên liên tục $X$ có phân phối Đều (*Uniform Distribution*) trên khoảng $(a; b)$ nếu hàm mật độ xác suất của $X$ có dạng:

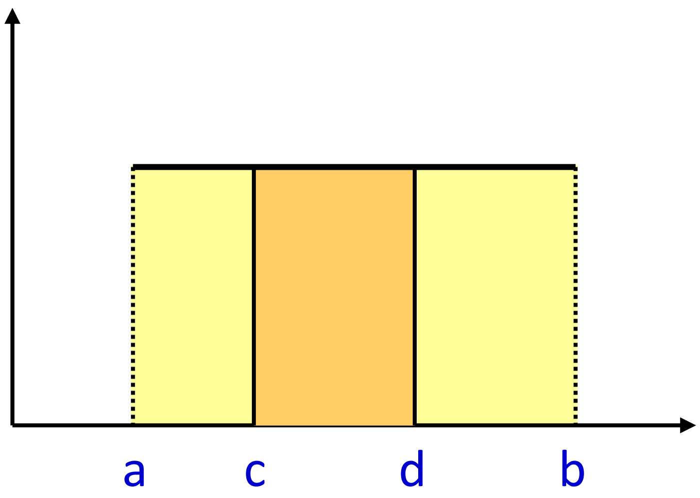
$$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in (a;b) \\ 0 & x \notin (a;b) \end{cases}$$

- Ký hiệu: $X \sim U(a; b)$
- Đồ thị hàm $f(x)$



# Phân phối đều

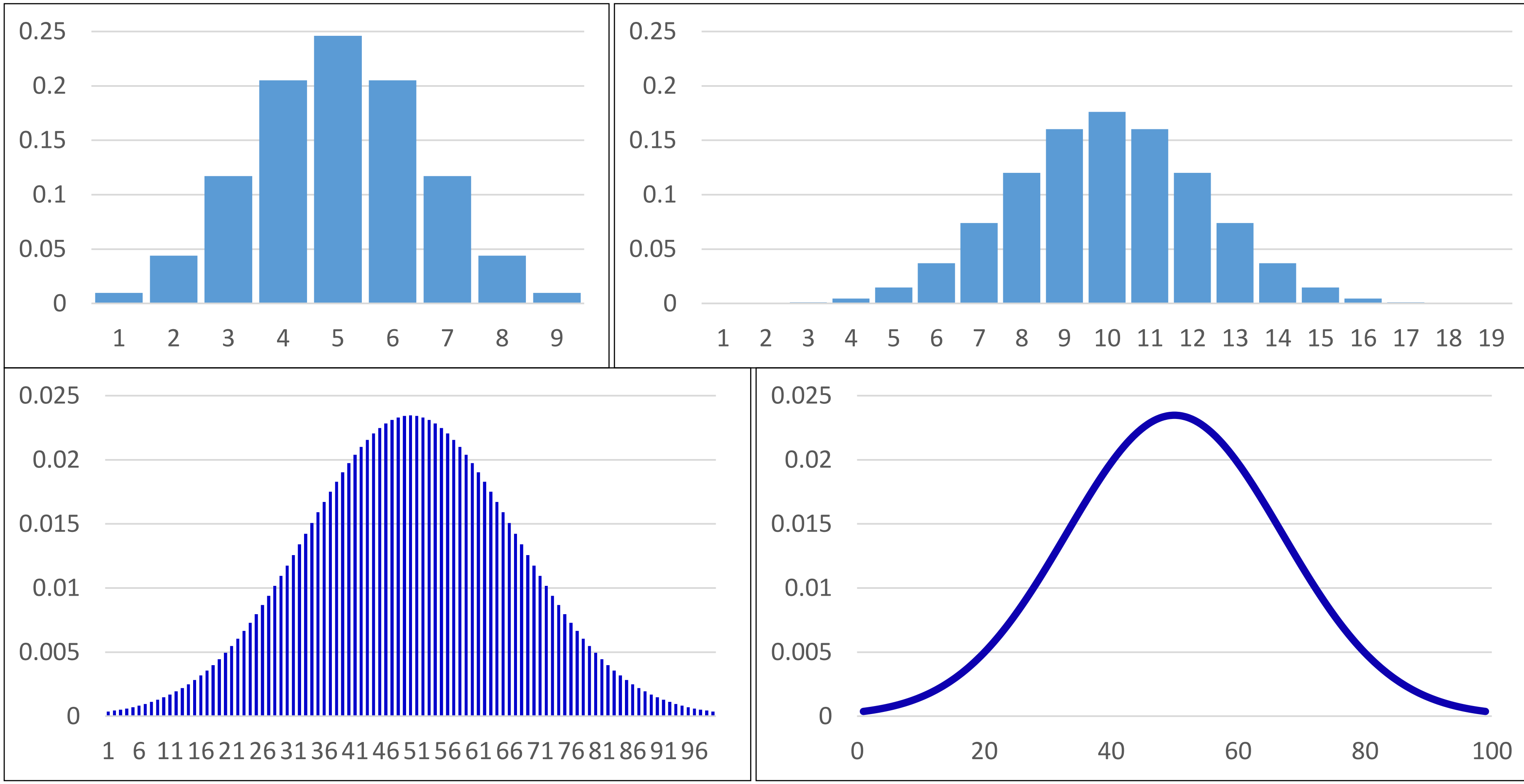
- $E(X) = \frac{a+b}{2}$
- $V(X) = \frac{(b-a)^2}{12}$
- $P(c < X < d) = \frac{d-c}{b-a}$




- **Ví dụ 4.4.** Thống kê cho thấy, thời gian một chuyến xe buýt A đi hết một hành trình ít nhất là 40 phút và nhiều nhất là 60 phút. Tìm hàm mật độ xác suất của thời gian đi hết hành trình của xe buýt A.

# 4.6. PHÂN PHỐI CHUẨN

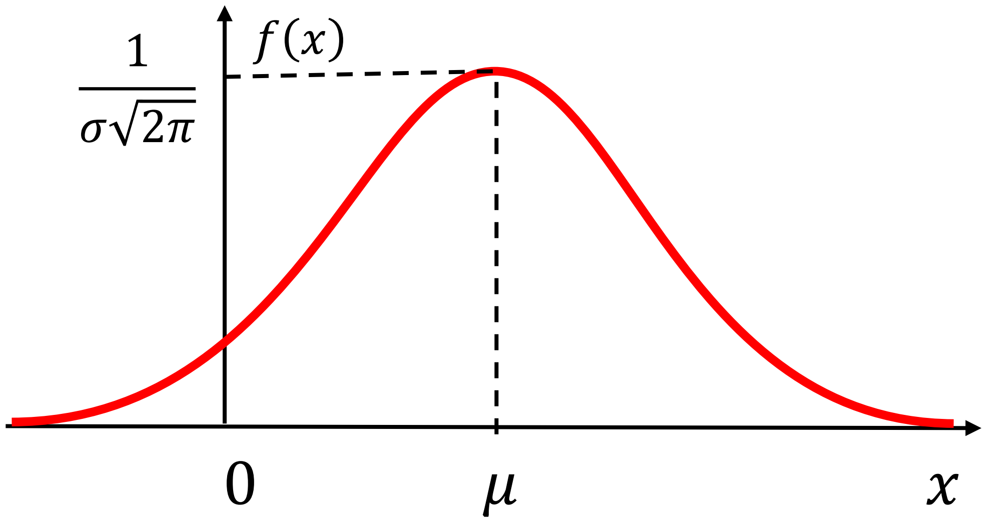
- $B\ (n;\ p\ =\ 0,5)$ với $n = 10;\ 20;\ 100$



# Phân phối Chuẩn

- Biến ngẫu nhiên liên tục $X$ có phân phối Chuẩn (*Normal Distribution*) nếu hàm mật độ có dạng:

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

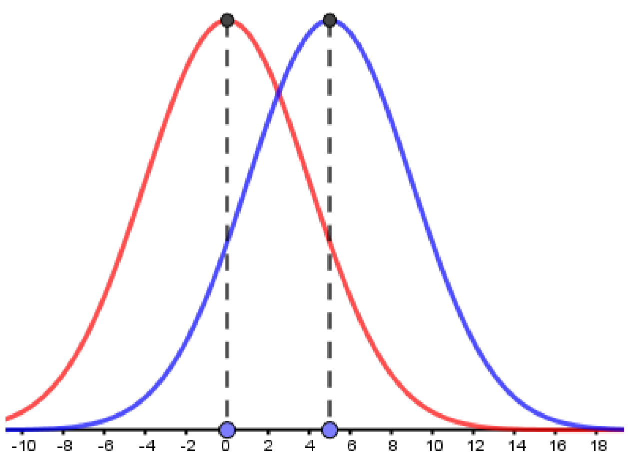



- Đồ thị $f(x)$ dạng quả chuông, đối xứng qua đường $x = \mu$
- Ký hiệu: $X \sim N(\mu, \sigma^2)$

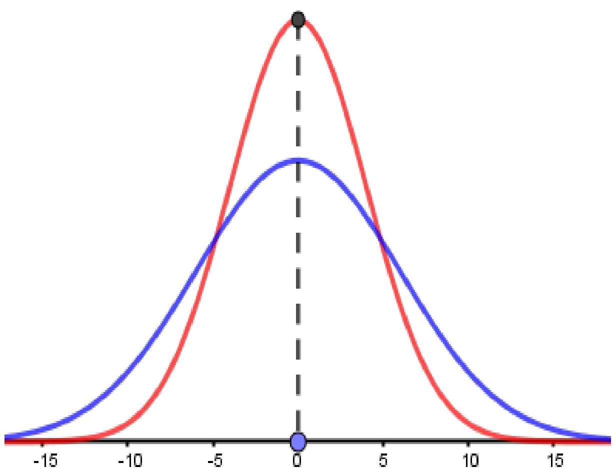
# Các tham số

$X \sim N(\mu, \sigma^2)$

- $E(X) = \mu$
- $V(X) = \sigma^2$




Khi $\mu$ tăng thì đồ thị của $f(x)$ dịch sang phải



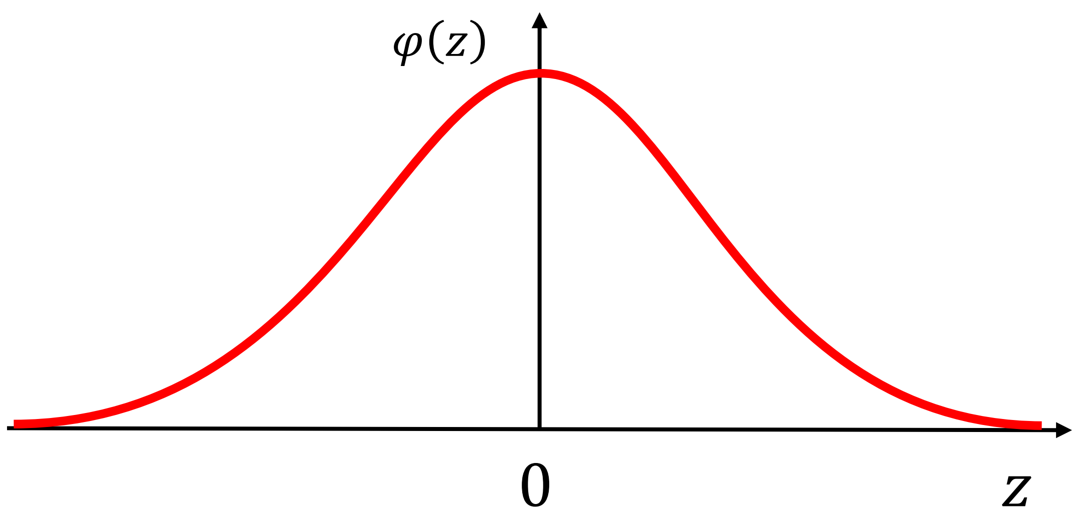

Khi $\sigma$ tăng thì đồ thị của $f(x)$ thấp xuống và rộng ra

# Phân phối Chuẩn hóa

- Biến $Z$ phân phối Chuẩn hóa (*Standardized Normal*), nếu $Z$ phân phối Chuẩn với $\mu = 0$ và $\sigma^2 = 1$.

$$Z \sim N(0;1)$$

- Hàm mật độ xác suất: $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- Đồ thị $\phi(z)$ đối xứng qua trục tung.



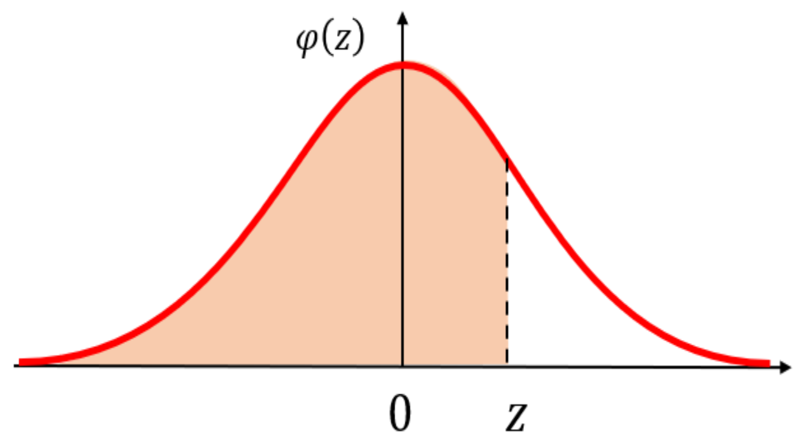
# Phân phối Chuẩn hóa

- Hàm phân phối xác suất: xác suất về bên trái một điểm

$$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \varphi(t)dt$$

Tính chất: $\Phi(-z) + \Phi(z) = 1$

Bảng giá trị $\Phi(z)$: bảng số

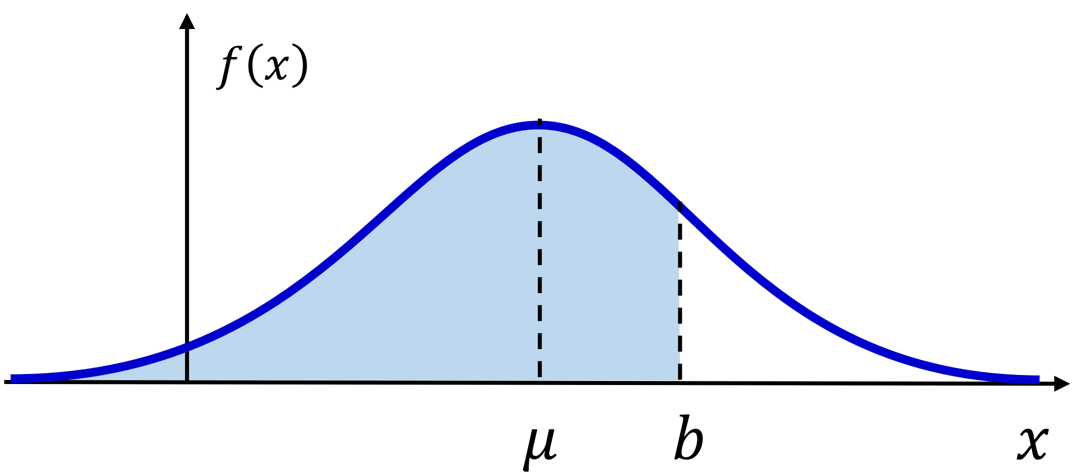


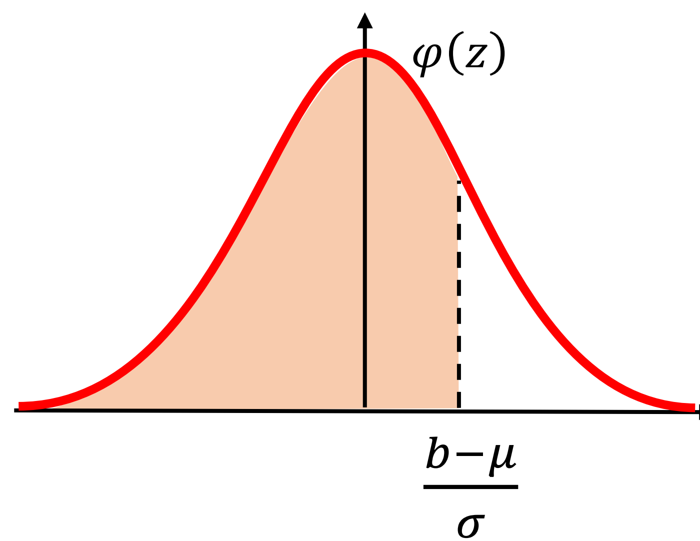

- Nếu $X \sim N(\mu; \sigma^2)$ thì $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

# Công thức tính xác suất

$X \sim N(\mu;\sigma^2)$

- $P(X<b)=P\left(\frac{X-\mu}{\sigma}<\frac{b-\mu}{\sigma}\right)=P\left(Z<\frac{b-\mu}{\sigma}\right)=\Phi\left(\frac{b-\mu}{\sigma}\right)$
- $P(a<X)=1-\Phi\left(\frac{a-\mu}{\sigma}\right)$
- $P(a<X<b)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$



# Công thức tính xác suất

**Ví dụ 4.5.** Lợi nhuận (đv: triệu) của một dự án là biến ngẫu nhiên phân phối Chuẩn, với trung bình bằng 500, phương sai bằng 400. Tính xác suất để:

a) Lợi nhuận cao hơn 540.

b) Lợi nhuận thấp hơn 570.

c) Lợi nhuận từ 480 đến 550.

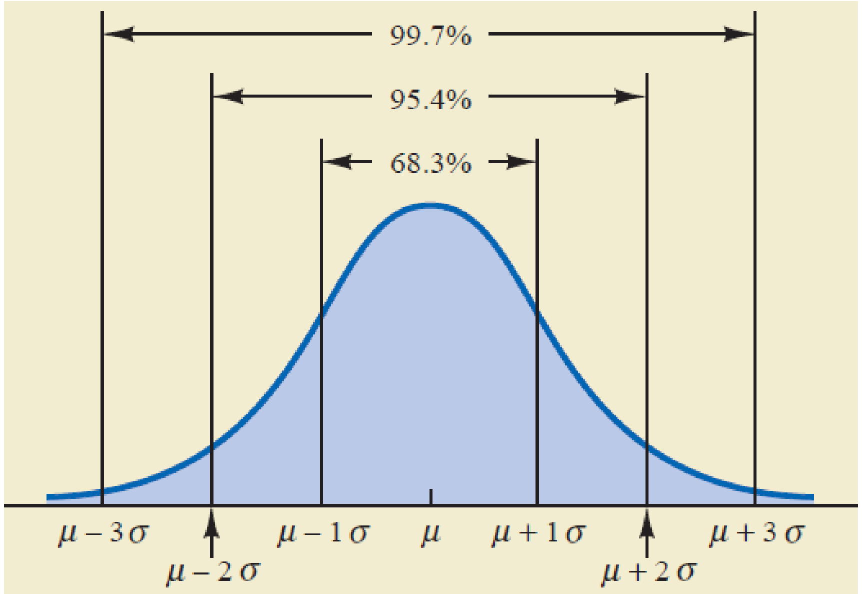
# Xác suất sai lệch so với kì vọng

$X \sim N(\mu; \sigma^2)$

$$P(|X-\mu| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right) - 1$$

Ba trường hợp:

- Quy tắc 1 - sigma:
  $P(|X-\mu| < \sigma) = 0{,}6826$
- Quy tắc 2 - sigma:
  $P(|X-\mu| < 2\sigma) = 0{,}9544$
- Quy tắc 3 - sigma:
  $P(|X-\mu| < 3\sigma) = 0{,}9974$



# Tổ hợp của các BNN phân phối Chuẩn

- $X_1, X_2, \dots, X_n$ độc lập
- $X_i \sim N\left(\mu_i, \sigma_i^2\right), i = 1, 2, \dots, n$ thì:
- Thì: $Y = \sum_{i=1}^{n} a_i X_i \sim N(\mu_Y, \sigma_Y^2)$

Trong đó:

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \quad ; \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

# Hội tụ của phân phối B(n;p) về Chuẩn

- $X \sim B(n;p)$, khi $n \geq 100$ có thể xấp xỉ $X \sim N(\mu, \sigma^2)$
- Với $\mu = np$ và $\sigma^2 = np(1 - p)$

**Ví dụ 4.6.**

Xác suất để một khách hàng vào siêu thị mua sản phẩm của hãng A là 0,3. Tính xác suất để trong 200 khách vào siêu thị có:

a) Nhiều hơn 70 khách mua hàng của hãng A.

b) Từ 40 đến 50 khách mua hàng của hãng A.

# Giá trị tới hạn Chuẩn

- Giá trị tới hạn Chuẩn mức $\alpha$, ký kiệu là $z_\alpha$, được xác định bởi:

$$P(Z > z_\alpha) = \alpha$$

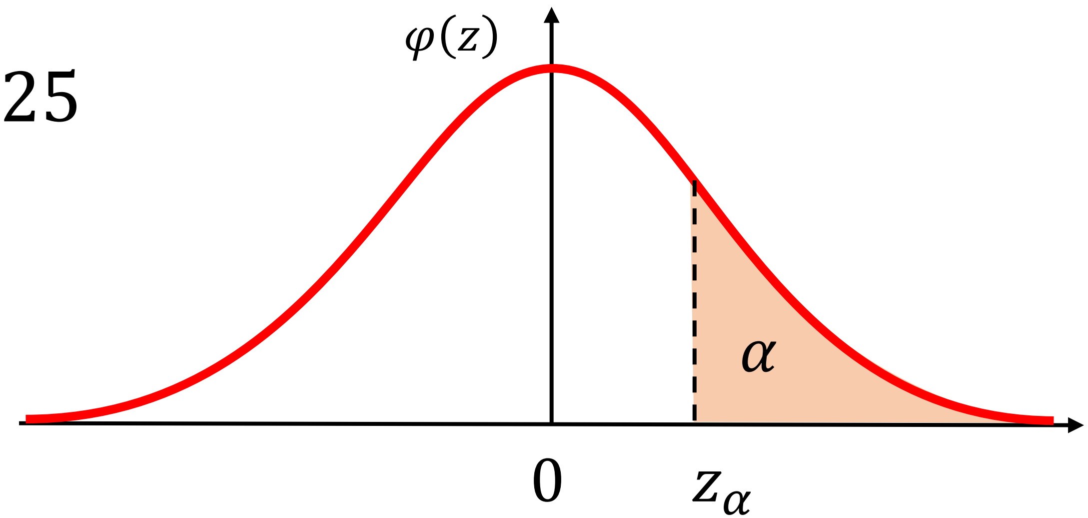
- **Tra bảng**

$$P(Z > 1{,}96) = 0{,}025$$

$$\Rightarrow z_{0{,}025} = 1{,}96$$

- **Tính chất:**

$$z_{1-\alpha} = -z_\alpha$$



# 4.7. PHÂN PHỐI KHÁC

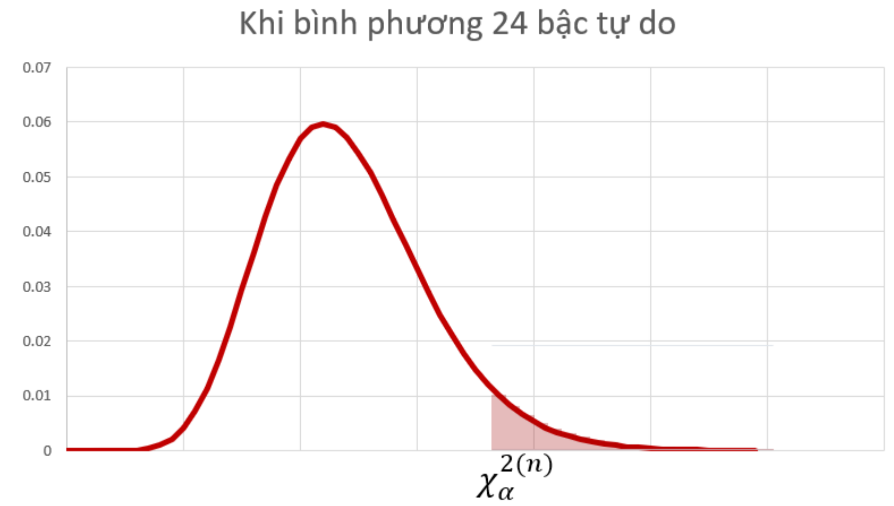
Phân phối khác dùng cho thống kê, cần nắm cách tìm giá trị tới hạn để làm bài tập

- Phân phối Khi bình Phương
- Phân phối Student
- Phân phối Fisher – Snedecor

# Phân phối Khi bình phương

- BNN liên tục $\chi^2$ phân phối Khi bình phương với $n$ bậc tự do (*degree of freedom: df*), kí hiệu: $\chi^2 \sim \chi^2(n)$
- Tham số: $E(\chi^2) = n; V(\chi^2) = 2n$
- Giá trị tới hạn mức $\alpha$ là $\chi_\alpha^{2(n)}$

$$P\left(\chi^2(n) > \chi_\alpha^{2(n)}\right) = \alpha$$



# Phân phối Khi bình phương

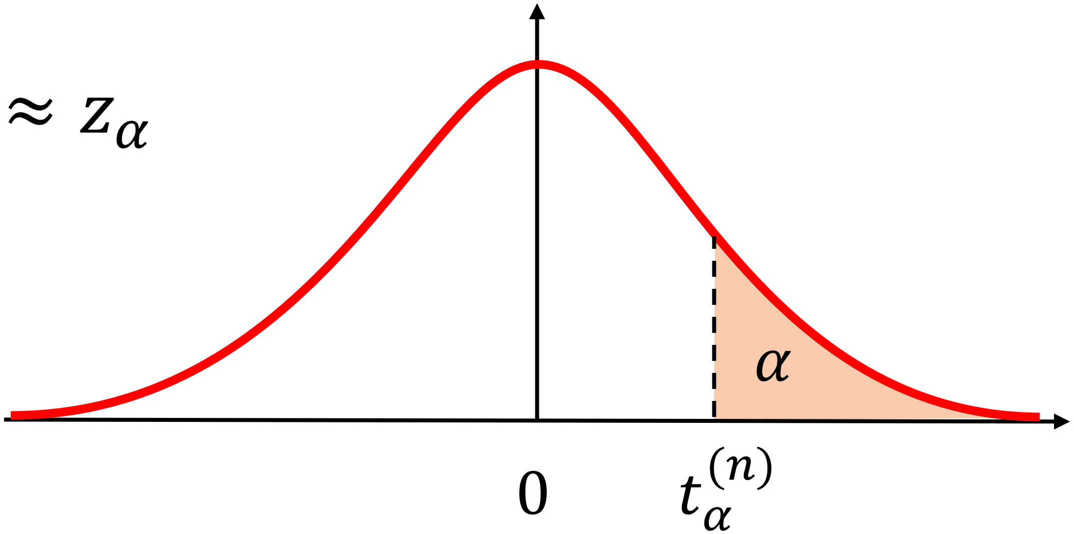
- Bảng đầy đủ: phụ lục 7 giáo trình
- Bảng giản lược $\chi^{2(10)}_{0,05}$ và $\chi^{2(50)}_{0,95}$

| n \ α | 0.975 | 0.95 | 0.05 | 0.025 |
|---|---|---|---|---|
| **1** | 0.001 | 0.004 | 3.841 | 5.024 |
| **2** | 0.051 | 0.103 | 5.991 | 7.378 |
| **3** | 0.216 | 0.352 | 7.815 | 9.348 |
| **4** | 0.484 | 0.711 | 9.488 | 11.14 |
| **5** | 0.831 | 1.145 | 11.07 | 12.83 |
| **10** | 3.247 | 3.940 | $\boxed{18.31}$ | 20.48 |
| **15** | 6.262 | 7.261 | 25.00 | 27.49 |

| n \ α | 0.975 | 0.95 | 0.05 | 0.025 |
|---|---|---|---|---|
| **20** | 9.591 | 10.85 | 31.41 | 34.17 |
| **24** | 12.40 | 13.85 | 36.42 | 39.36 |
| **30** | 16.79 | 18.49 | 43.77 | 46.98 |
| **39** | 23.65 | 25.70 | 54.57 | 58.12 |
| **50** | 32.36 | $\boxed{34.76}$ | 67.50 | 71.42 |
| **99** | 73.36 | 77.05 | 123.2 | 128.4 |
| **120** | 91.57 | 95.70 | 146.6 | 152.2 |

# Phân phối Student

- BNN liên tục T phân phối Student với $n$ bậc tự do, kí hiệu: $T \sim T(n)$
- Tham số: $E(T) = 0; V(T) = \frac{n}{n-2}$
- Giá trị tới hạn mức $\alpha$ là $t_{\alpha}^{(n)}: P\left(T(n) > t_{\alpha}^{(n)}\right) = \alpha$
  - $t_{1-\alpha}^{(n)} = -t_{\alpha}^{(n)}$
  - $n > 30$ thì $t_{\alpha}^{(n)} \approx z_{\alpha}$



# Phân phối Student

- Bảng đầy đủ: phụ lục 8 giáo trình
- Bảng giản lược $t_{0,05}^{(15)}$ và $t_{0,025}^{(32)} \approx z_{0,025}$

| n \ α | 0.1 | 0.05 | 0.025 |
|---|---|---|---|
| **10** | 1.372 | 1.812 | 2.228 |
| **11** | 1.363 | 1.796 | 2.201 |
| **12** | 1.356 | 1.782 | 2.179 |
| **13** | 1.350 | 1.771 | 2.160 |
| **14** | 1.345 | 1.761 | 2.145 |
| **15** | 1.341 | 1.753 | 2.131 |
| **16** | 1.337 | 1.746 | 2.120 |

| n \ α | 0.1 | 0.05 | 0.025 |
|---|---|---|---|
| **17** | 1.333 | 1.740 | 2.110 |
| **18** | 1.330 | 1.734 | 2.101 |
| **19** | 1.328 | 1.729 | 2.093 |
| **20** | 1.325 | 1.725 | 2.086 |
| **24** | 1.318 | 1.711 | 2.064 |
| **30** | 1.310 | 1.697 | 2.042 |
| **∞** | 1.282 | 1.645 | 1.960 |

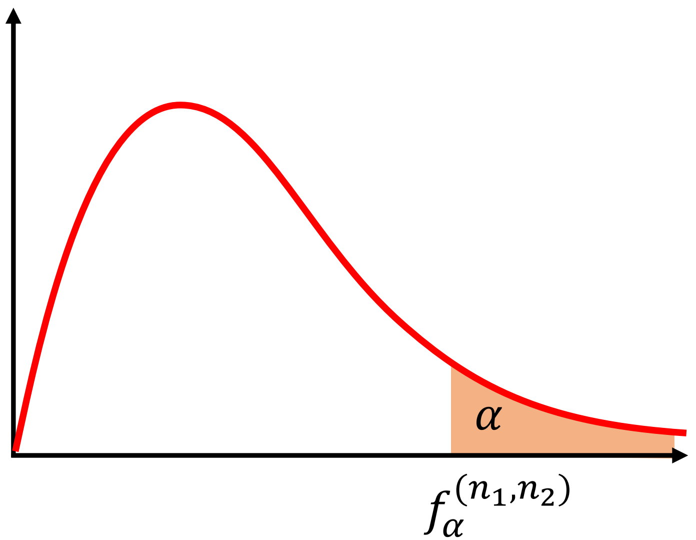
# Phân phối Fisher

- BNN liên tục $F$ tuân theo quy luật Fisher-Snedecor (gọi tắt là Fisher) với hai bậc tự do $n_1$ và $n_2$
- Ký hiệu: $F \sim F(n_1, n_2)$
- Giá trị tới hạn mức $\alpha$ là $f_\alpha^{(n_1,n_2)}$

  $$P\left(F(n_1, n_2) > f_\alpha^{(n_1,n_2)}\right) = \alpha$$

- Tính chất:

  $$f_{1-\alpha}^{(n_1,n_2)} = \frac{1}{f_\alpha^{(n_2,n_1)}}$$



# Phân phối Fisher

- Bảng đầy đủ: phụ lục 9 giáo trình; Bảng giản lược: $f_{0,05}^{(39,59)}$

| $n_2$ | $\alpha$ \ $n_1$ | 24 | 39 | 59 | 99 | 120 |
|---|---|---|---|---|---|---|
| 24 | ***0.025*** | 2.27 | 2.15 | 2.08 | 2.03 | 2.01 |
| | ***0.05*** | 1.98 | 1.90 | 1.84 | 1.80 | 1.79 |
| 39 | ***0.025*** | 2.02 | 1.89 | 1.82 | 1.75 | 1.74 |
| | ***0.05*** | 1.80 | 1.70 | 1.65 | 1.60 | 1.58 |
| 49 | ***0.025*** | 1.94 | 1.81 | 1.73 | 1.66 | 1.65 |
| | ***0.05*** | 1.74 | 1.64 | 1.58 | 1.53 | 1.52 |
| 59 | ***0.025*** | 1.89 | 1.75 | 1.67 | 1.60 | 1.59 |
| | ***0.05*** | 1.70 | $\boxed{1.60}$ | 1.54 | 1.49 | 1.47 |
| 99 | ***0.025*** | 1.79 | 1.65 | 1.56 | 1.49 | 1.46 |
| | ***0.05*** | 1.63 | 1.52 | 1.45 | 1.39 | 1.38 |

# BÀI 5 – BIẾN NGẪU NHIÊN HAI CHIỀU

- **5.1. Khái niệm biến ngẫu nhiên nhiều chiều**
- **5.2. Bảng phân phối xác suất hai chiều**
- **5.3. Tham số đặc trưng**

[1] Chương 4, trang 209 – 270

[3] Chapter 4, pp 176 - 190

# 5.1. BIẾN NGẪU NHIÊN NHIỀU CHIỀU

- Chiều cao X (m) của sinh viên VN là biến ngẫu nhiên 1 chiều.
- Điểm 10%, 20%, 20%, 50% của môn LT Xác suất&TKT: $(X, Y, Z, T)$ là biến ngẫu nhiên 4 chiều.
- Hệ $n$ biến ngẫu nhiên 1 chiều được xét một cách đồng thời tạo nên biến ngẫu nhiên $n$ chiều.
- Kí hiệu: $(X_1, X_2, \ldots . X_n)$

# Biến ngẫu nhiên hai chiều

- Hệ hai biến ngẫu nhiên 1 chiều được xét một cách đồng thời tạo nên biến ngẫu nhiên 2 chiều.
- Kí hiệu: $(X, Y)$
- **Ví dụ 5.1.** Thu nhập và tiêu dùng của hộ gia đình; chiều dài và chiều rộng của 1 sản phẩm, …
- **Phân loại**
  - BNN 2 chiều rời rạc: nếu $X, Y$ đều rời rạc
  - BNN 2 chiều liên tục: nếu $X, Y$ đều liên tục

# Ví dụ

- **Ví dụ 5.2.** Cho các biến ngẫu nhiên:
  $X$: Số lao động (người), $Y$: Số người ăn theo (người)

| $X$ \ $Y$ | 0 | 1 | 2 |
|---|---|---|---|
| 1 | 0,1 | 0,02 | 0,01 |
| 2 | 0,16 | 0,12 | 0,08 |
| 3 | 0,2 | 0,17 | ? |

- a) $P(X = 2, Y = 2) =$
- b) $P(Y = 2 \mid X = 1) =$

# 5.2. BẢNG PHÂN PHỐI XÁC SUẤT

Bảng phân phối xác suất đồng thời với các xác suất biên của các thành phần

| **X** \ **Y** | $y_1$ | $y_2$ | $\cdots$ | $y_m$ | **P(X)** |
|---|---|---|---|---|---|
| $x_1$ | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1m}$ | $P(x_1)$ |
| $x_2$ | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2m}$ | $P(x_2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_n$ | $p_{n1}$ | $p_{n2}$ | $\cdots$ | $p_{nm}$ | $P(x_n)$ |
| **P(Y)** | $P(y_1)$ | $P(y_2)$ | $\cdots$ | $P(y_m)$ | **1** |

# Bảng phân phối xác suất biên

- Bảng phân phối xác suất biên của $X$

| $X$ | $x_1$ | ... | $x_i$ | ... | $x_n$ |
|---|---|---|---|---|---|
| $P$ | $P(x_1)$ | ... | $P(x_i)$ | ... | $P(x_n)$ |

- Bảng phân phối xác suất biên của $Y$

| $Y$ | $y_1$ | ... | $y_j$ | ... | $y_m$ |
|---|---|---|---|---|---|
| $P$ | $P(y_1)$ | ... | $P(y_j)$ | ... | $P(y_m)$ |

# Bảng phân phối xác suất có điều kiện

- Bảng phân phối của $(X \mid Y = y_j)$:

$$P(x_i|y_j) = \frac{P(x_i,y_j)}{P(y_j)}$$

| $(X \mid Y = y_j)$ | $x_1$ | $x_2$ | … | $x_n$ |
|---|---|---|---|---|
| $P$ | $P(x_1\|y_j)$ | $P(x_2\| y_j)$ | … | $P(x_n \| y_j)$ |

- Bảng phân phối của $(Y \mid X = x_i)$

| $(Y \mid X = xi)$ | $y_1$ | $y_2$ | … | $y_m$ |
|---|---|---|---|---|
| $P$ | $P(y_1 \|x_i)$ | $P(y_2\|x_i)$ | … | $P(y_m \|x_i)$ |

# 5.3. THAM SỐ ĐẶC TRƯNG

- **Kì vọng, phương sai: $\boldsymbol{E(X), V(X), E(Y), V(Y)}$**
- Kì vọng có điều kiện
  - $E(Y|x_i) = \sum_{j=1}^{m} y_j\, P(y_j|x_i)$
  - $E(X|y_j) = \sum_{i=1}^{n} x_i\, P(x_i|y_j)$
- Tính độc lập: $X, Y$ độc lập nếu

$$P(x_i, y_j) = P(x_i) \cdot P(y_j) \quad \text{với } \forall\, i, j$$

- Nếu tồn tại $i, j$ để $P(x_i, y_j) \neq P(x_i) \cdot P(y_j)$ thì $X, Y$ không độc lập

# Hiệp phương sai và hệ số tương quan

- **Hiệp phương sai** (*covariance*):

$$Cov(X,Y) = E([X - E(X)][Y - E(Y)])$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{m} x_i\, y_j P(x_i, y_j) - E(X)E(Y)$$

- **Hệ số tương quan** (*correlation*) $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sigma_X.\sigma_Y}$

| | | |
|---|---|---|
| $Cov > 0$ | $\rho = +1$ | $X, Y$ tương quan tuyến tính dương |
| | $0 < \rho < 1$ | $X, Y$ tương quan dương |
| $Cov = 0$ | $\rho = 0$ | $X, Y$ không tương quan |
| $Cov < 0$ | $-1 < \rho < 0$ | $X, Y$ tương quan âm |
| | $\rho = -1$ | $X, Y$ tương quan tuyến tính âm |

# Hiệp phương sai và hệ số tương quan

- Tính chất
  - $Cov(X,Y) = Cov(Y,X)$
  - $\rho_{X,Y} = \rho_{Y,X}$
  - $X,Y$ độc lập $\Rightarrow$ $Cov(X,Y) = 0$ và $\rho_{X,Y} = 0$
- Phương sai tổng hiệu tổng quát

$$V(aX \pm bY) = a^2 V(X) + b^2 V(Y) \pm 2abCov(X,Y)$$

# BÀI 6. LUẬT SỐ LỚN

- **Định lý giới hạn trung tâm**
- $X_1, X_2, \dots, X_n$ là các BNN độc lập có cùng phân phối xác suất, kỳ vọng và phương sai hữu hạn
- Đặt $Y = \sum_{i=1}^{n} X_i$ và $Z = \frac{Y - E(Y)}{\sqrt{V(Y)}}$
- Khi $n \to \infty$ thì $Z$ sẽ hội tụ về quy luật $N(0, 1)$
- Trong ứng dụng, $n \geq 30$ được coi là đủ lớn để áp dụng quy luật Chuẩn (dù biến ngẫu nhiên gốc không phân phối Chuẩn)

# BÀI 7 – MẪU NGẪU NHIÊN

- **7.1. Các khái niệm**
- **7.2. Trung bình mẫu**
- **7.3. Phương sai mẫu**
- **7.4. Tỉ lệ mẫu**

[1] Chương 6, trang 295 – 347, 361 – 363, 367 – 369

[3] Chapter 6, pp.244 - 283

# 7.1. CÁC KHÁI NIỆM

- Tổng thể
- Tham số đặc trưng của tổng thể
- Mẫu ngẫu nhiên
- Mẫu cụ thể
- Thống kê (tham số đặc trưng mẫu)

# Tổng thể

- Tập hợp toàn bộ các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hay định lượng nào đó được gọi là **tổng thể** (*population*)
- Kích thước tổng thể (*population size*): là số phần tử $\boldsymbol{N}$
- Dấu hiệu lượng hóa được: $\boldsymbol{X}$- Biến ngẫu nhiên gốc
- $X = \{x_1, x_2, \dots, x_N\}$
- Các tham số đặc trưng của $X$ là **tham số đặc trưng của tổng thể**

# Mô tả tổng thể

- Nếu $X$ chỉ gồm $k$ giá trị khác nhau: $x_1, x_2, \dots, x_k$
- Số lượng tương ứng là $N_1, N_2, \dots, N_k$
- gọi là tần số tổng thể của $x_i$
- Đặt $p_i = \frac{N_i}{N}$, gọi là tỉ lệ tổng thể

| Giá trị | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| Tần số | $N_1$ | $N_2$ | … | $N_k$ |
| Tỉ lệ | $p_1$ | $p_2$ | … | $p_k$ |

$$\begin{cases} 0 \le N_i \le N \\ \sum_{i=1}^{k} N_i = N \end{cases}$$

$$\begin{cases} 0 \le p_i \le 1 \\ \sum_{i=1}^{k} p_i = 1 \end{cases}$$

# Tham số đặc trưng của tổng thể

- Trung bình tổng thể (*population mean*): $\boldsymbol{\mu}$ (*hoặc* $m$)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = E(X)$$

- Phương sai tổng thể (*population variance*): $\boldsymbol{\sigma^2}$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2 = V(X)$$

- Độ lệch chuẩn tổng thể: $\sigma = \sqrt{\sigma^2}$
- Tỉ lệ tổng thể (*population proportion*): $\boldsymbol{p}$

  Tần số dấu hiệu (hay biến cố) $A$ là $M_A$

$$p = \frac{M_A}{N} = P(A)$$

# Nhận xét

- Nghiên cứu **Tổng thể:** qua tham số đặc trưng tổng thể, nghiên cứu toàn bộ các phần tử:
  - Chi phí lớn, có thể không khả thi
  - Sai sót khi thu thập, có thể phá hủy tập hợp
- Nghiên cứu một số phần tử đại diện: **Mẫu**
  - Từ tổng thể rút $n$ phần tử (mẫu kích thước $n$)
  - Xác định tham số đặc trưng mẫu (thống kê)
  - Rút ra kết luận liên quan đến tổng thể

# Mẫu ngẫu nhiên

- **Mẫu ngẫu nhiên** kích thước $n$ là tập hợp của $n$ biến ngẫu nhiên độc lập $X_1, X_2, \dots, X_n$ được thành lập từ biến ngẫu nhiên gốc $X$ và có cùng quy luật phân phối xác suất với $X$.
- Ký hiệu: $W = (X_1, X_2, \dots, X_n)$
- $E(X_i) = E(X) = m$
- $V(X_i) = V(X) = \sigma^2, i = \overline{1, n}$

# Mẫu cụ thể

- Gồm $n$ quan sát ($n$ con số): $w = (x_1, x_2, \dots, x_n)$
- Nếu chỉ gồm $k$ giá trị khác nhau: $x_1, x_2, \dots, x_k$ với số lần xuất hiện tương ứng : $n_1, n_2, \dots, n_k$
- $n_i$ là tần số mẫu của $x_i$ (*frequency*)
- Đặt $\hat{p}_i = \frac{n_i}{n}$ : tỉ lệ mẫu (*sample proportion*)

| Giá trị | $x_1$ | $x_2$ | ... | $x_k$ |
|---|---|---|---|---|
| Tần số | $n_1$ | $n_2$ | ... | $n_k$ |
| Tỉ lệ | $\hat{p}_1$ | $\hat{p}_2$ | ... | $\hat{p}_k$ |

$\sum_{i=1}^{k} n_i = n$

$\sum_{i=1}^{k} \hat{p}_i = 1$

# Ví dụ

- **Ví dụ 7.1.**
- Nghiên cứu về khối lượng sản phẩm $(X)$
- Mẫu ngẫu nhiên kích thước $n = 10$,
- $W = (X_1, X_2, \dots, X_{10})$
  - $E(X_1) = E(X_2) = \cdots = E(X_{10}) = E(X)$
  - $V(X_1) = V(X_2) = \cdots = V(X_{10}) = V(X)$
- Mẫu cụ thể $w = (20, 21, 20, 23, 23, 24, 22, 24, 22, 22)$

| Khối lượng (g) | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|
| Số sản phẩm | 2 | 1 | 3 | 2 | 2 |

# Thống kê đặc trưng mẫu

- Một hàm của các biến ngẫu nhiên $X_i$ trong mẫu là một thống kê (*statistic*)

$$G = G(X_1, X_2, \dots, X_n)$$

- Vì mẫu ngẫu nhiên nên $G$ là biến ngẫu nhiên với phân phối xác suất xác định
- Mẫu cụ thể: thống kê là số cụ thể, giá trị quan sát

$$G_{qs} = g = G(x_1, x_2, \dots, x_n)$$

- Thống kê trong mẫu thường tương ứng với một tham số trong tổng thể

# 7.2. TRUNG BÌNH MẪU

- Trung bình mẫu ngẫu nhiên (*sample mean*)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- $\bar{X}$ là biến ngẫu nhiên:

$$E(\bar{X}) = \mu; \;\; V(\bar{X}) = \frac{\sigma^2}{n}; \;\; \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Nếu $X \sim N(\mu, \sigma^2)$ thì:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right); \;\; Z = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$$

# Ví dụ

- **Ví dụ 7.2.** Chiều dài của một loại sản phẩm là một biến ngẫu nhiên phân phối Chuẩn với trung bình là 100 cm và phương sai là 16 $cm^2$.

Kiểm tra ngẫu nhiên 25 sản phẩm.

- a) Tính khả năng để chiều dài trung bình của 25 sản phẩm này lớn hơn 102 cm.
- b) Với xác suất là 0,95 thì chiều dài trung bình của 25 sản phẩm trên tối đa là bao nhiêu?

# 7.3. PHƯƠNG SAI MẪU

- Độ lệch bình phương trung bình (*mean of squares*)

$$MS = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

- Phương sai mẫu (*sample variance*) $\boldsymbol{S^2}$

$$\boldsymbol{S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \frac{n}{n-1}MS$$

- Độ lệch chuẩn mẫu: $S = \sqrt{S^2}$

- Ta có : $E(S^2) = \sigma^2$ và $E(MS) = \frac{n-1}{n}\sigma^2$

# 7.4. TỈ LỆ MẪU

- Trong mẫu kích thước $n$ có $X_A$ phần tử có dấu hiệu (biến cố) $A$ thì tỉ lệ mẫu

$$\hat{p} = \frac{X_A}{n}$$

- Nếu $P(A) = p$ thì: $E(\hat{p}) = p$ ; $V(\hat{p}) = \frac{p(1-p)}{n}$
- Mẫu kích thước $n \geq 100$

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \quad \text{và} \quad Z = \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$$

# Ví dụ

- **Ví dụ 7.3.** Tỉ lệ phế phẩm của nhà máy A là 10%. Lấy ngẫu nhiên 100 sản phẩm của nhà máy này.
- a) Tính xác suất để trong số sản phẩm lấy ra có ít nhất là 15% phế phẩm.
- b) Với xác suất là 0,9 thì trong mẫu trên có tối đa là bao nhiêu phế phẩm?

# 7.5. PHÂN PHỐI XÁC SUẤT

- Sử dụng thống kê trong mẫu để phản ánh về tham số trong tổng thể.
- Phân phối xác suất thể hiện mối liên hệ giữa các đại lượng này, phụ thuộc vào phân phối xác suất của biến ngẫu nhiên gốc $X$
- Dấu hiệu định lượng: thường qua biến phân phối Chuẩn $N(\mu, \sigma^2)$
- Dấu hiệu định tính: qua biến Không một $A(p)$
- **Các phân phối xác suất phục vụ suy diễn thống kê, ước lượng và kiểm định tham số**

# Phân phối xác suất

| Đại lượng | Suy đoán về mẫu | Ước lượng và kiểm định tham số |
|---|---|---|
| Trung bình | $Z = \dfrac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \sim N(0,1)$ | $T = \dfrac{(\bar{X}-\mu)\sqrt{n}}{S} \sim T(n-1)$ |
| Phương sai | $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^{2(n-1)}$ | $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^{2(n-1)}$ |
| Tỉ lệ $n \geq 100$ | $Z = \dfrac{(\hat{p}-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$ | $Z = \dfrac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \sim N(0,1)$ |

# Suy đoán về trung bình mẫu (Tự đọc)

- Ta có: $X \sim N(\mu, \sigma^2)$ rút ra mẫu ngẫu nhiên kích thước $n$ xây dựng được thống kê $\bar{X}$ và $S^2$ với $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Các công thức suy diễn trung bình mẫu:

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \bar{X} < \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-\infty < \bar{X} < \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha}\right) = 1 - \alpha$$

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} z_{\alpha} < \bar{X} < +\infty\right) = 1 - \alpha$$

# Suy đoán về tỉ lệ mẫu (tự đọc)

- Các công thức suy diễn trung bình mẫu:

$$P\left(p-\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha/2}<\hat{\boldsymbol{p}}<p+\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha/2}\right)=1-\alpha$$

$$P\left(0<\hat{\boldsymbol{p}}<p+\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha}\right)=1-\alpha$$

$$P\left(p-\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha}<\hat{\boldsymbol{p}}<1\right)=1-\alpha$$

# BÀI 8 - ƯỚC LƯỢNG THAM SỐ

- **8.1. Ước lượng tham số tổng thể**
- **8.2. Ước lượng điểm**
- **8.3. Ước lượng bằng khoảng tin cậy**

[1] Chương 7, trang 389 – 420, 431 – 437, 440 – 445

[2] Mục 6, trang 46 - 54

[3] Chapter 7, pp. 284 - 327

# 8.1. KHÁI NIỆM

- Trong tổng thể, $X$ đã biết phân phối xác suất nhưng tham số $\boldsymbol{\theta}$ (tham số tổng thể) **chưa biết.**
- Từ mẫu $\rightarrow$ ước lượng tham số $\theta$ (*parameter estimate*)
- Ước lượng điểm (*point estimate*) một giá trị $\hat{\theta}$
  - Mẫu ngẫu nhiên: ƯL ngẫu nhiên (*estimator*)
  - Mẫu cụ thể: ƯL cụ thể (*estimate*), hay giá trị quan sát (*observed value*)
- Ước lượng khoảng (*interval estimate*)

# 8.2. ƯỚC LƯỢNG ĐIỂM

- **Tính không chệch** (*unbiased*):
  - $\hat{\theta}$ là ước lượng không chệch của $\theta \Leftrightarrow E(\hat{\theta}) = \theta$
  - Nếu $E(\hat{\theta}) \neq \theta$ : ước lượng chệch
- **Tính hiệu quả** (*efficient*):

  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}^*, \hat{\theta}$ là ƯL không chệch
  - $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ thì $\hat{\theta}_1$ là **hiệu quả hơn** $\hat{\theta}_2$
  - $V(\hat{\theta}^*) < V(\hat{\theta})$ với mọi $\hat{\theta}$ thì $\hat{\theta}^*$ là **ƯL hiệu quả**
- Ước lượng không chệch, hiệu quả $\rightarrow$ ƯL tốt nhất
- **Tính vững** (*consistent*): khi $n \rightarrow \infty$ thì $\hat{\theta} \rightarrow \theta$ (theo nghĩa xác suất)

# Ví dụ

- **Ví dụ 8.1.**
- Tổng thể $X$ có trung bình là $\mu$, phương sai là $\sigma^2$
- Mẫu kích thước $n = 3$, trong các thống kê sau, đâu là ước lượng **không chệch**, đâu là ước lượng **hiệu quả hơn** cho $\mu$:

$$G_1 = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3$$
$$G_2 = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3$$
$$G_3 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$

# Bất đẳng thức Cramer - Rao

- Nếu BNN $X$ có công thức tính xác suất hoặc hàm mật độ là $f(x,\theta)$ thì với mọi $\hat{\theta}$ là ước lượng không chệch của $\theta$, luôn có:

$$V(\hat{\theta}) \geq \frac{1}{nE\left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^2}$$

- Do đó nếu $\hat{\theta}^*$ là ước lượng không chệch và có phương sai bằng vế phải của bất đẳng thức thì nó là ước lượng hiệu quả nhất

# Một số kết luận

- Khi $X \sim N(\mu, \sigma^2)$ thì
  - $\bar{X}$ là ước lượng không chệch, hiệu quả của $\mu$
  - $S^2$ là ước lượng không chệch của $\sigma^2$
  - $MS$ là ước lượng chệch, nhưng là ước lượng vững của $\sigma^2$
- Khi $X \sim A(p)$ thì $\hat{p}$ là ước lượng không chệch, hiệu quả của $p$.
- Cách thay thế $\boldsymbol{\mu}, \boldsymbol{p}, \boldsymbol{\sigma}^2$ bởi $\overline{\boldsymbol{X}}, \widehat{\boldsymbol{p}}, \boldsymbol{S}^2$ tương ứng gọi là ***tìm ước lượng điểm*** *theo hàm ước lượng.*

# Ước lượng hợp lý tối đa

- Mẫu $W = (X_1, X_2, \dots, X_n)$, mẫu cụ thể $(x_1, x_2, \dots, x_n)$
- Hàm hợp lý (*likelihood function*) với mẫu

$$L(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta)$$

- Giá trị $\hat{\theta}$ làm $L$ đạt cực đại là **ước lượng hợp lý tối đa** của $\theta$ (*maximum likelihood estimator*: *MLE*)
- Có thể tìm cực đại $L$ qua logarit của $L$
  - $X \sim N(\mu, \sigma^2)$ thì $\bar{X}$ là MLE của $\mu$; *MS* là MLE của $\sigma^2$
  - Khi $X \sim A(p)$ thì $\hat{p}$ là MLE của $p$

# Ví dụ

- **Ví dụ 8.2.**

a) Xác suất sinh viên đi làm thêm là $p = 0{,}4$. Trong các mẫu sau, mẫu nào hợp lý nhất? Biết 1 là có đi làm, 0 là ngược lại

$$w_1 = (0, 0, 0, 1, 0)$$
$$w_2 = (0, 1, 0, 0, 1)$$
$$w_3 = (1, 0, 0, 1, 1)$$

b) Có mẫu $(0, 1, 1, 0, 1)$ rút ra từ biến $A(p)$ chưa biết $p$. Trong các ước lượng sau cho $p$, giá trị nào hợp lý nhất

$$\hat{p}_1 = 0{,}5; \qquad \hat{p}_2 = 0{,}6; \qquad \hat{p}_3 = 0{,}7$$

# 8.3. ƯỚC LƯỢNG KHOẢNG

- Còn gọi là ước lượng bằng **khoảng tin cậy**
- Với mẫu ngẫu nhiên, tìm khoảng ngẫu nhiên $(G_1, G_2)$ để khả năng khoảng đó chứa $\theta$ bằng một mức xác suất cho trước:

$$P(G_1 < \theta < G_2) = 1 - \alpha$$

- Mức xác suất $(1 - \alpha)$ là **độ tin cậy** (*confidence level*)
- $(G_1, G_2)$ là **khoảng tin cậy** (*confidence interval, CI*)
- $I = G_2 - G_1$ là **độ dài khoảng tin cậy** (*width*)

# Xây dựng khoảng tin cậy

- Xét thống kê $G$ liên kết giữa tham số và thống kê trong mẫu, $G$ có phân phối xác suất xác định
- Với $(1-\alpha) \rightarrow$ xác định $\alpha_1$ và $\alpha_2 : \alpha_1 + \alpha_2 = \alpha$
- Xác định giá trị tới hạn $g_{1-\alpha_1}$ và $g_{\alpha_2}$ sao cho:

$$P\left(g_{1-\alpha_1} < G < g_{\alpha_2}\right) = 1 - \alpha$$

- Biến đổi thu được khoảng $(G_1, G_2)$ thỏa mãn:

$$P(G_1 < \theta < G_2) = 1 - \alpha$$

# Ước lượng trung bình tổng thể

- Tổng thể phân phối Chuẩn, $X \sim N(\mu, \sigma^2)$
- Tham số $\mu$ là trung bình tổng thể chưa biết
- Ước lượng khoảng cho $\mu$ với độ tin cậy $(1-\alpha)$
- Mẫu $W = (X_1, X_2, \dots, X_n)$
- Chia hai trường hợp:
  - Khi đã biết $\sigma \to$ dùng thống kê $Z$
  - Khi chưa biết $\sigma \to$ Sử dụng $S$ để thay, dùng thống kê

$$T = \frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim T(n-1)$$

# Ước lượng $\mu$ khi biết $\sigma^2$

- Do $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ với $\begin{cases} 0 \le \alpha_1, \alpha_2 \le \alpha \\ \alpha_1 + \alpha_2 = \alpha \end{cases}$

$$\to P\left(z_{1-\alpha_1} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha_2}\right) = 1-\alpha$$

$$\to P\left(\bar{X} - z_{\alpha_2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha_1}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Có 3 khoảng tin cậy thông dụng tương ứng với:

- (1) $\alpha_1 = \alpha, \alpha_2 = 0$
- (2) $\alpha_1 = 0, \alpha_2 = \alpha$
- (3) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

# Ước lượng $\mu$ khi biết $\sigma^2$

- Khoảng tin cậy tối đa (phía trái: *left-tailed*)

$$\mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Khoảng tin cậy tối thiểu (phía phải: *right-tailed*)

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu$$

- Khoảng tin cậy hai phía (đối xứng: *two-tailed*)

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

# Ước lượng $\mu$ khi biết $\sigma^2$

- Khoảng tin cậy đối xứng có dạng: $\bar{x} \pm \varepsilon$ hay $\bar{x} \pm ME$
- $\varepsilon$ là sai số biên (*ME*: *marginal error*): $\varepsilon = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$
- Độ dài khoảng tin cậy: $I = 2\varepsilon = 2\frac{\sigma}{\sqrt{n}} z_{\alpha/2}$
- Xác định kích thước mẫu $n_0$ thỏa mãn yêu cầu về sai số hoặc độ dài khoảng tin cậy:

$$\varepsilon \leq \varepsilon_0 \Leftrightarrow n_0 \geq \sigma^2 \frac{\left(z_{\alpha/2}\right)^2}{\varepsilon_0^2}$$

$$I \leq I_0 \Leftrightarrow n_0 \geq 4\sigma^2 \frac{\left(z_{\alpha/2}\right)^2}{I_0^2}$$

# Ước lượng $\mu$ khi không biết $\sigma^2$

- Khoảng tin cậy tối đa

$$\mu < \bar{x} + t_{\alpha}^{(n-1)} \frac{S}{\sqrt{n}}$$

- Khoảng tin cậy tối thiểu

$$\bar{x} - t_{\alpha}^{(n-1)} \frac{S}{\sqrt{n}} < \mu$$

- Khoảng tin cậy hai phía (đối xứng)

$$\bar{x} - t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}}$$

# Ước lượng $\mu$ khi không biết $\sigma^2$

- Khoảng tin cậy đối xứng: $\bar{x} \pm \varepsilon$ hay $\bar{x} \pm ME$
- Với $\varepsilon = ME = t_{\alpha/2}^{(n-1)} \dfrac{s}{\sqrt{n}}$
- Độ dài khoảng tin cậy: $I = 2\varepsilon = 2t_{\alpha/2}^{(n-1)} \dfrac{s}{\sqrt{n}}$
- Xác định kích thước mẫu $n_0$ theo mẫu sơ bộ kích thước $n$ đã có

$$\varepsilon \le \varepsilon_0 \Rightarrow n_0 \ge \frac{s^2 (t_{\alpha/2}^{(n-1)})^2}{\varepsilon_0^2}$$

$$I \le I_0 \Rightarrow n_0 \ge \frac{4s^2 (t_{\alpha/2}^{(n-1)})^2}{I_0^2}$$

# Ví dụ

- **Ví dụ 8.3.** Cân ngẫu nhiên 25 sản phẩm khối lượng trung bình là 25,32g và phương sai là 5,28g$^2$. Giả sử khối lượng phân phối chuẩn. Với độ tin cậy 95%
- a) Ước lượng khối lượng trung bình của tất cả các sản phẩm bằng khoảng tin cậy tối đa
- b) Tìm khoảng tin cậy đối xứng cho khối lượng trung bình
- c) Muốn sai số trong câu (b) còn không quá 0,5g thì cần cân thử thêm ít nhất bao nhiêu sản phẩm?

# Ước lượng phương sai tổng thể

- Tổng thể phân phối Chuẩn $X \sim N(\mu, \sigma^2)$
- Khoảng tin cậy tối đa $\sigma^2 < \dfrac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha}}$
- Khoảng tin cậy tối thiểu $\dfrac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha}} < \sigma^2$
- Khoảng tin cậy hai phía

$$\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha/2}}$$

# Ví dụ

- **Ví dụ 8.4.** Cân ngẫu nhiên 25 sản phẩm khối lượng trung bình là 25,32g và phương sai là 5,28g$^2$. Giả sử khối lượng của sản phẩm phân phối chuẩn.

  Với độ tin cậy 95%

- a) Độ dao động của khối lượng đo bởi phương sai tối đa là bao nhiêu?

- b) Tìm khoảng tin cậy cho độ lệch chuẩn của khối lượng sản phẩm

# Ước lượng tỉ lệ tổng thể $p$

- Khoảng tin cậy tối đa

$$p < \hat{p} + z_{\alpha}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

- Khoảng tin cậy tối thiểu

$$\hat{p} - z_{\alpha}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < p$$

- Khoảng tin cậy hai phía (đối xứng)

$$\hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < p < \hat{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

# Ước lượng tỉ lệ tổng thể $p$

- Khoảng tin cậy đối xứng: $\hat{p} \pm ME$ hay $\hat{p} \pm \varepsilon$

$$ME = \varepsilon = z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}};$$

$$I = 2ME = 2z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

- Xác định kích thước mẫu $n_0$ thỏa mãn yêu cầu về sai số hoặc độ dài KTC:

$$\varepsilon \le \varepsilon_0 \Leftrightarrow n_0 \ge \frac{\hat{p}(1-\hat{p})z_{\alpha/2}^2}{\varepsilon_0^2}$$

$$I \le I_0 \Leftrightarrow n_0 \ge \frac{4\hat{p}(1-\hat{p})z_{\alpha/2}^2}{I_0^2}$$

# Ví dụ

- **Ví dụ 8.5.** Quan sát ngẫu nhiên 400 người vào cửa hàng thì có 144 người mua hàng. Với độ tin cậy 95%:
- a) Ước lượng tỉ lệ khách mua hàng bằng khoảng tin cậy đối xứng
- b) Muốn độ dài khoảng tin cậy trong câu giảm xuống còn một nửa thì cần quan sát tối thiểu bao nhiêu người?
- c) Nếu trong một ngày có 5000 người vào cửa hàng thì có tối đa bao nhiêu người mua hàng?

# BÀI 9. KIỂM ĐỊNH GIẢ THUYẾT

- **9.1. Khái niệm**
- **9.2. Kiểm định tham số một tổng thể**
- **9.3. Kiểm định tham số hai tổng thể**
- **9.4. Kiểm định phi tham số**

[1] Chương 8, trang 465-470, trang 479

[2] Mục 7, trang 55 - 58

[3] Chapter 9, pp.346-384

# 9.1. KHÁI NIỆM

- **Giả thuyết thống kê**: Mệnh đề về một vấn đề thống kê nào đó về tổng thể.
- **Kiểm định tham số**: Kết luận về tính đúng / sai của một giả thuyết thống kê đối với tham số tổng thể dựa vào các bằng chứng thực nghiệm.
- **Ví dụ**

- Thu nhập trung bình của người lao động là trên 2000 USD/năm
- Tỷ lệ khách quay lại mua hàng lần hai là 50%
- Độ dao động của giá vàng trên thị trường tư nhân trong năm qua là chưa đến 30 USD

# Cặp giả thuyết, thống kê kiểm định

- Kiểm định so sánh tham số $\theta$ (của tổng thể) và số thực $\theta_0$ cho trước, có ba cặp giả thuyết

$$(1)\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases} \quad (2)\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta > \theta_0 \end{cases} \quad (3)\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{cases}$$

- Với mỗi cặp giả thuyết có một thống kê $G$ để kiểm định
- Thống kê tính trên mẫu cụ thể là $G_{qs}$

# Ví dụ

**Ví dụ 9.1.** Nêu cặp giả thuyết phù hợp để kiểm định giả thuyết tương ứng với các nhận định sau:

a) Thu nhập trung bình của người lao động là trên 3000 USD/năm.

b) Tỉ lệ khách quay lại mua hàng lần hai là 50%.

c) Độ dao động của giá vàng trên thị trường chưa đến 30 USD.

d) Trong số khách hàng, tỉ lệ nữ ít hơn nam.

e) Độ biến động của giá vàng vượt quá 500 $USD^2$.

# Các loại sai lầm

- Sai lầm loại I: bác bỏ một điều đúng (*type I error*)
- Sai lầm loại II: chấp nhận một điều sai (*type II error*)

| **Quyết định** | **Bản chất** $H_0$ đúng | $H_0$ sai |
|---|---|---|
| Chấp nhận $H_0$ | Đúng<br>Xác suất $= 1 - \alpha$ | Sai lầm loại II<br>Xác suất $= \beta$ |
| Bác bỏ $H_0$ | Sai lầm loại I<br>Xác suất $= \alpha$ | Đúng<br>Xác suất $= 1 - \beta$ |

- $\alpha$ và $\beta$ thay đổi trái chiều nhau.

# Mức ý nghĩa và miền bác bỏ

- Xác định một miền $W_\alpha$ sao cho nếu $H_0$ đúng thì xác suất $G$ thuộc miền đó là một mức α đủ nhỏ:

$$P(G \in W_\alpha \,|H_0 \text{ đúng}) = \alpha$$

- $W_\alpha$ gọi là miền bác bỏ (*reject area*)
- $W_\alpha$ được xác định bởi các giá trị tới hạn (*critical value*)
- $\alpha$ gọi là mức ý nghĩa (*significant level*)
- Mức $\alpha$ thường dùng là 1%, 5%, 10%

# Các bước kiểm định

- Từ mệnh đề $\rightarrow$ cặp giả thuyết $H_0, H_1$
- Tính thống kê $G_{qs}$ với mẫu cụ thể
- Với mức ý nghĩa $\alpha$ cho trước, xác định giá trị tới hạn và miền bác bỏ $W_\alpha$
- Quy tắc
  - $G_{qs} \in W_\alpha \rightarrow$ bác bỏ $H_0$ ($H_0$ sai)
  - $G_{qs} \notin W_\alpha \rightarrow$ chưa bác bỏ $H_0$ ($H_0$ đúng)
- Kết luận về mệnh đề ban đầu

# Kiểm định qua *P*-value

- *P*-value là “mức xác suất thấp nhất để bác bỏ $H_0$”
- *P*-value thường được tính sẵn qua các phần mềm chuyên dụng

- Quy tắc kiểm định theo *P*-value

Với mức ý nghĩa $\alpha$ cho trước:

  - Nếu *P*-value $< \alpha$ thì bác bỏ $H_0$
  - Nếu *P*-value $\geq \alpha$ thì chưa bác bỏ $H_0$

# 9.2. KIỂM ĐỊNH THAM SỐ MỘT TỔNG THỂ

- **Kiểm định trung bình tổng thể**
- **Kiểm định phương sai tổng thể**
- **Kiểm định tỉ lệ  tổng thể**

[1] Chương 8, trang 471-491, 512-519, 524-527

[3] Chapter 9, pp. 346 - 384

# Kiểm định trung bình tổng thể

- Tổng thể phân phối chuẩn $X \sim N(\mu, \sigma^2)$
- Tham số $\mu$ chưa biết, kiểm định so sánh $\mu$ với số $\mu_0$

Ba cặp giả thuyết

$$(1)\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \quad (2)\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases} \quad (3)\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

**Xét hai trường hợp**:

- Phương sai tổng thể $\sigma^2$ đã biết (lý thuyết)
- Phương sai tổng thể $\sigma^2$ chưa biết (thực tế)

# Kiểm định $\mu$ khi biết $\sigma^2$ (tự đọc)

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ | P-value |
|---|---|---|---|
| $Z_{qs} = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ | $H_0: \mu = \mu_0$<br>$H_1: \mu \neq \mu_0$ | $\lvert Z_{qs} \rvert > z_{\alpha/2}$ | $2P(Z > \lvert Z_{qs} \rvert)$ |
| | $H_0: \mu = \mu_0$<br>$H_1: \mu > \mu_0$ | $Z_{qs} > z_{\alpha}$ | $P(Z > Z_{qs})$ |
| | $H_0: \mu = \mu_0$<br>$H_1: \mu < \mu_0$ | $Z_{qs} < -z_{\alpha}$ | $P(Z < Z_{qs})$ |

# Ví dụ

**Ví dụ 9.2.** Biết kích thước sản phẩm là biến ngẫu nhiên phân phối chuẩn với phương sai là $36mm^2$. Đo ngẫu nhiên 50 sản phẩm tính được trung bình mẫu là 122mm. Với mức ý nghĩa 5%

a) Kiểm định giả thuyết kích thước trung bình là trên 120mm

b) P-value của cặp giả thuyết trong câu (a) thuộc khoảng nào?

c)* Tìm P-value của cặp giả thuyết trong câu (a)

d) Có thể nói kích thước trung bình chưa đến 123mm?

# Kiểm định $\mu$ khi chưa biết $\sigma^2$

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $T_{qs} = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ | $H_0: \mu = \mu_0$<br>$H_1: \mu \neq \mu_0$ | $\lvert T_{qs} \rvert > t_{\alpha/2}^{(n-1)}$ |
| | $H_0: \mu = \mu_0$<br>$H_1: \mu > \mu_0$ | $T_{qs} > t_{\alpha}^{(n-1)}$ |
| | $H_0: \mu = \mu_0$<br>$H_1: \mu < \mu_0$ | $T_{qs} < -t_{\alpha}^{(n-1)}$ |

**Ví dụ 9.3.** Cân ngẫu nhiên 25 sản phẩm khối lượng trung bình là 25,32g và phương sai là 5,28g$^2$. Giả sử khối lượng phân phối chuẩn. Với mức ý nghĩa 5%

a) Kiểm định giả thuyết trung bình tổng thể lớn hơn 24g

b) Có thể nói khối lượng trung bình là chưa đến 26g hay không? Nếu mức ý nghĩa là 10% thì sao?

c) Nhận xét ý kiến cho rằng khối lượng trung bình là khác 26,5g

# Kiểm định phương sai tổng thể $\sigma^2$

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$ | $H_0: \sigma^2 = \sigma_0^2$<br>$H_1: \sigma^2 \neq \sigma_0^2$ | $\chi_{qs}^2 > \chi_{\alpha/2}^{2(n-1)}$<br>hoặc $\chi_{qs}^2 < \chi_{1-\alpha/2}^{2(n-1)}$ |
| | $H_0: \sigma^2 = \sigma_0^2$<br>$H_1: \sigma^2 > \sigma_0^2$ | $\chi_{qs}^2 > \chi_{\alpha}^{2(n-1)}$ |
| | $H_0: \sigma^2 = \sigma_0^2$<br>$H_1: \sigma^2 < \sigma_0^2$ | $\chi_{qs}^2 < \chi_{1-\alpha}^{2(n-1)}$ |

# Ví dụ

**Ví dụ 9.4.** Tiêu chuẩn cho độ dao động  của khối lượng một loại quả đóng hộp là không được vượt quá 5g. Kiểm tra ngẫu nhiên 50 quả thu hoạch tại một vườn thấy phương sai mẫu của khối lượng quả là 30g$^2$.

Với mức ý nghĩa 5%, cho biết mức dao động của khối lượng loại quả tại vườn này là đạt tiêu chuẩn hay không? Giả thiết rằng khối lượng quả là biến ngẫu nhiên phân phối Chuẩn.

# Kiểm định tỉ lệ tổng thể $p$

| **Thống kê** | **Cặp giả thuyết** | **Bác bỏ $H_0$** |
|---|---|---|
| $Z_{qs} = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$<br><br>$n \geq 100$ | $H_0: p = p_0$<br>$H_1: p \neq p_0$ | $\lvert Z_{qs} \rvert > z_{\alpha/2}$ |
| | $H_0: p = p_0$<br>$H_1: p > p_0$ | $Z_{qs} > z_\alpha$ |
| | $H_0: p = p_0$<br>$H_1: p < p_0$ | $Z_{qs} < -z_\alpha$ |

# Ví dụ

**Ví dụ 9.5.** Trong số 400 người vào cửa hàng thì có 224 nữ và 176 nam.

Trong 224 nữ có 108 người mua hàng; trong 176 nam có 94 người mua hàng. Với mức ý nghĩa 5%:

a) Có thể nói tỷ lệ nữ chiếm trên một nửa số người vào cửa hàng hay không?

b) Có thể cho rằng tỷ lệ mua hàng của nữ là ít hơn của nam hay không?

# 9.3. KIỂM ĐỊNH THAM SỐ HAI TỔNG THỂ

- Khái niệm mẫu độc lập và mẫu phụ thuộc
- Kiểm định trung bình hai tổng thể
- Kiểm định phương sai hai tổng thể
- Kiểm định tần suất hai tổng thể

[1] Chương 8, trang 465 – 550

[2] Mục 7, trang 58 - 68

[3] Chapter 10, pp. 385 - 416

# Mẫu độc lập – mẫu phụ thuộc

- Mẫu độc lập: quan sát thu được từ các đối tượng độc lập hay khác nhau.
  - Số quan sát có thể khác nhau
  - Không quan trọng thứ tự các quan sát
- Mẫu phụ thuộc: là hai mẫu được chọn theo cách một quan sát bất kì ở mẫu thứ nhất tương ứng duy nhất với một quan sát ở mẫu thứ hai (quan hệ theo cặp)
  - Số quan sát phải bằng nhau
  - Thứ tự của các quan sát là cố định

# Ví dụ

- Hai mẫu phụ thuộc

| Cửa hàng | Trước quảng cáo | Sau quảng cáo |
|---|---|---|
| 1 | 72 | 76 |
| 2 | 75 | 79 |
| 3 | 70 | 77 |
| 4 | 82 | 80 |
| 5 | 70 | 75 |
| 6 | 83 | 89 |

**Quảng cáo thành công?**

- Hai mẫu độc lập

| Công ty A | Công ty B |
|---|---|
| 76 | 90 |
| 79 | 82 |
| 77 | 85 |
| 80 | 90 |
| 75 | 80 |
| 89 | 79 |
| | 87 |
| | 88 |

**Doanh số của A và B có sự khác biệt?**

# Kiểm định trung bình hai tổng thể

| | 1 | 2 |
|---|---|---|
| Tổng thể | $X_1 \sim N(\mu_1, \sigma_1^2)$ | $X_2 \sim N(\mu_2, \sigma_2^2)$ |
| Mẫu độc lập | $W_1$ kích thước $n_1$ | $W_2$ kích thước $n_2$ |
| Trung bình, phương sai | $\bar{x}_1, s_1^2$ | $\bar{x}_2, s_2^2$ |

- Với mức ý nghĩa $\alpha$, kiểm định so sánh $\mu_1$ và $\mu_2$
- Hai trường hợp:
  - Giả sử $\sigma_1^2 \neq \sigma_2^2$
  - Giả sử $\sigma_1^2 = \sigma_2^2$ : tự đọc trong giáo trình

# Kiểm định trung bình hai tổng thể

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $T_{qs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ | $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 \neq \mu_2$ | $\lvert T_{qs} \rvert > z_{\alpha/2}$ |
| | $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 > \mu_2$ | $T_{qs} > z_\alpha$ |
| $n_1, n_2 > 30$ | $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 < \mu_2$ | $T_{qs} < -z_\alpha$ |

# Kiểm định phương sai hai tổng thể

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $F_{qs} = \dfrac{s_1^2}{s_2^2}$ <br> $f_{1-\alpha}^{(m_1,m_2)} = \dfrac{1}{f_\alpha^{(m_2,m_1)}}$ | $H_0: \sigma_1^2 = \sigma_2^2$ <br> $H_1: \sigma_1^2 \neq \sigma_2^2$ | $F_{qs} > f_{\alpha/2}^{(n_1-1,n_2-1)}$ <br> hoặc $F_{qs} < f_{1-\alpha/2}^{(n_1-1,n_2-1)}$ |
| | $H_0: \sigma_1^2 = \sigma_2^2$ <br> $H_1: \sigma_1^2 > \sigma_2^2$ | $F_{qs} > f_{\alpha}^{(n_1-1,n_2-1)}$ |
| | $H_0: \sigma_1^2 = \sigma_2^2$ <br> $H_1: \sigma_1^2 < \sigma_2^2$ | $F_{qs} < f_{1-\alpha}^{(n_1-1,n_2-1)}$ |

# Kiểm định phương sai hai tổng thể

- Giả thuyết $\sigma_1^2 < \sigma_2^2$ hoán vị thành $\sigma_2^2 > \sigma_1^2$
- Chỉ xét với $s_1^2 > s_2^2$ thì bảng quyết định:

| $X_1 \sim N(\mu_1, \sigma_1^2)$<br>$X_2 \sim N(\mu_2, \sigma_2^2)$ | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $\boldsymbol{s_1^2 > s_2^2}$<br>$F_{qs} = \dfrac{s_1^2}{s_2^2}$ | $\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$<br>$\mathrm{H}_1: \sigma_1^2 > \sigma_2^2$ | $F_{qs} > f_{\alpha}^{(n_1-1, n_2-1)}$ |
| | $\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$<br>$\mathrm{H}_1: \sigma_1^2 \neq \sigma_2^2$ | $F_{qs} > f_{\alpha/2}^{(n_1-1, n_2-1)}$ |

# Ví dụ

- **Ví dụ 9.6.** Khảo sát ngẫu nhiên 40 khách hàng nam và 40 khách hàng nữ thấy khách nam chi trung bình là 230 nghìn và độ lệch chuẩn là 50 nghìn; khách nữ chi trung bình là 205 nghìn và độ lệch chuẩn là 60 nghìn. Giả sử chi tiêu phân phối chuẩn.
- a) Với mức ý nghĩa 5%, kiểm định giả thuyết mức chi trung bình của nam nhiều hơn nữ
- b) Với mức ý nghĩa 10%, mức chi của nam có đồng đều như nữ không?

# Kiểm định tỉ lệ hai tổng thể

| Thống kê | Cặp giả thuyết | Bác bỏ $H_0$ |
|---|---|---|
| $Z_{qs} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ | $H_0: p_1 = p_2$<br>$H_1: p_1 \neq p_2$ | $\lvert Z_{qs} \rvert > z_{\alpha/2}$ |
| $\bar{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ | $H_0: p_1 = p_2$<br>$H_1: p_1 > p_2$ | $Z_{qs} > z_{\alpha}$ |
| | $H_0: p_1 = p_2$<br>$H_1: p_1 < p_2$ | $Z_{qs} < -z_{\alpha}$ |

# Ví dụ

- **Ví dụ 9.7.** Quan sát số khách hàng vào ba cửa hàng A, B, C

| | A | B | C |
|---|---|---|---|
| Nữ | 55 | 115 | 85 |
| Nam | 45 | 55 | 35 |
| Tổng | 100 | 170 | 120 |

- a) Với mức ý nghĩa 5%, kiểm định giả thuyết "tỉ lệ khách hàng là nữ ở hai cửa hàng A và B là như nhau"
- b) Với mức ý nghĩa là 1% thì tỉ lệ khách hàng là nữ ở cửa hàng B nhỏ hơn tỉ lệ khách hàng là nữ ở cửa hàng C?

# 9.4. KIỂM ĐỊNH PHI THAM SỐ

Kiểm định phi tham số được sử dụng trong những trường hợp dữ liệu không có phân phối Chuẩn hoặc cho các mẫu nhỏ có ít quan sát.

- **Kiểm định Jarque - Bera (JB)**
- **Kiểm định tính độc lập**

[1] Chương 9, trang 552 – 556, 575 - 580

[2] Mục 8, trang 69 - 75

# Kiểm định Jarque – Bera (JB)

$H_0$: Biến $X$ phân phối chuẩn

$H_1$: Biến $X$ không phân phối chuẩn

- Hệ số bất đối xứng: $a_3 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^3/n}{S^3}$
- Hệ số nhọn: $a_4 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^4/n}{S^4}$
- Thống kê: $JB = n\left(\dfrac{a_3^2}{6} + \dfrac{(a_4 - 3)^2}{24}\right)$
- Bác bỏ $H_0$: $JB > \chi_{\alpha}^{2(2)}$

# Ví dụ

- **Ví dụ 9.8.** Với số liệu sau:

| Khối lượng (g) | 20-22 | 22-24 | 24-26 | 26-28 | 28-30 |
|---|---|---|---|---|---|
| Số sản phẩm | 2 | 5 | 8 | 7 | 3 |

- Tính được: $\bar{x} = 25{,}32$ và $s = 2{,}286$
- $\sum(x_i - \bar{x})^3 = -38{,}56; \sum(x_i - \bar{x})^4 = 568{,}63$
- Với mức ý nghĩa 5%, kiểm định giả thuyết khối lượng sản phẩm là biến ngẫu nhiên phân phối Chuẩn.

# Kiểm định tính độc lập

Hai dấu hiệu định tính A và B và bảng tiếp liên

- A gồm $h$ phạm trù: $A_1, A_2, \dots, A_h$
- B gồm $k$ phạm trù: $B_1, B_2, \dots, B_k$

| | $\mathbf{B_1}$ | $\mathbf{B_2}$ | ... | $\mathbf{B_k}$ | $\Sigma$ |
|---|---|---|---|---|---|
| $\mathbf{A_1}$ | $n_{11}$ | $n_{12}$ | ... | $n_{1k}$ | $\boldsymbol{n_1}$ |
| $\mathbf{A_2}$ | $n_{21}$ | $n_{22}$ | ... | $n_{2k}$ | $\boldsymbol{n_2}$ |
| ... | ... | ... | ... | ... | |
| $\mathbf{A_h}$ | $n_{h1}$ | $n_{h2}$ | ... | $n_{hk}$ | $\boldsymbol{n_h}$ |
| $\Sigma$ | $\boldsymbol{m_1}$ | $\boldsymbol{m_2}$ | ... | $\boldsymbol{m_k}$ | $n$ |

# Kiểm định tính độc lập

- Kiểm định cặp giả thuyết
  - $H_0$: A và B độc lập
  - $H_1$: A và B không độc lập
- Tiêu chuẩn

$$\chi^2 = n\left(\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i m_j} - 1\right)$$

- Miền bác bỏ: $W_\alpha = \left\{\chi^2 : \chi^2 > \chi_\alpha^{2((h-1)\times(k-1))}\right\}$

# Ví dụ

**Ví dụ 9.9.** Với mức ý nghĩa 5%, kiểm định tính độc lập của giới tính và loại tốt nghiệp của các cử nhân

| Loại TN / Giới | Trung bình | Khá | Giỏi | ∑ |
|---|---|---|---|---|
| Nữ | 90 | 150 | 40 | |
| Nam | 100 | 100 | 20 | |
| ∑ | | | | |