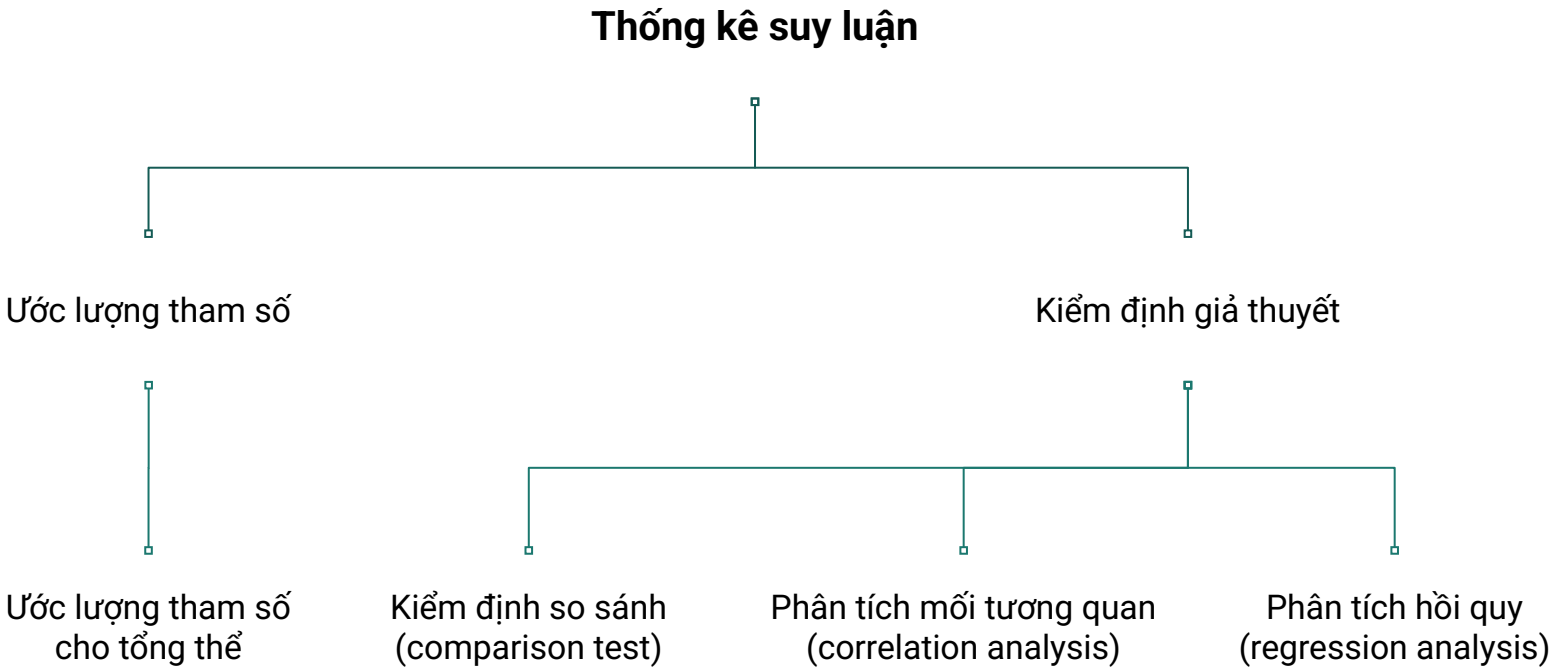
Thống kê suy luận
Inferential statistic

TS. Lưu Phúc Lợi
Trưởng phòng Nghiên cứu khoa học
Viện ARiHA

Các loại thống kê suy luận

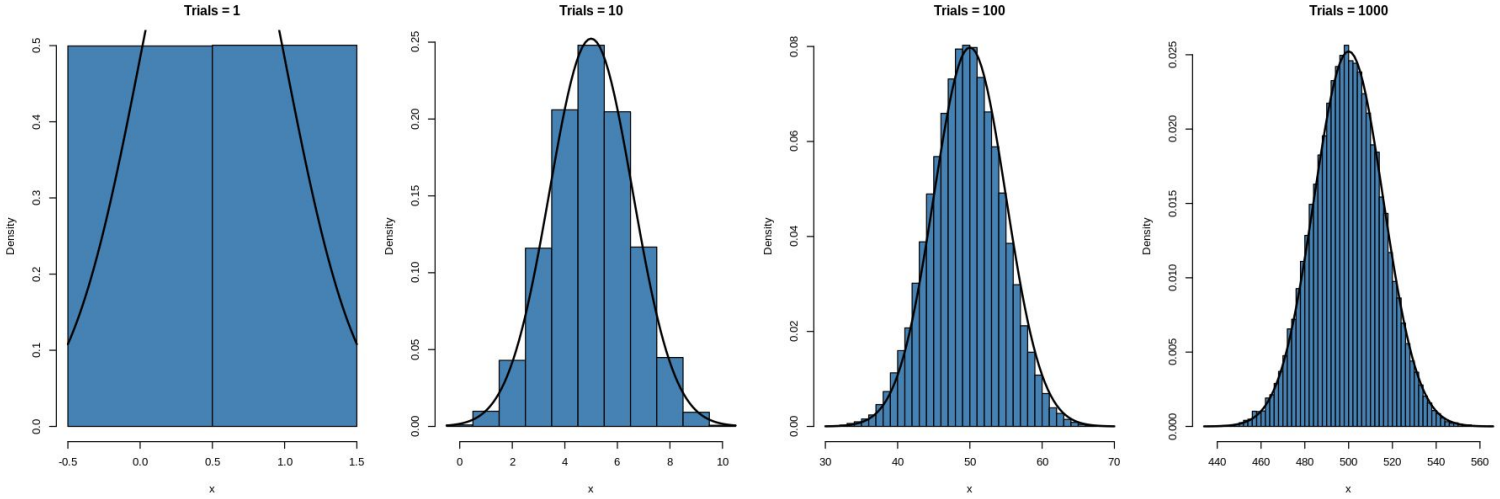


Cơ sở của thống kê suy luận

1. Định lý giới hạn trung tâm

Central Limit Theorem - CLT

"*Dưới các điều kiện thích hợp thì một phân phối xác suất khi được chuẩn hóa của trung bình mẫu sẽ hội tụ đến phân phối chuẩn tắc, dù phân phối ban đầu có thể không chuẩn.*"



Pólya, G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift, 8*, 171–181. https://doi.org/10.1007/BF01206525

2. Xác suất trung bình của mẫu

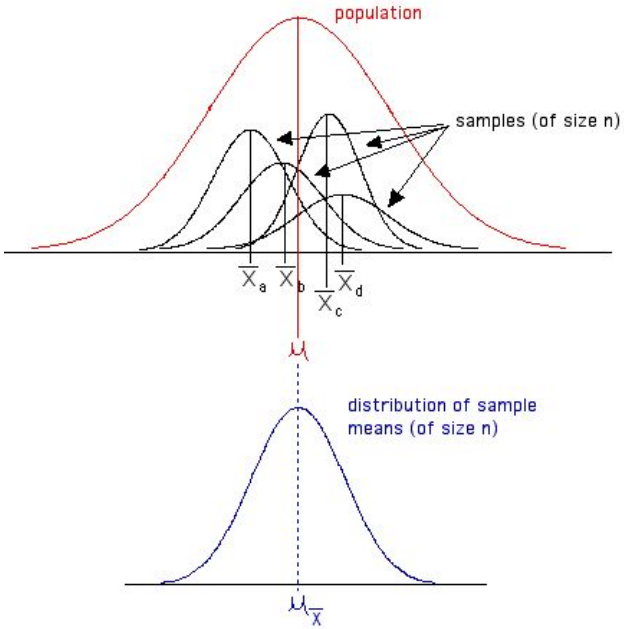
Sampling distribution of the mean

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

*Nếu ta lấy rất nhiều mẫu kích thước **n** từ cùng một quần thể, mỗi mẫu cho ra một trung bình. Tập hợp các trung bình đó tạo thành một phân phối.*

*Khi **n** tăng, trung bình mẫu có xu hướng tiến gần μ (trung bình thật của quần thể).*

Thống kê suy luận dùng phân phối này để ước lượng μ và định lượng độ không chắc chắn (SE, CI, p-value).



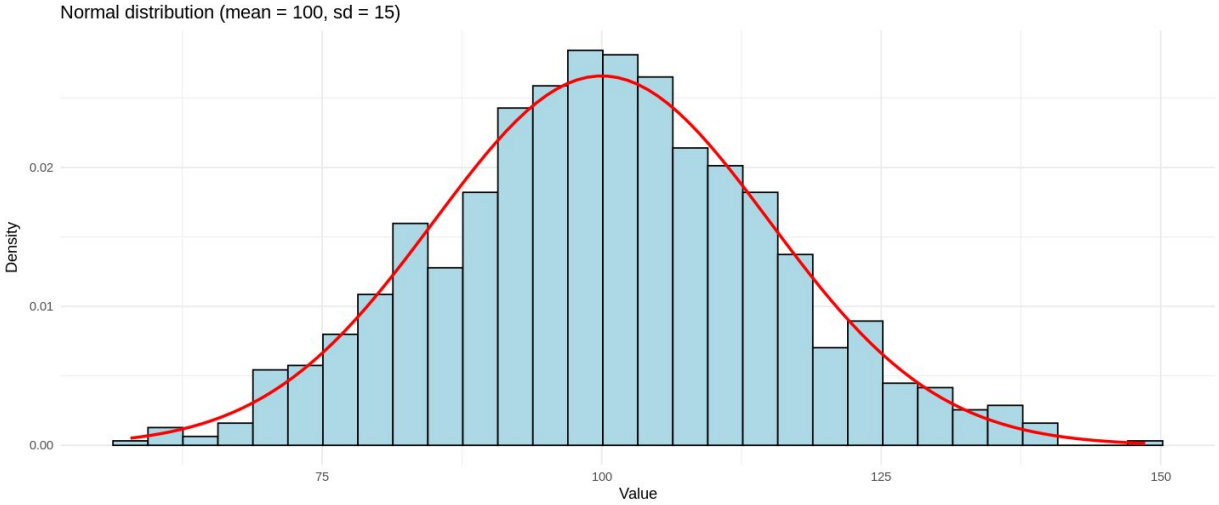
Bernoulli, J. (1713). *Ars Conjectandi*. Basel: Thurnisius.

Thế nào là phân phối chuẩn?

Phân phối chuẩn (normal distribution) là một phân bố xác suất liên tục có hình chuông đối xứng, tập trung quanh giá trị trung bình.

Đặc điểm:

- Đường cong đối xứng quanh mean (μ)
- Độ rộng phụ thuộc vào độ lệch chuẩn (σ)
- 68% giá trị nằm trong ±**1σ**, 95% nằm trong ±**2σ**, 99.7% trong ±**3σ** (quy tắc 68–95–99.7)

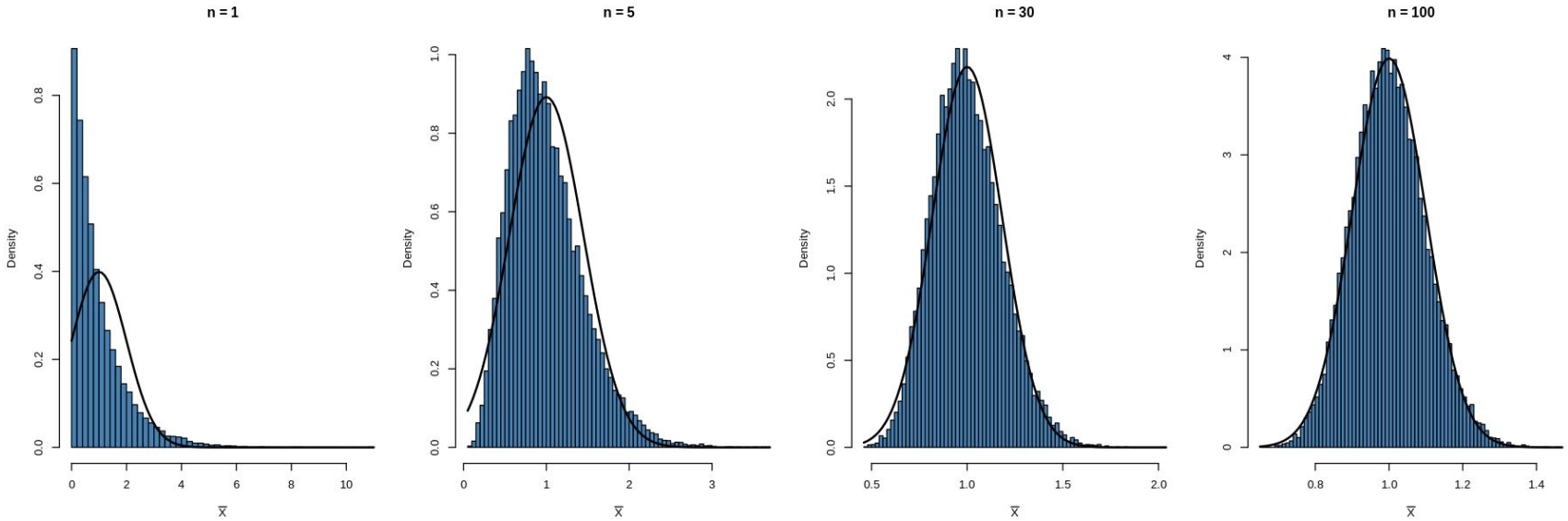


Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg: Sumptibus F. Perthes et I. H. Besser.

Thế nào là phân phối gần chuẩn?

Phân phối gần chuẩn (approximately normal) là phân phối của một đại lượng nào đó không đúng chuẩn tuyệt đối, nhưng hình dạng/CDF của nó rất gần với một phân phối chuẩn nào đó.

Mức "gần" đến đâu phụ thuộc vào: độ lệch (skewness), đuôi nặng (heavy tails), ngoại lệ, và **n**.



Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society, 49*(1), 122–136.
Esseen, C.-G. (1942). On the Liapounoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik, A28*, 1–19.

Giá trị của thống kê suy luận

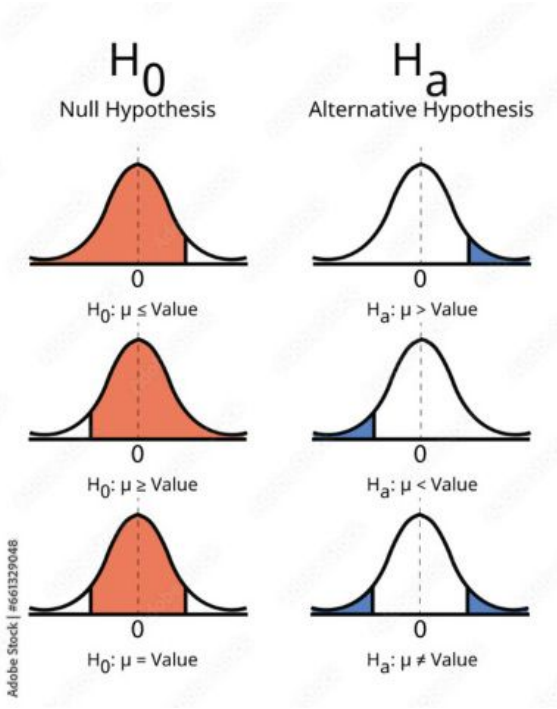
1. Ý nghĩa thống kê

p value

***H₀**: Giả thuyết không có khác biệt*
(mọi sai khác quan sát được trong mẫu chỉ do ngẫu nhiên/lấy mẫu và sai số đo chứ không phản ánh khác biệt thật ở quần thể)

- Ý nghĩa thống kê dưới dạng **p < α (0,05)**

- Vì **p < α** nên kết quả quan sát có ít hơn 5% giả định **H₀** đúng, ta nói kết quả có ý nghĩa thống kê và thường bác bỏ **H₀** => có sự khác biệt giữa hai nhóm nghiên cứu



Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Series 5, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Fisher, R. A. (1958). *Statistical methods for research workers* (13th rev. ed.). Hafner Publishing Company.

Ý nghĩa thống kê *(p-value)*

p nhỏ → dữ liệu **khó xảy ra** nếu H0 đúng → bằng chứng chống lại H0 mạnh hơn.

p lớn → dữ liệu chi thấy H0 xảy ra cao hơn → chưa có bằng chứng để bác bỏ H0.

Loại dữ liệu

1. Age: [years] -> numeric
2. Sex: [M: Male, F: Female] -> categorical
3. ChestPainType: [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic] -> categorical
4. RestingBP: [mm Hg] -> numeric
5. Cholesterol: [mm/dl] -> numeric
6. FastingBS: [1: if FastingBS > 120 mg/dl, 0: otherwise] -> categorical
7. RestingECG: [Normal: Normal, ST: having ST-T wave abnormality, LVH: left ventricular hypertrophy] -> categorical
8. MaxHR: [Numeric value between 60 and 202] -> numeric
9. ExerciseAngina: [Y: Yes, N: No] -> categorical
10. Oldpeak: [Numeric value measured in depression] -> numeric
11. ST_Slope: [Up: upsloping, Flat: flat, Down: downsloping] -> categorical
12. HeartDisease: [1: heart disease, 0: Normal] -> categorical

Lựa chọn phép kiểm định phù hợp

Biến phụ thuộc

1. Biến phụ thuộc là biến liên tục

- Nếu X là 2 nhóm → independent t-test (hoặc paired t-test)
- Nếu X là ≥3 nhóm → ANOVA

2. Biến phụ thuộc phân loại/nhị phân

- Nếu X cũng phân loại → Chi-square độc lập

Lựa chọn phép kiểm định phù hợp

Biến độc lập là kiểu phân loại hay liên tục?

1. Biến độc lập là kiểu phân loại: nhóm (Sex, nhóm tuổi, ChestPainType…) → t-test/ANOVA/Chi-square

2. Biến độc lập là kiểu liên tục: tuổi, cholesterol… → thường là correlation/regression.

1.1. T-test

Kiểm định sự khác biệt trung bình (mean) của một biến định lượng giữa hai điều kiện/nhóm.

- One-sample t-test: so sánh mean mẫu với một giá trị chuẩn μ_0.
- Independent two-sample t-test: so sánh mean giữa 2 nhóm độc lập (Group A vs Group B).
- Paired t-test: so sánh mean của chênh lệch trong thiết kế cặp đôi/ trước–sau (cùng đối tượng đo 2 lần).

Hãy tính t-test và diễn giải kết quả thống kê sau:

Biến	Giá trị
MaxHR ~ HeartDisease	
MaxHR ~ Sex	
MaxHR ~ FastingBS	

Ví dụ:

```
df <- read.csv("/content/Heart_Data.csv")
df$Cholesterol[df$Cholesterol=0] <- NA
t.test(Cholesterol ~ HeartDisease, data=df)$p.value
```

> 0.00474182506868301

Diễn giải: “So sánh nồng độ cholesterol huyết thanh giữa nhóm có bệnh tim và nhóm không có bệnh tim bằng kiểm định t độc lập (Welch). Kết quả cho thấy có sự khác biệt có ý nghĩa thống kê về giá trị trung bình cholesterol giữa hai nhóm (p = 0.0047)”

Diễn giải thống kê (ý nghĩa p-value)

Bạn đã thực hiện **kiểm định t hai mẫu độc lập** để so sánh **trung bình Cholesterol** giữa hai nhóm theo biến `HeartDisease` (có/không bệnh tim). Trong R, `t.test(y ~ group)` mặc định là **Welch's t-test** (không giả định phương sai hai nhóm bằng nhau).

- **Giả thuyết không (H0):** Trung bình Cholesterol ở hai nhóm `HeartDisease` là **bằng nhau**.
- **Giả thuyết đối (H1):** Trung bình Cholesterol ở hai nhóm là **khác nhau** (kiểm định hai phía).

Với **p-value = 0.00474**, nếu chọn mức ý nghĩa thường dùng **α = 0.05** (thậm chí α = 0.01), ta có:

- **Bác bỏ H0**.
- Có **bằng chứng thống kê** cho thấy **trung bình Cholesterol khác nhau** giữa hai nhóm `HeartDisease`.

Diễn giải đúng của p-value:

> Nếu trong thực tế **không có khác biệt trung bình** Cholesterol giữa hai nhóm, thì xác suất quan sát được (hoặc quan sát được kết quả “cực đoan hơn”) như dữ liệu hiện tại chỉ khoảng **0.47%**.

1.2. ANOVA

Analysis of Variance: Kiểm định sự khác biệt trung bình của một biến định lượng giữa từ 3 nhóm trở lên (hoặc nhiều mức của một/ nhiều yếu tố).

Các biến thể:

- One-way ANOVA: 1 yếu tố phân nhóm (k mức).
- Two-way / factorial ANOVA: ≥2 yếu tố, có thể có tương tác (interaction).
- Repeated-measures ANOVA: đo lặp theo thời gian/điều kiện trên cùng đối tượng (tương tự paired nhưng nhiều hơn 2 lần đo).

Lưu ý quan trọng: ANOVA “tổng quát hóa” t-test. Nếu chỉ có 2 nhóm thì one-way ANOVA và independent t-test cho p-value tương đương (trong điều kiện chuẩn).

Ví dụ:

So sánh MaxHR trung bình giữa các nhóm ChestPainType (TA/ATA/NAP/ASY)

```
df <- read.csv( "/content/Heart_Data.csv" )
fit <- aov(MaxHR ~ ChestPainType, data = df)
summary(fit)
```

```
>                Df Sum Sq Mean Sq F value Pr(>F)
ChestPainType   3  79522   26507   47.05 <2e-16 ***
Residuals     914 514903     563
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Diễn giải: “Vì p-value < 0.05 ta bác bỏ H0. Điều này cho thấy có bằng chứng thống kê rằng MaxHR trung bình khác nhau giữa ít nhất hai nhóm ChestPainType.”

Giải thích ý nghĩa các đại lượng trong bảng

- **Sum Sq (tổng bình phương)**:
  - ChestPainType: 79,522 là phần biến thiên của MaxHR “giải thích được” bởi khác biệt giữa các nhóm.
  - Residuals: 514,903 là phần biến thiên còn lại “trong nhóm” (không giải thích bởi ChestPainType).
- **Mean Sq (bình phương trung bình)**:
  - Giữa nhóm: 26,507
  - Trong nhóm: 563

  ANOVA dùng tỉ lệ hai giá trị này để tạo F.
- **F value = 47.05**:

  Đây là tỉ số $\frac{\text{MS}_{between}}{\text{MS}_{within}}$. Giá trị lớn cho thấy biến thiên giữa nhóm lớn hơn nhiều so với biến thiên trong nhóm.

1.3. Chi-square

X^2

- Chi-square test of independence (độc lập):
 Mục tiêu: kiểm định xem hai biến phân loại có liên hệ với nhau hay không (ví dụ: Giới tính × Bệnh tim).

- Chi-square goodness-of-fit (phù hợp):
 Mục tiêu: kiểm định xem phân phối tần số quan sát có khớp với một phân phối kỳ vọng đã cho hay không (ví dụ: tỉ lệ 50–50).

Ví dụ:

Kiểm định Sex có liên quan đến HeartDisease hay không

```
df <- read.csv( "/content/Heart_Data.csv" )
tab <- table(df$Sex, df$HeartDisease)

chisq.test(tab, correct =  TRUE)
chisq.test(tab, correct =  FALSE)
```

```
>      Pearson's Chi-squared test with Yates'
continuity correction
data:  tab
X-squared = 84.145, df = 1, p-value < 2.2e-16

      Pearson's Chi-squared test
data:  tab
X-squared = 85.646, df = 1, p-value < 2.2e-16
```

Diễn giải: Có bằng chứng thống kê rất mạnh rằng Sex có liên quan đến HeartDisease (hai biến không độc lập). Nói cách khác, tỷ lệ mắc HeartDisease khác nhau giữa nam và nữ trong mẫu dữ liệu này.

1) Bối cảnh và giả thuyết

Bạn đang làm **kiểm định Chi-square độc lập** cho bảng 2×2 (Sex × HeartDisease).

- H_0: Sex và HeartDisease **độc lập** (không liên quan); phân bố HeartDisease như nhau ở nam và nữ.
- H_1: Sex và HeartDisease **không độc lập** (có liên quan); tỷ lệ HeartDisease khác nhau theo giới.

---

2) Đọc hai kết quả bạn có

(A) Pearson's Chi-squared test with Yates' continuity correction

- $\chi^2 = 84.145, df = 1,$ p < 2.2×10⁻¹⁶

(B) Pearson's Chi-squared test không hiệu chỉnh (uncorrected)

- $\chi^2 = 85.646, df = 1,$ p < 2.2×10⁻¹⁶

Nhận xét thống kê:

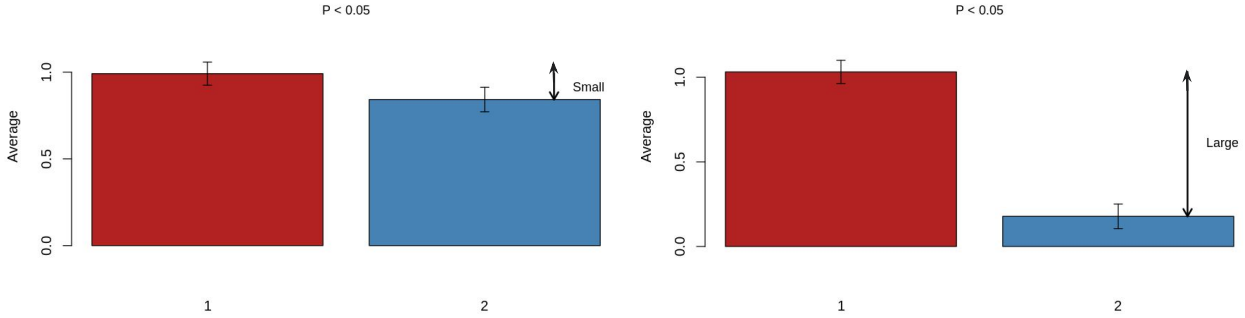
- Cả hai kiểm định đều cho p-value cực nhỏ → **bác bỏ** H_0 rất mạnh.
- Hai thống kê χ^2 hơi khác nhau vì:
  - **Yates' continuity correction** (hiệu chỉnh liên tục) thường áp dụng cho bảng 2×2 nhằm làm kiểm định "bảo thủ hơn" (giảm χ^2 một chút, p tăng nhẹ).
  - Dữ liệu của bạn lớn, nên hiệu chỉnh này **không làm thay đổi kết luận**.

So sánh tổng thể

Tiêu chí	t-test	ANOVA	Chi-square
Câu hỏi chính	Mean có khác nhau không?	Mean có khác nhau giữa ≥3 nhóm không?	Hai biến phân loại có liên hệ không? / Phân phối có phù hợp không?
Biến phụ thuộc (Y)	Định lượng (liên tục)	Định lượng (liên tục)	Phân loại (dữ liệu đếm)
Biến giải thích (X)	Nhóm 2 mức (hoặc trước–sau)	Nhóm ≥3 mức (hoặc nhiều yếu tố)	Phân loại (≥2 mức)
Thống kê kiểm định	t	F	χ^2
Khi nào dùng	2 nhóm (hoặc paired)	≥3 nhóm/đa yếu tố	Bảng tần số giữa các nhóm
Post-hoc	Không cần (vì 2 nhóm)	Thường cần nếu có ý nghĩa	Phân tích residual / so sánh cặp tỉ lệ (cẩn trọng)

Hạn chế

- Không cho biết mức độ khác nhau/tương quan
- Không cho biết độ chính xác của ước lượng
- Giá trị phụ thuộc vào cỡ mẫu

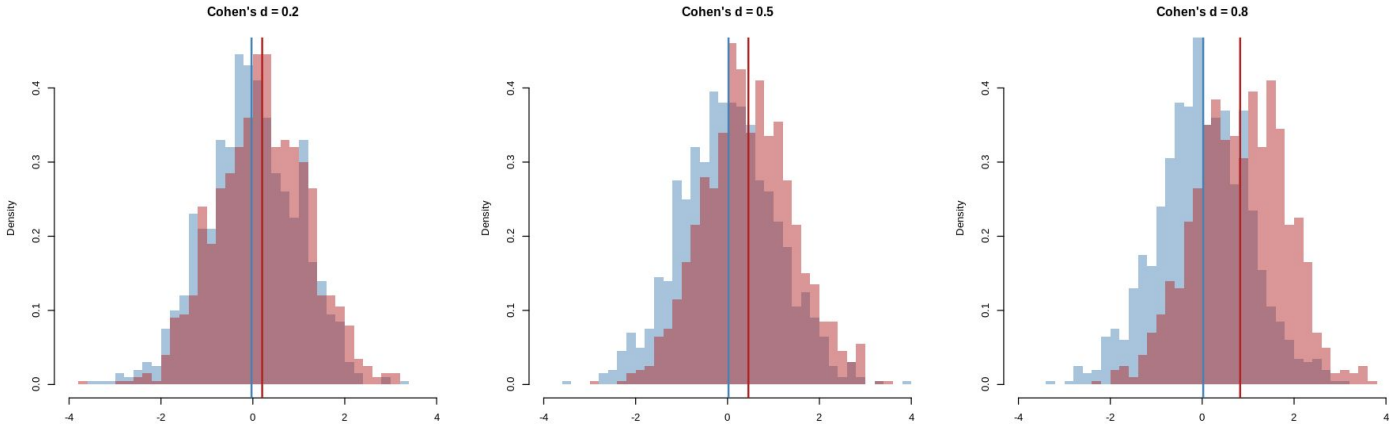


Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Series 5, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Fisher, R. A. (1958). *Statistical methods for research workers* (13th rev. ed.). Hafner Publishing Company.

2. Hệ số ảnh hưởng

2.1. Cohen's d - d



Cohen's d là một chỉ số thống kê dùng để đo mức độ khác biệt giữa hai nhóm theo đơn vị độ lệch chuẩn

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Sp là độ lệch chuẩn gộp (pooled SD)

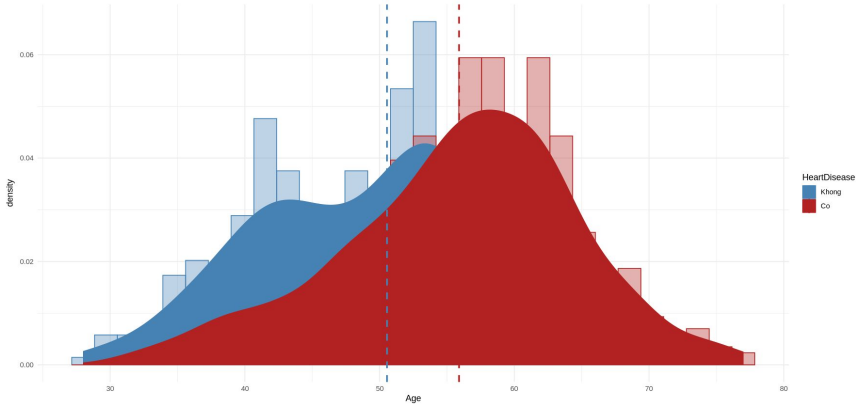
2.1. Cohen's d

Ví dụ: Tính hệ số d giữa biến Age và biến bệnh suy tim (HeartDisease)

```
df <- read.csv("/content/Heart_Data.csv")
df$HeartDisease <- factor(df$HeartDisease, levels = c(0,1), labels = c("Khong","Co"))
age1 <- df$Age[df$HeartDisease =="Khong"]
age2 <- df$Age[df$HeartDisease =="Co"]
x1 <- mean(age1)
x2 <- mean(age2)
s1 <- sd(age1)
s2 <- sd(age2)
sp <- sqrt( ((length(age1) - 1)*s1^2 + (length(age2) - 1)*s2^2) / (length(df$Age)- 2) )
d <- (x2 - x1) / sp
```

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

```
> n1 (Khong) = 410
> n2 (Co)    = 508
> x1 = 50.55122
> x2 = 55.89961
> s1 = 9.444915
> s2 = 8.727056
> sp = 9.05462
> d  = 0.5906804
```

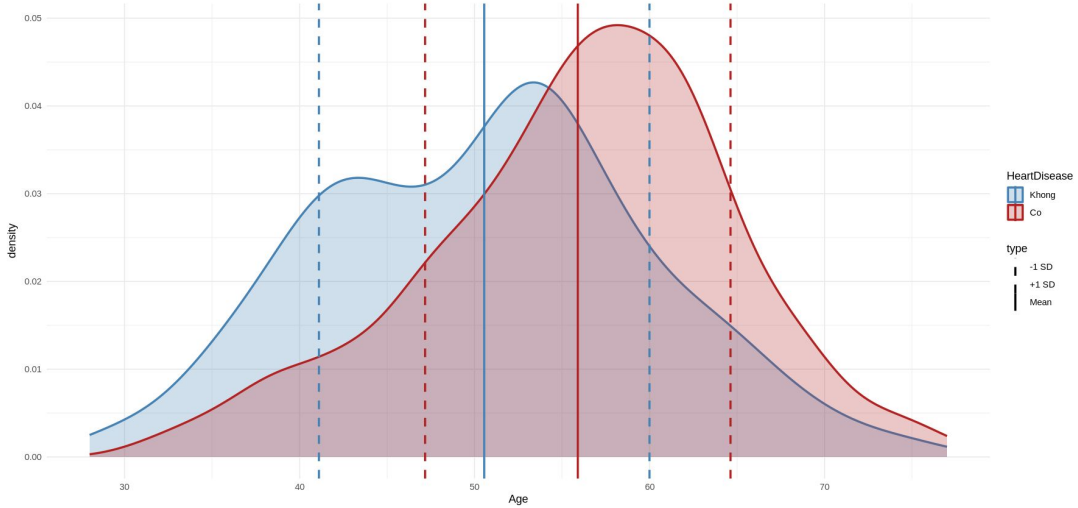


2.1. Cohen's d

Ví dụ: Tính hệ số d giữa biến Age và biến bệnh suy tim (HeartDisease)

Package: lsr

```
library(lsr)
heart = read.csv("/content/Heart_Data.csv")
cohensD(formula = Age ~ HeartDisease, data = heart)
> 0.590680436457599
```

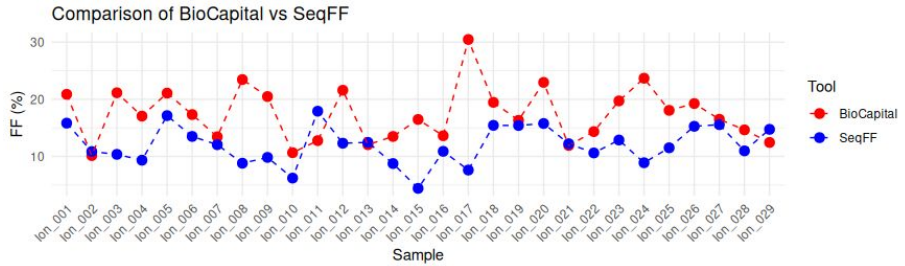


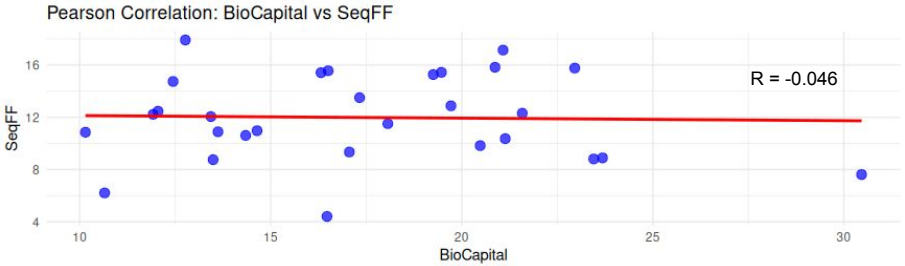
2.2. Pearson correlation

r

Hệ số đo mức độ và chiều hướng liên hệ tuyến tính giữa hai biến định lượng X và Y. Về bản chất, nó là hiệp phương sai được chuẩn hoá theo độ lệch chuẩn của từng biến, nên không có đơn vị đo và nằm trong khoảng [−1, 1]:

- **r = 1**: tuyến tính dương hoàn hảo
- **r = −1**: tuyến tính âm hoàn hảo
- **r = 0**: không có liên hệ tuyến tính





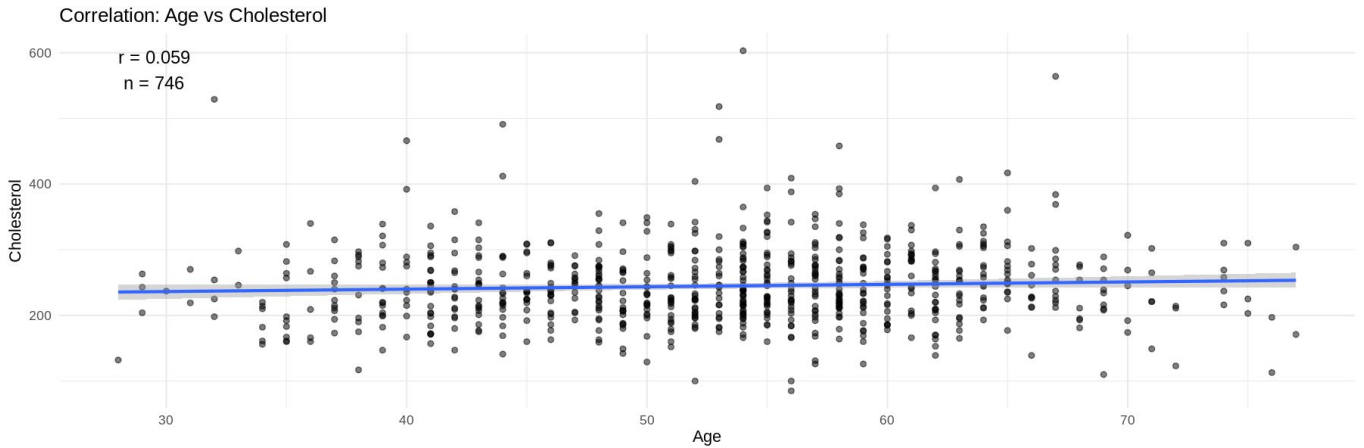
Pearson, K. (1896). *Mathematical contributions to the theory of evolution—III. Regression, heredity, and panmixia*. *Philosophical Transactions of the Royal Society of London. Series A*, *187*, 253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson correlation - *r*

Ví dụ: Tính hệ số r giữa biến Age và biến Cholesterol

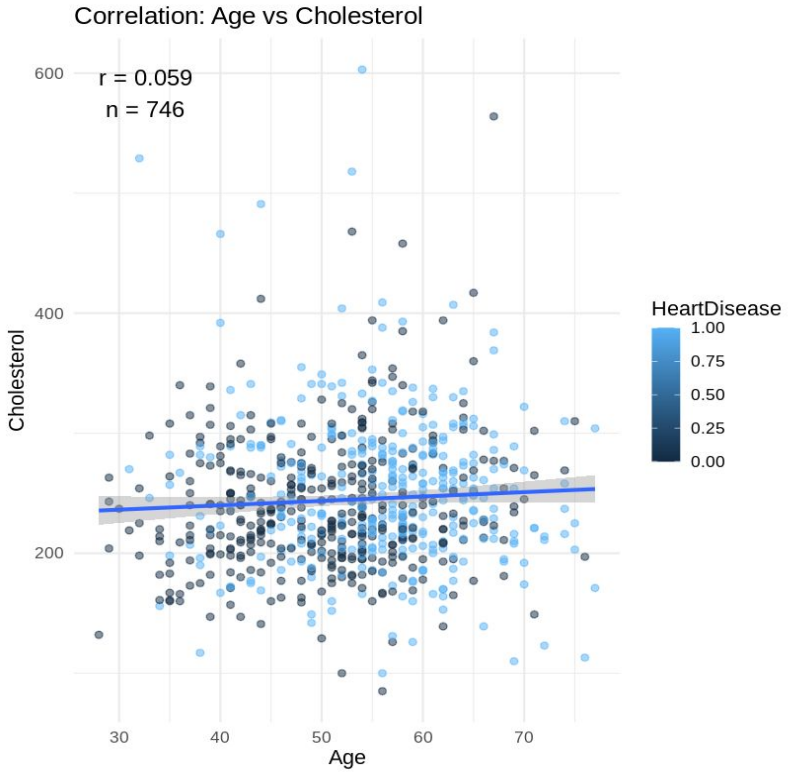
```
df <- read.csv("/content/Heart_Data.csv")
df$Cholesterol[df$Cholesterol == 0] <- NA
##fuction cor()
r <- cor(df$Age, df$Cholesterol, method = "pearson", use = "complete.obs")
cat("Pearson r =", r, "\n")
```

```
> Pearson r = 0.05875824
```



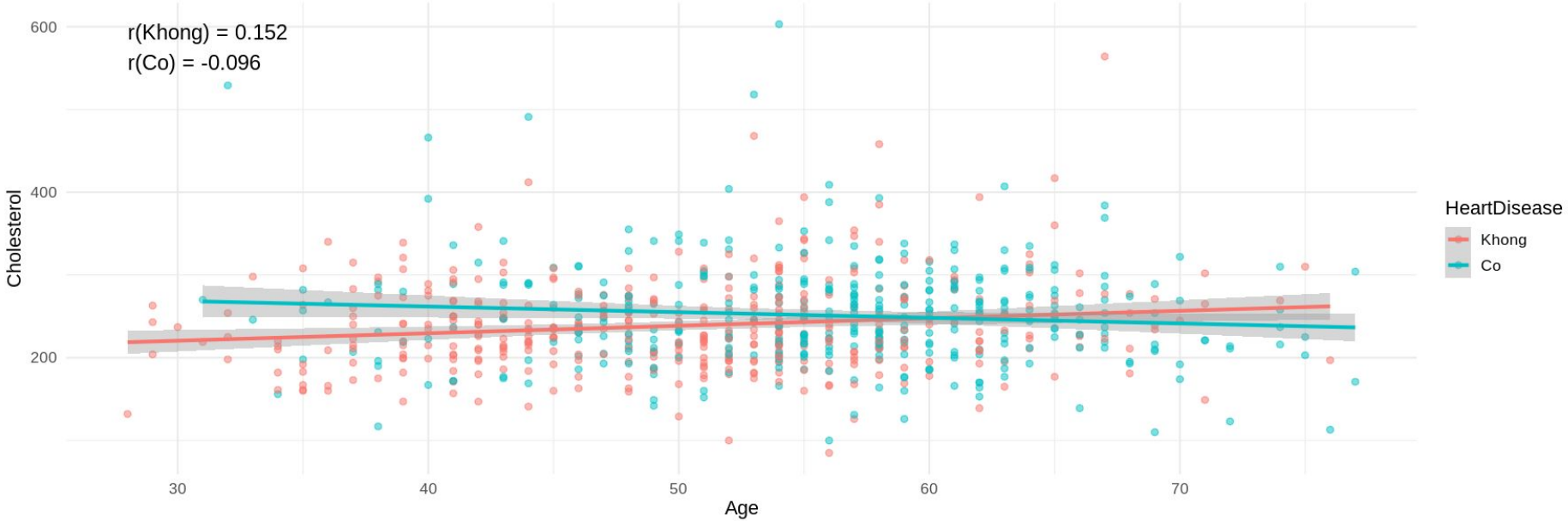
Pearson correlation - *r*

Ví dụ: Tính hệ số r giữa biến Age và biến Cholesterol, các điểm giá trị thể hiện mức độ bệnh/không bệnh



Pearson correlation - *r*

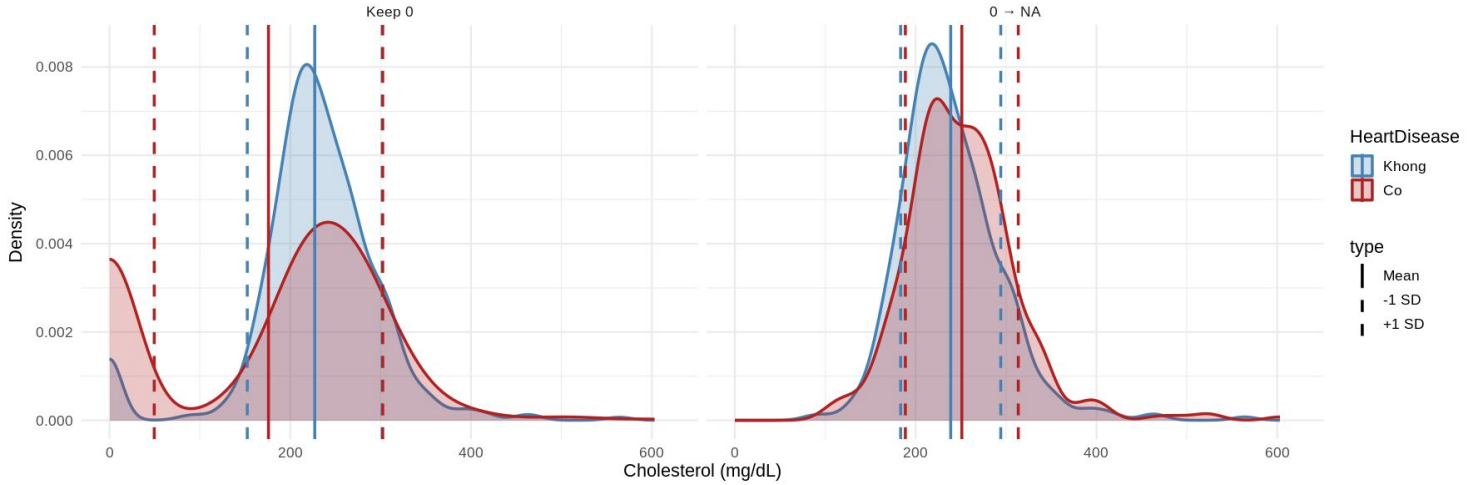
Ví dụ: Tính hệ số r giữa biến Age, biến Cholesterol với HeartDisease



Hạn chế

- Nhạy với outlier và phân phối lệch
- Không cho biết khác/giống nhau có ý nghĩa thống kê hay không
- **r** cao không đảm bảo dự đoán chính xác theo nghĩa lâm sàng; cần xem thêm hồi quy, RMSE, …

Giữ giá trị Cholesterol = 0
d = -0.481
Bỏ giá trị Cholesterol = 0 (NA)
d = 0.209
Số Cholesterol = 0 theo nhóm
0=No: 20
1=Yes: 152



3. Khoảng tin cậy

Neyman (1937) là người đưa ra khung lý thuyết "confidence limits / confidence coefficient / confidence interval": ông định nghĩa **confidence limits** là hai hàm của dữ liệu và **confidence coefficient** là xác suất bao phủ (coverage) của thủ tục đó; khoảng giữa 2 giới hạn là **confidence interval**.

"Tham số thật là hằng số, còn khoảng CI là biến ngẫu nhiên; vì vậy sau khi tính xong CI, việc "tham số nằm trong khoảng hay không" (với khoảng vừa tính) về mặt logic là 0 hoặc 1, không phải "95%""

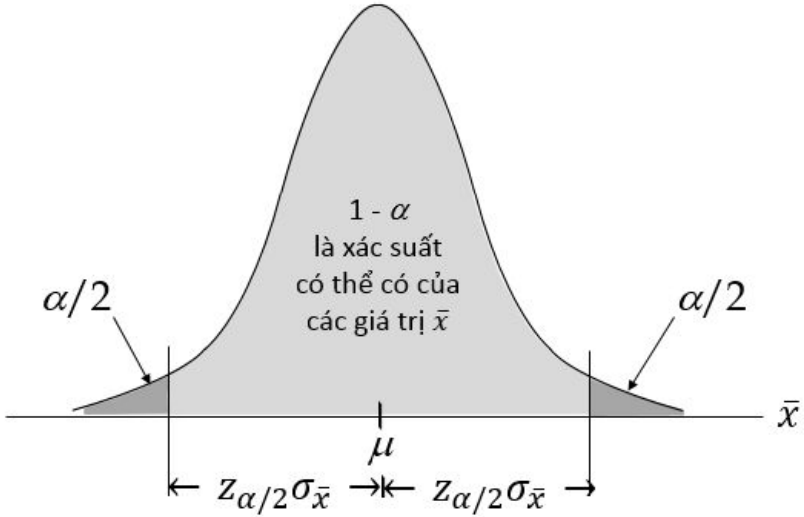
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333–380. https://doi.org/10.1098/rsta.1937.0005

3. Khoảng tin cậy

Confidence interval - CI

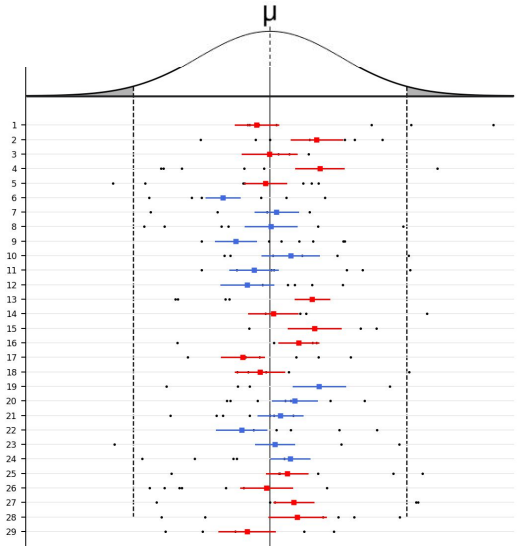
Là khoảng mà trung bình của các mẫu lấy ra từ tổng thể sẽ nằm trong khoảng đó (với độ tin cậy nhất định, thường chọn 95%)

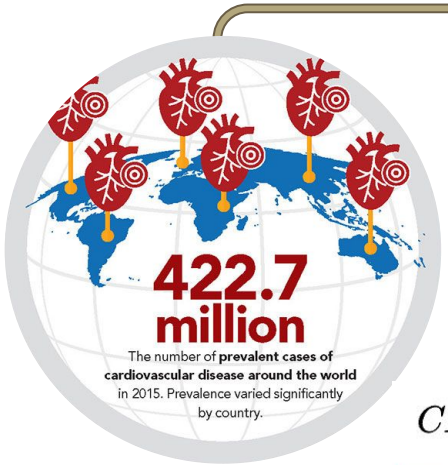
Nên hiểu rằng: nếu lặp lại lấy mẫu và dựng CI vô hạn lần, **~95% các khoảng** sẽ chứa tham số thật



Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333–380. https://doi.org/10.1098/rsta.1937.0005

Ví dụ





$$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

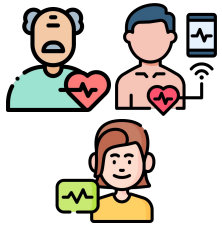
Mean value

Lower/Upper limit

z-value for the confidence level

Standard deviation

Sample size



Heart Rate

n = 918

X = 136.809

sd = 25.460

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333–380. https://doi.org/10.1098/rsta.1937.0005

Khoảng tin cậy - *CI*

Ví dụ: Tính hệ số tin cậy của MaxHR (CI95%) với 10 bin Cholesterol

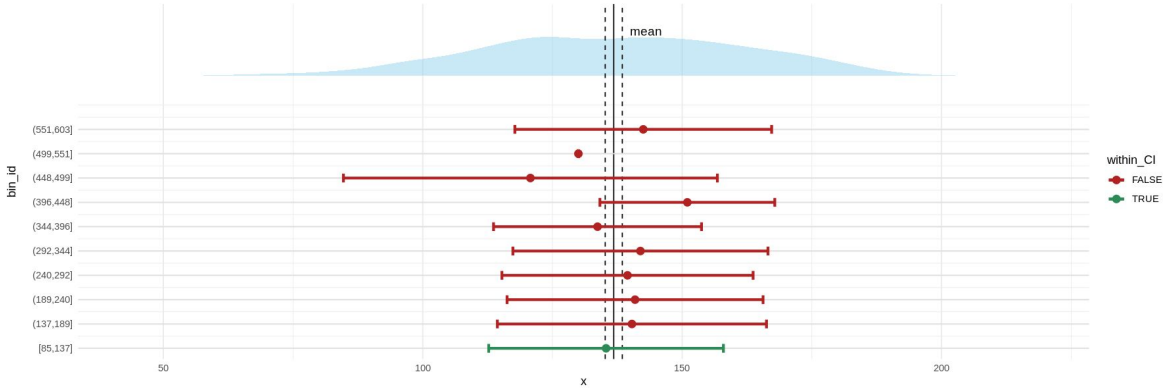
```
n = 918
xbar = 136.809
s = 25.460
se <- s / sqrt(n)

alpha <- 0.05
tcrit <- qt(1 - alpha/2, df = n - 1)
ME <- tcrit * se

CI_lower <- xbar - ME
CI_upper <- xbar + ME
```

```
> n = 125591
  mean = 136.809368191721
  sd = 25.4603341382503
  se = 0.0718430629550688
  tcrit = 1.9599828737327
  CI_lower = 136.668557018733
  CI_upper = 136.95017936471
```

Function t-test

```
df <- read.csv("/content/Heart_Data.csv")
t.test(df$MaxHR, conf.level = 0.95)$conf.int
```

```
> 135.16020190543    138.458534478012
```



Hạn chế

- CI trả lời “ước lượng + độ chắc chắn”, **không trả lời trực tiếp** “có đáng kể về mặt sinh học/lâm sàng không”

- Kết quả không ổn định với tham số không tuyến tính hoặc mô hình phức tạp

Lưu ý:

- Với chênh lệch trung bình: nếu CI 95% không chứa 0 → thường tương ứng $p < 0.05$

- Với OR/RR: nếu CI 95% không chứa 1 → thường tương ứng $p < 0.05$