KHÓA BỒI DƯỠNG GIẢNG VIÊN KHU VỰC MIỀN BẮC NĂM 2023

Nguyễn Thị Nhung

TRƯỜNG ĐẠI HỌC THĂNG LONG

Phần III

Kiểm định tính độc lập
Kiểm định về phân phối tổng thể

Nội dung trình bày

1. Kiểm định tính độc lập

2. Kiểm chứng mức phù hợp của một phân phối
   - Kiểm định Khi- bình phương
   - Kiểm định Kolmogorov-Smirnov
   - Kiểm định phân phối chuẩn

Bài toán tình huống

- *Kể từ khi xuất hiện, virus SARS-CoV-2 đã gây ra hàng loạt những ca tử vong trên toàn cầu. Đây được xem là đại dịch lớn nhất trên thế giới kể từ đại dịch cúm năm 1918.*
- *Qua một số nghiên cứu ở các nước Trung Quốc, Mỹ, Hàn Quốc cho thấy*
  - *Tỷ lệ mắc Covid-19 ở nam giới cao hơn nữ giới;*
  - *Nam giới mắc Covid-19 có xu hướng nặng hơn nữ giới;*
  - *Tỷ lệ tử vong ở nam giới cao hơn nữ giới.*

(1) (2)

---

[1] https://biomedic.com.vn/covid-19-va-gioi-tinh-dau-la-nhom-gap-nguy-co-lon-hon/

[2] Nguyên nhân: Do miễn dịch theo giới (hệ miễn dịch ở nữ giới phản ứng mạnh mẽ hơn với các bệnh truyền nhiễm; Do lối sống (nam giới hút thuốc nhiều hơn, ít rửa tay bằng xà phòng hơn,...))

? Câu hỏi

- Có sự liên hiện giữa giới tính và mức độ tử vong do Covid-19?
- Có sự liên hệ giữa số lượng tử vong do Covid-19 và độ tuổi? Tỷ lệ mức độ tử vong cao phân theo độ tuổi như thế nào?

Dữ liệu Covid-19 ở Mỹ

Một số thông tin về dữ liệu Covid-19 được dùng để minh họa trong những ví dụ dưới đây.

- `Link: https://www.kaggle.com/datasets/ramjasmaurya/covid-19-deaths-by-age-and-sextill-20-nov-2021`[(3)];
- `File: COVID-19_Deaths_by_Sex_and_Age(USA).csv`
- Dữ liệu về số ca tử vong do Covid-19 ở Mỹ được điều tra từ 1/1/2020 đến 20/11/2021 cho những thông tin chính về: Độ tuổi, Giới tính, Số ca tử vong do covid-19 và tử vong do một số bệnh khác.

Covid 19 deaths by Age and Sex(till 20 Nov 2021)

USA covid 19 deaths from 2020 to nov2021 by(Age,Sex and Disease)

[(3)] Data is taken from NHCS of USA(www.data.cdc.gov)

Ví dụ

Bài toán

Thống kê số ca tử vong do Covid-19 ở các bang của Mỹ trong thời gian 1/1/2020 đến 20/11/2021 theo giới tính và mức độ tử vong ta có bảng dữ liệu sau.

Giới tính	Mức độ tử vong				Tổng
	< 100	100-500	500-1000	> 1000	
Nữ	1067	495	217	391	2170
Nam	989	546	239	472	2246
Tổng	2056	1041	456	863	4416

Ở mức ý nghĩa $\alpha = 5\%$ hãy kiểm tra xem mức độ tử vong do Covid-19 có phụ thuộc vào giới tính không.

Kiểm định tính độc lập

Bài toán

Ta xét hai biến định tính và muốn kiểm tra xem mối quan hệ giữa chúng là độc lập hay phụ thuộc. Để thực hiện việc này ta kiểm định cặp giả thuyết sau:

H_0 : *Hai biến định tính độc lập (không có mối liên hệ giữa hai biến này);*
H_1 : *Hai biến định tính không độc lập (có mối liên hệ giữa hai biến này).*

Qui trình thực hiện

- Giả sử biến định tính thứ nhất gồm r loại, biến định tính thứ hai gồm c loại. Chọn từ tổng thể ra mẫu gồm n phần tử xếp chéo thành $r \times c$ giá trị $Oij, i = 1, \ldots, r; j = 1, \ldots, c$, trong đó O_{ij} là số quan sát có thuộc tính thứ i của biến thứ nhất và thuộc tính thứ j của biến thứ hai. Khi đó ta có bảng sau:

Biến thứ nhất	Biến thứ hai					Tổng
	1	2	3	...	c	
1	O_{11}	O_{12}	O_{13}	...	O_{1c}	R_1
2	O_{21}	O_{22}	O_{23}	...	O_{2c}	R_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	R_r
Tổng	C_1	C_2	C_3	...	C_c	n

Qui trình thực hiện

- Giả sử H_0 đúng, khi đó tần số lí thuyết E_{ij} của ô ở địa chỉ ij được tính theo công thức:

$$E_{ij} = \frac{R_i \times C_j}{n}.$$

Ta có giá trị kiểm định:

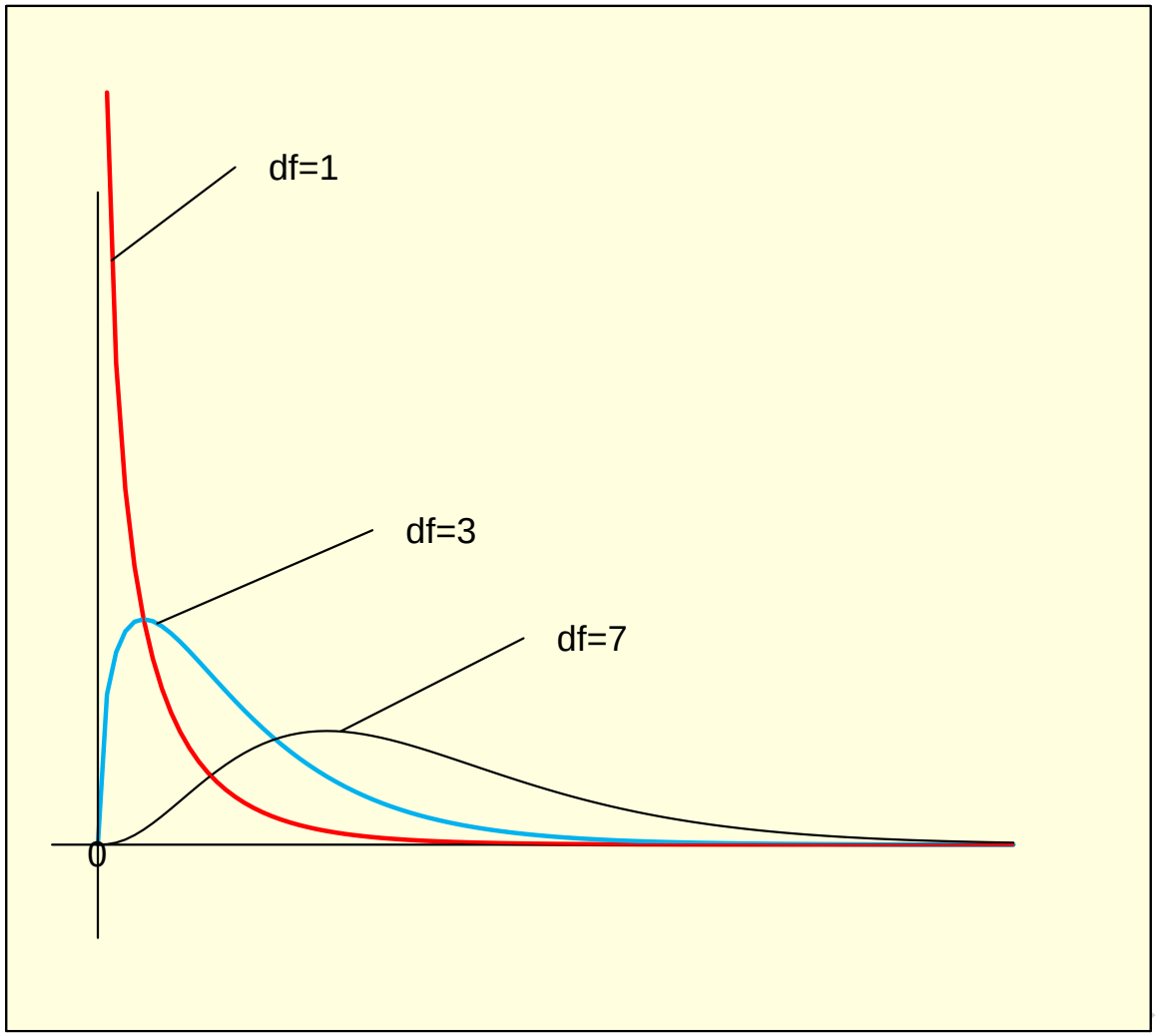
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- Nếu giả thuyết H_0 đúng và $E_{ij} \geqslant 5, \forall i, j$ thì χ^2 tuân theo phân phối khi- bình phương với $(r-1)(c-1)$ bậc tự do.
- So sánh giá trị kiểm định với $\chi^2_{(r-1)(c-1),\alpha}$, tại mức ý nghĩa α ta đưa ra quyết định bác bỏ H_0 nếu

$$\chi^2 > \chi^2_{(r-1)(c-1),\alpha}.$$

Phân phối Khi-bình phương

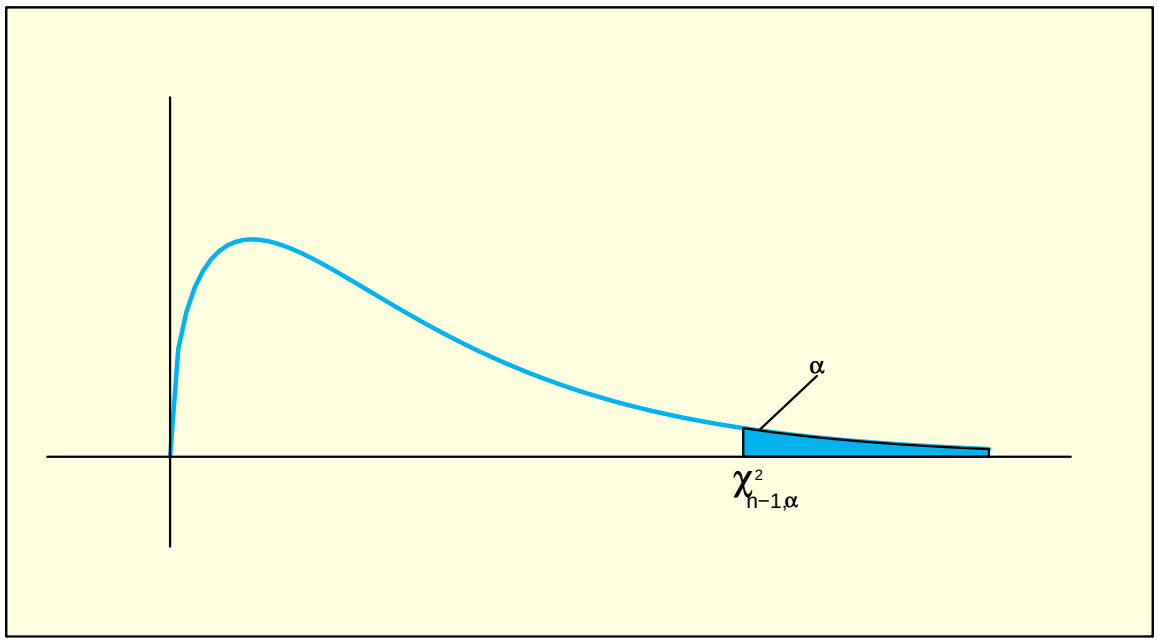
Mot so duong phan phoi khi binh phuong



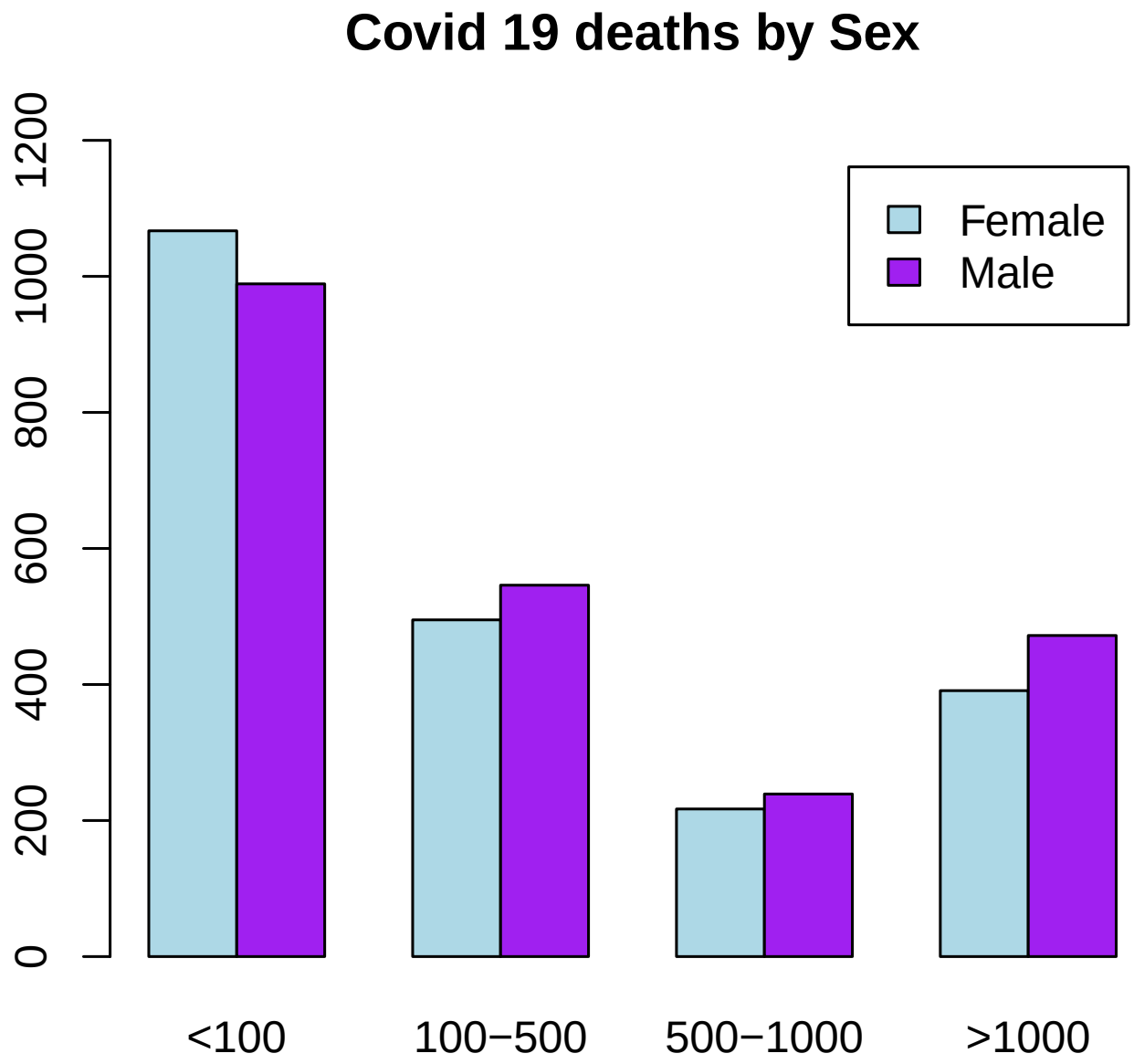
Giá trị tới hạn trong phân phối Khi- bình phương

Ký hiệu χ^2_n là biến ngẫu nhiên có phân phối khi-bình phương với bậc tự do là n. Hai giá trị tới hạn $\chi^2_{n,\alpha}$ và $\chi^2_{n,1-\alpha}$ [(4)] được định nghĩa bởi:

$$P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha, \qquad (\chi^2_n < \chi^2_{n,1-\alpha}) = \alpha$$



Sự phụ thuộc của Giới tính và Mức độ tử vong Covid-19



Ví dụ

- Cặp giả thuyết H_0, H_1 cần kiểm định:
  H_0 : Mức độ tử vong Covid-19 không phụ thuộc vào giới tính
  H_1 : Mức độ tử vong Covid-19 phụ thuộc vào giới tính

- Các giá trị tần số lí thuyết E_{ij} được cho tương ứng trong bảng sau:

Giới tính	Mức độ tử vong			
	< 100	100-500	500-1000	> 1000
Nữ	1010.31	511.54	224.08	424.08
Nam	1045.69	529.46	231.92	438.93

Ví dụ

- Tính toán giá trị kiểm định

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 12.817.$$

- Giá trị tới hạn $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{3,0.05} = 7.8$.
- Do $\chi^2 = 12.817 > 7.8 = \chi^2_{3,0.05}$ nên bác bỏ H_0. Như vậy ta có đủ bằng chứng thống kê để cho rằng giới tính ảnh hưởng đến mức độ tử vong Covid-19.

Kiểm định tính độc lập trong R bằng hàm `chisq.test()`[5]

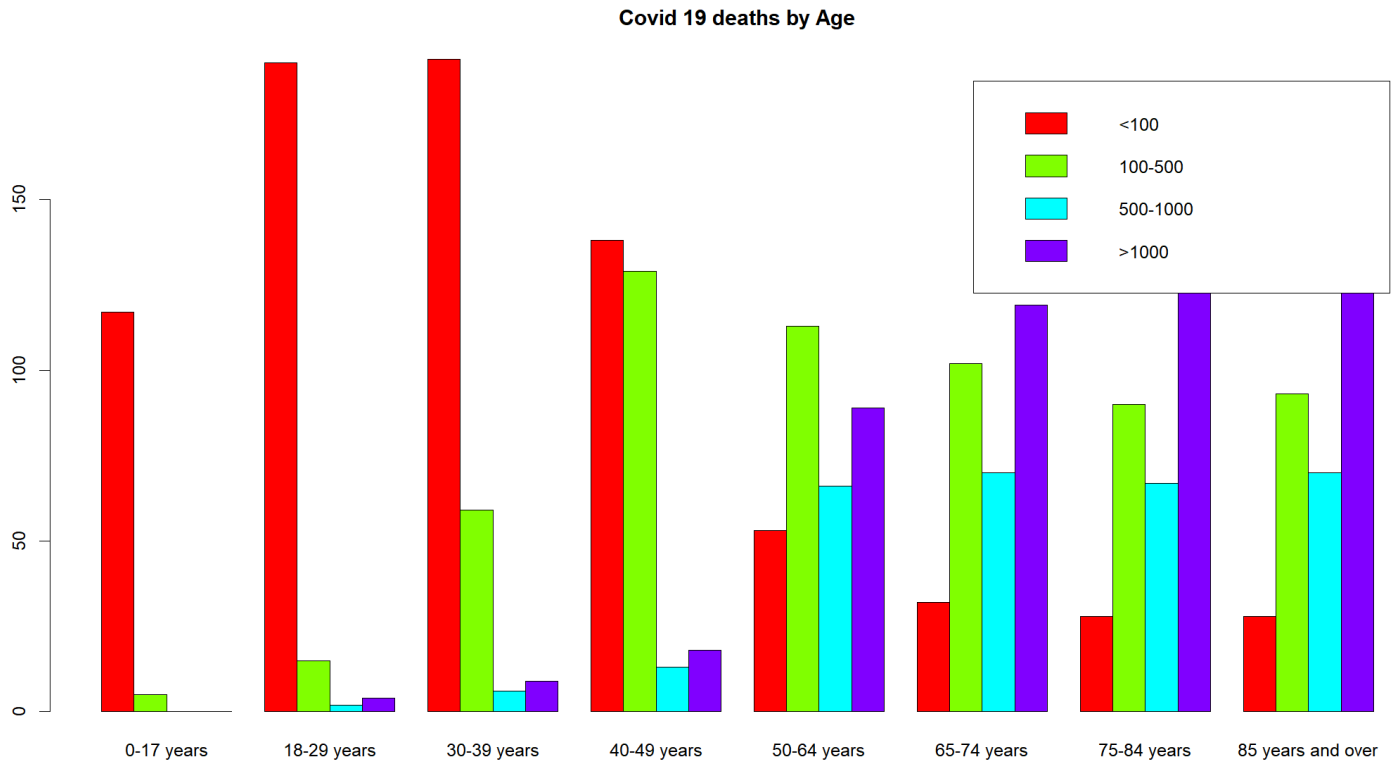
```
muc_tu_vong = c(1067, 495, 217, 391,  989, 546, 239,  472)
A = matrix(muc_tu_vong, nrow = 2, byrow=TRUE)
chisq.test(A)

        Pearson's Chi-squared test

data:  A
X-squared = 12.817, df = 3, p-value = 0.005048
```

[5] `chisq.test(A)`, A: ma trận các quan sát của hai biến cần kiểm định tính độc lập

Có sự liên hệ giữa số lượng tử vong do Covid-19 và độ tuổi? Tỷ lệ mức độ tử vong cao phân theo độ tuổi như thế nào?



Độ tuổi ảnh hưởng đến mức độ tử vong

```
chisq.test(table(COVID.19.Deaths_factor,  Age.Group))

        Pearson's Chi-squared test

data:  table(COVID.19.Deaths_factor, Age.Group)
X-squared = 1148.5, df = 21, p-value < 2.2e-16
```

Có thể cho rằng Mức độ tử vong cao do Covid-19 theo độ tuổi tuân theo một tỷ lệ cho trước?

Chẳng hạn, có thể cho rằng mức độ tử vong cao do covid-19 theo độ tuổi tuân theo tỷ lệ sau không?

18-29 years	30-39	40-49	50-64	65-74	$\geqslant$75 years
1	3	6	20	30	40

Nội dung trình bày

1. Kiểm định tính độc lập

2. Kiểm chứng mức phù hợp của một phân phối
   - Kiểm định Khi- bình phương
   - Kiểm định Kolmogorov-Smirnov
   - Kiểm định phân phối chuẩn

Kiểm định về qui luật phân phối xác suất của tổng thể

Bài toán

Giả sử ta chưa biết qui luật phân phối xác suất của tổng thể, ta cần kiểm định xem phân phối của tổng thể có tuân theo một qui luật xác suất A nào đó hay không bằng cách kiểm định cặp giả thuyết sau:

H_0 : *Tổng thể tuân theo qui luật xác suất A;*
H_1 : *Tổng thể không tuân theo qui luật xác suất A.*

Qui trình thực hiện

- Chọn một mẫu ngẫu nhiên gồm n phần tử, mỗi phần tử được xếp vào đúng một trong k lớp. Gọi O_i là số phần tử rơi vào lớp thứ i, trong đó $i = 1, 2, \ldots, k$.
- Nếu H_0 đúng, tức là tổng thể tuân theo qui luật xác suất A, thì xác suất để một phần tử rơi vào lớp $1, 2, \ldots, k$ lần lượt là $p_1, p_2, \ldots, p_k$ với $p_1 + p_2 + \ldots + p_k = 1$. Khi đó số phần tử kì vọng theo k lớp đó sẽ là $E_i = np_i, i = 1, 2, \ldots, k$.

Lớp	1	2	...	k	Tổng
Số phần tử quan sát	O_1	O_2	...	O_k	n
Xác suất theo H_0	p_1	p_2	...	p_k	1
Số phần tử theo H_0	$E_1 = np_1$	$E_2 = np_2$	...	$E_k = np_k$	n

Qui trình thực hiện

- Nếu H_0 đúng và cỡ mẫu lớn sao cho $E_i = np_i \geqslant 5, \forall i = \overline{1,k}$ thì biến ngẫu nhiên

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

tuân theo phân phối xấp xỉ phân phối khi- bình phương với $k - m - 1$ bậc tự do, trong đó m là số tham số tổng thể ước lượng từ dữ liệu mẫu.

- Tại mức ý nghĩa α, giả thuyết H_0 bị bác bỏ nếu

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-m-1,\alpha}.$$

Ví dụ

Bài toán

Thống kê 400 khoảng thời gian có mức độ tử vong cao (trên 1000 người) phân theo độ tuổi, ta được bảng dữ liệu sau.

Khoảng tuổi	18-39 years	40-49	50-64	65-74	⩾75 years	Tổng
Số khoảng	13	18	86	115	168	400

Tại mức ý nghĩa $\alpha = 5\%$*, có thể cho rằng số tử vong cao theo các độ tuổi trên lần lượt tuân theo tỷ lệ* $4 : 6 : 20 : 30 : 40$ *không?*

Ví dụ

Lời giải: Cặp giả thuyết H_0, H_1 cần kiểm định:
H_0 : Số tử vong cao theo các độ tuổi tuân theo tỷ lệ 4: 6: 20: 30: 40.
H_1 : Số tử vong cao theo các độ tuổi không tuân theo TL 4: 6: 20: 30: 40.

- Gọi $p_1, \ldots, p_6$ lần lượt là tỷ lệ tử vong cao theo các độ tuổi 18-39, $\ldots \geqslant 75$.
- Giả sử H_0 đúng, khi đó ta có

$$p_1 = 0.04, p_2 = 0.06, p_3 = 0.2, p_4 = 0.3, p_5 = 0.4.$$

- Số tử vong cao kỳ vọng theo từng độ tuổi sẽ là:

$$E_1 = 400 \times 0.04 = 16, E_2 = 400 \times 0.06 = 24,$$

$$E_3 = 400 \times 0.2 = 80, E_4 = 400 \times 0.3 = 120, E_5 = 400 \times 0.4 = 160.$$

Ví dụ

Lời giải:

Khoảng tuổi	18-39 years	40-49	50-64	65-74	⩾75 years	Tổng
Số khoảng	13	18	86	115	168	400
Xác suất theo H_0	0.04	0.06	0.2	0.3	0.4	1
Số kỳ vọng theo H_0	16	24	80	120	160	400

- Giá trị thống kê được cho bởi:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(13-16)^2}{16} + \frac{(9-12)^2}{12} + \ldots + \frac{(168-160)^2}{160} = 3.12.$$

- Bậc tự do của kiểm định khi- bình phương là k-1 = 5-1 = 4. Với $\alpha = 5\%$ ta có $\chi^2_{4,0.05} = 9.49$. Do $3.12 < 9.49$ nên không bác bỏ H_0. Như vậy ta có cơ sở thống kê để cho rằng số tử vong cao theo các độ tuổi trên lần lượt tuân theo tỷ lệ $4 : 6 : 20 : 30 : 40$.

Thực hiện kiểm chứng mức phù hợp trong R bằng hàm chisq.test [6]

```
x <- c(13,  18, 86, 115, 168)
p0 <- c(0.04, 0.06, 0.2,  0.3, 0.4)
chisq.test(x,  p=p0)
```

```
        Chi-squared test for given probabilities

data:  x
X-squared = 3.1208, df = 4, p-value = 0.5378
```

[6] `chisq.test(x, p = p0)`, trong đó x là véc tơ chỉ các quan sát, p_0 là véc tơ xác suất chỉ qui luật phân phối của tổng thể.

Một số lưu ý khi dùng kiểm định Khi-bình phương

Khi sử dụng kiểm định Khi-bình phương để kiểm định xem dữ liệu có được chọn từ một tổng thể tuân theo những phân phối phổ biến như phân phối chuẩn, phân phối Poisson, phân phối mũ,. . . , ta cần ước lượng các tham số chưa biết để xác định hoàn toàn phân phối cần kiểm định. Khi đó số bậc tự do của phân phối Khi-bình phương sử dụng trong kiểm định phụ thuộc vào m (số tham số cần ước lượng từ mẫu).

- Phân phối chuẩn ta cần ước lượng hai tham số là μ và σ^2 nên m=2, số bậc tự do k-3
- Phân phối Poisson ta cần ước lượng một tham số là λ nên m=1, số bậc tự do k-2
- Phân phối mũ ta cần ước lượng một tham số là μ nên m=1, số bậc tự do k-2,
- v.v..

Nội dung trình bày

1. Kiểm định tính độc lập

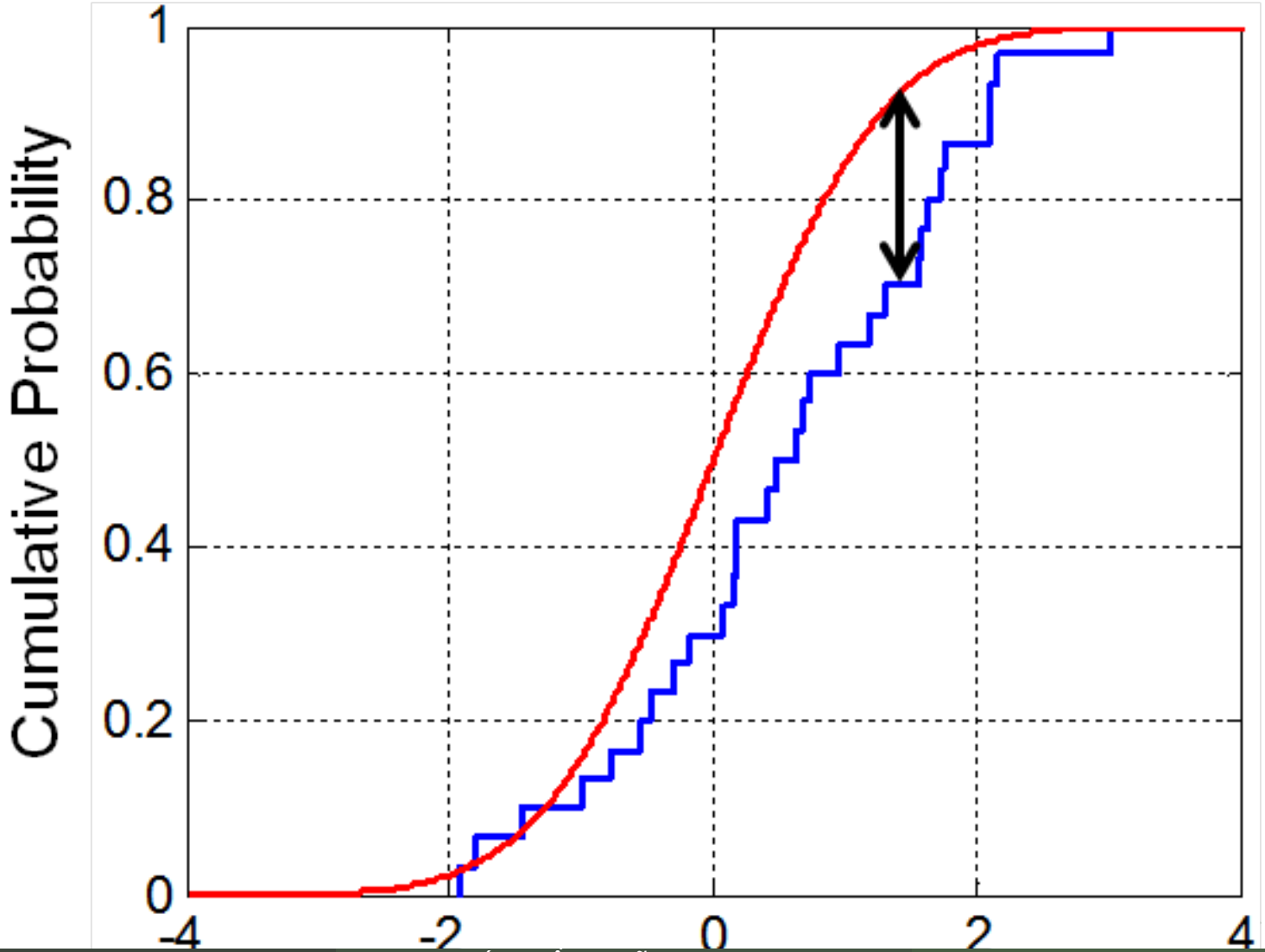
2. Kiểm chứng mức phù hợp của một phân phối
   - Kiểm định Khi- bình phương
   - Kiểm định Kolmogorov-Smirnov
   - Kiểm định phân phối chuẩn

Kiểm định Kolmogorov-Smirnov (K–S test or KS test)

- Kiểm định Kolmogorov-Smirnov kiểm tra xem một mẫu có phải được chọn từ một tổng thể có phân phối lý thuyết cho trước (one-sample K–S test) hay hai mẫu có phải được chọn ra từ cùng một tổng thể (có phân phối lý thuyết chưa biết) hay không.
- Kiểm định Kolmogorov-Smirnov được đặt tên theo hai nhà toán học Andrey Kolmogorov và Nikolai Smirnov.
- Thống kê Kolmogorov–Smirnov xem xét khoảng cách giữa hàm phân phối thực nghiệm của mẫu và hàm phân phối của phân phối tham chiếu hoặc giữa các hàm phân phối thực nghiệm của hai mẫu.

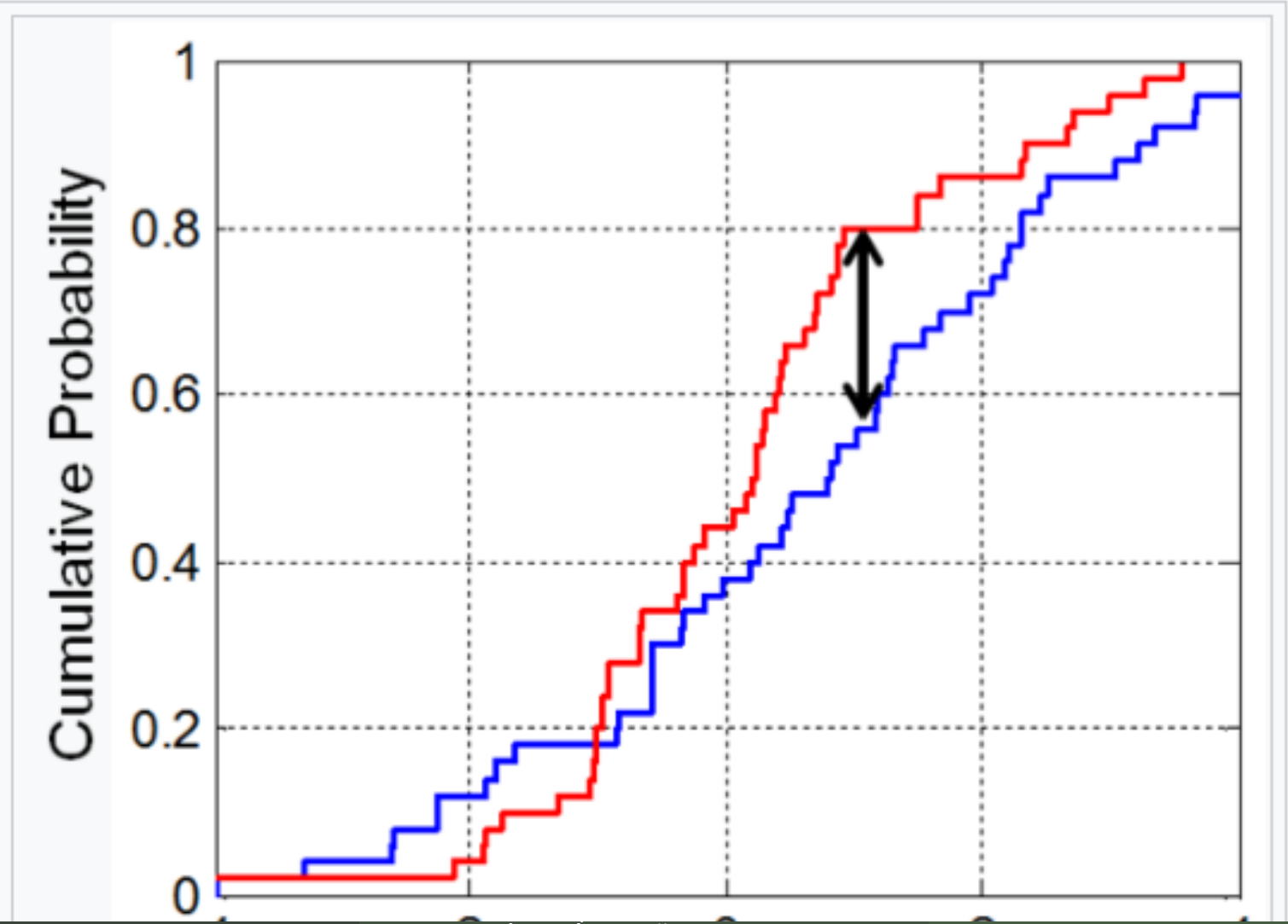
Minh họa thống kê Kolmogorov-Smirnov

The red line is a model CDF, the blue line is an empirical CDF, and the black arrow is the KS statistic.



Minh họa thống kê Kolmogorov-Smirnov

Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.



Kiểm định Kolmogorov-Smirnov (One sample K–S test)

Bài toán

Cho một phân phối lý thuyết (liên tục) có hàm phân phối xác suất là $F(x)$, ta cần kiểm định xem một mẫu có được chọn ra từ một tổng thể tuân theo phân phối lý thuyết đã cho hay không bằng cách kiểm định cặp giả thuyết sau:

H_0 : *Mẫu được chọn ra từ tổng thể tuân theo phân phối xác suất lý thuyết đã cho;*
H_1 : *Mẫu không được chọn từ tổng thể tuân theo phân phối xác suất lý thuyết đã cho.*

Kiểm định Kolmogorov-Smirnov

Xét mẫu ngẫu nhiên cỡ n, gồm các quan sát được sắp thứ tự $X_1, \ldots, X_n$.

- Hàm phân phối thực nghiệm F_n^* cho $\{X_i\}$ được cho bởi công thức

$$F_n^*(x) = \frac{\text{Số phần tử trong mẫu} \leqslant x}{n}$$
$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty, x]}(X_i), \forall x \in (-\infty, +\infty),$$

trong đó $\mathbb{I}_{(-\infty,x]}(X_i)$ là hàm chỉ báo, bằng 1 nếu $X_i \leqslant x$, bằng 0 trong các trường hợp còn lại.

Kiểm định Kolmogorov-Smirnov

- Thống kê Kolmogorov–Smirnov cho một hàm phân phối cho trước $F(x)$ được xác định bởi

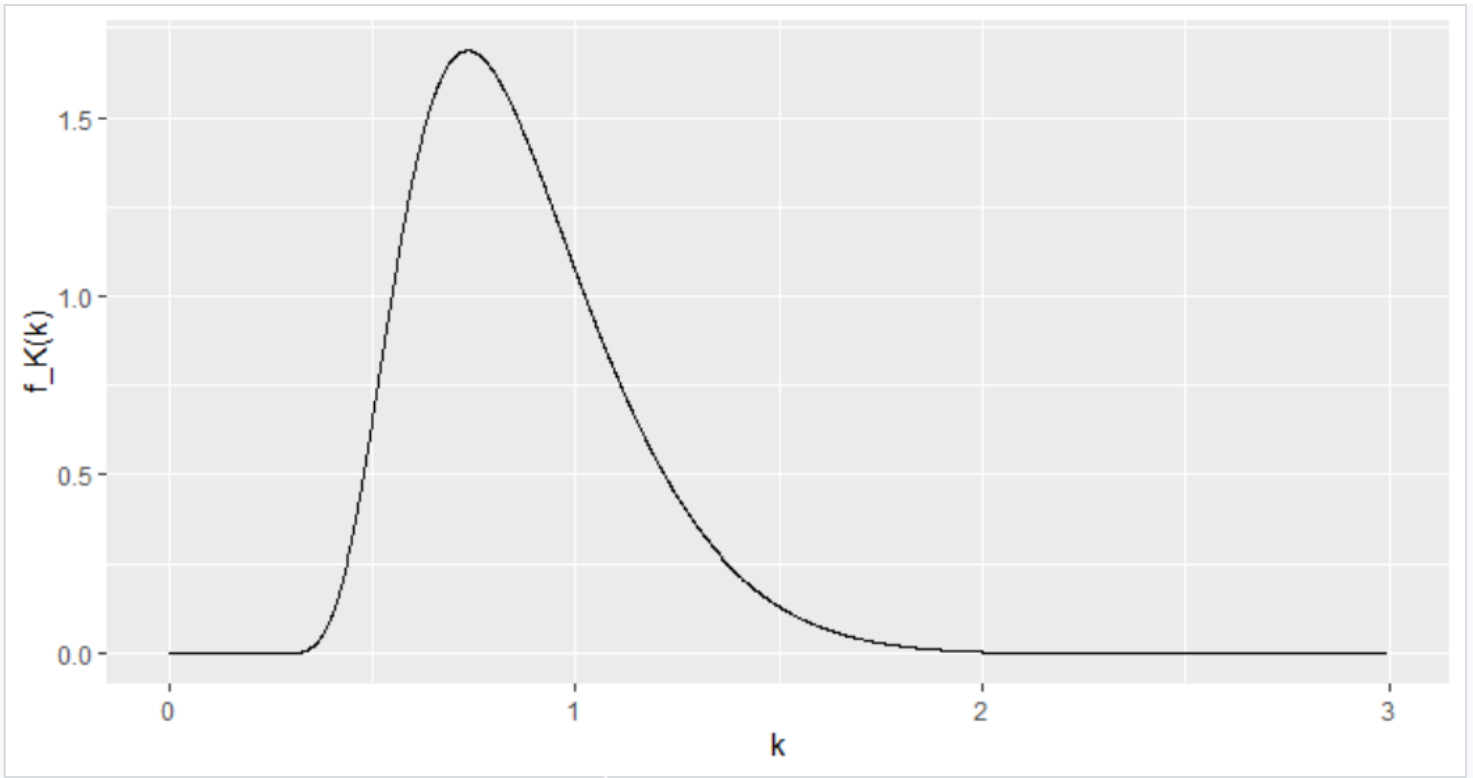
$$D_n = \sup_x |F_n^*(x) - F(x)|,$$

trong đó $\sup_x$ lấy qua tập các khoảng cách.
Một cách trực giác, thống kê D là khoảng cách lớn nhất của 2 hàm phân phối.

Kiểm định Kolmogorov-Smirnov

- Nếu $F(x)$ là liên tục thì $\sqrt{n}D$ hội tụ đến phân phối Kolmogorov mà không phụ thuộc vào dạng của $F(x)$.
  Phân phối Kolmogorov là phân phối xác suất có hàm phân phối cho bởi

$$P(K \leqslant x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$



Kiểm định Kolmogorov-Smirnov

- Quy tắc bác bỏ H_0 tại mức ý nghĩa α là

$$\sqrt{n}D > K_\alpha,$$

trong đó K_α xác định bởi $P(K > K_\alpha) = \alpha$, với K là biến ngẫu nhiên tuân theo phân phối Kolmogorov.

Kiểm định Kolmogorov-Smirnov trong R

- Kiểm định Kolmogorov-Smirnov trong R ta dùng hàm `ks.test(x , ...)`, trong đó x là véc tơ dữ liệu, phần dấu ba chấm gõ tên hàm phân phối liên tục nào đó, nêu rõ tham số.
  Ví dụ `ks.test(x, "pnorm",12,1)` , kiểm định xem x có tuân theo phân phối chuẩn với $\mu = 12$ và $\sigma = 1$ hay không.
- Hàm `ks.test` còn được xây dựng để kiểm định xem hai tập số liệu có phải được lấy ra từ hai tổng thể cùng phân phối hay không `ks.test(x, y)`.

Nội dung trình bày

1. Kiểm định tính độc lập

2. Kiểm chứng mức phù hợp của một phân phối
   - Kiểm định Khi- bình phương
   - Kiểm định Kolmogorov-Smirnov
   - Kiểm định phân phối chuẩn

Kiểm định phân phối chuẩn

? Có những cách nào có thể kiểm tra được một tập dữ liệu mẫu có rút ra từ một tổng thể có phân phối chuẩn

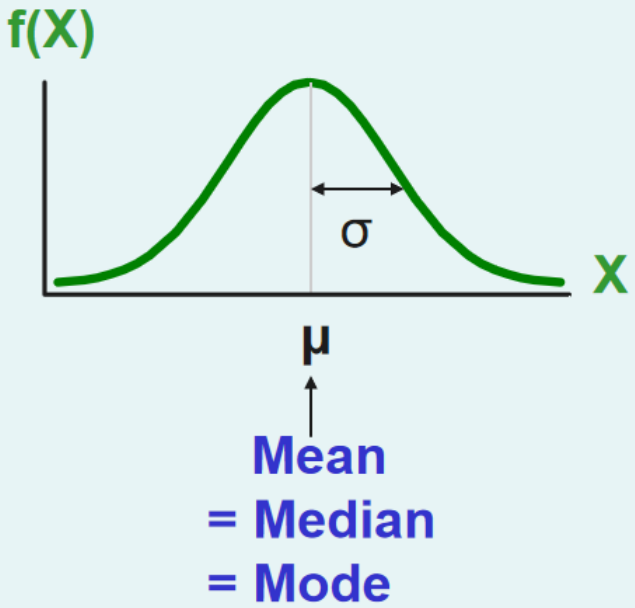
Kiểm định phân phối chuẩn

- Dùng biểu đồ, dùng đại lượng thống kê mô tả;
- Dùng một trong các kiểm định sau.
  - Kiểm định Khi-bình phương;
  - Kiểm định Kolmogorov-Smirnov;
  - Kiểm định Jarque–Bera test;
  - Kiểm định Shapiro-Wilk.

Đánh giá phân phối chuẩn: Số đo mô tả

? Trung bình, trung vị và mode bằng nhau

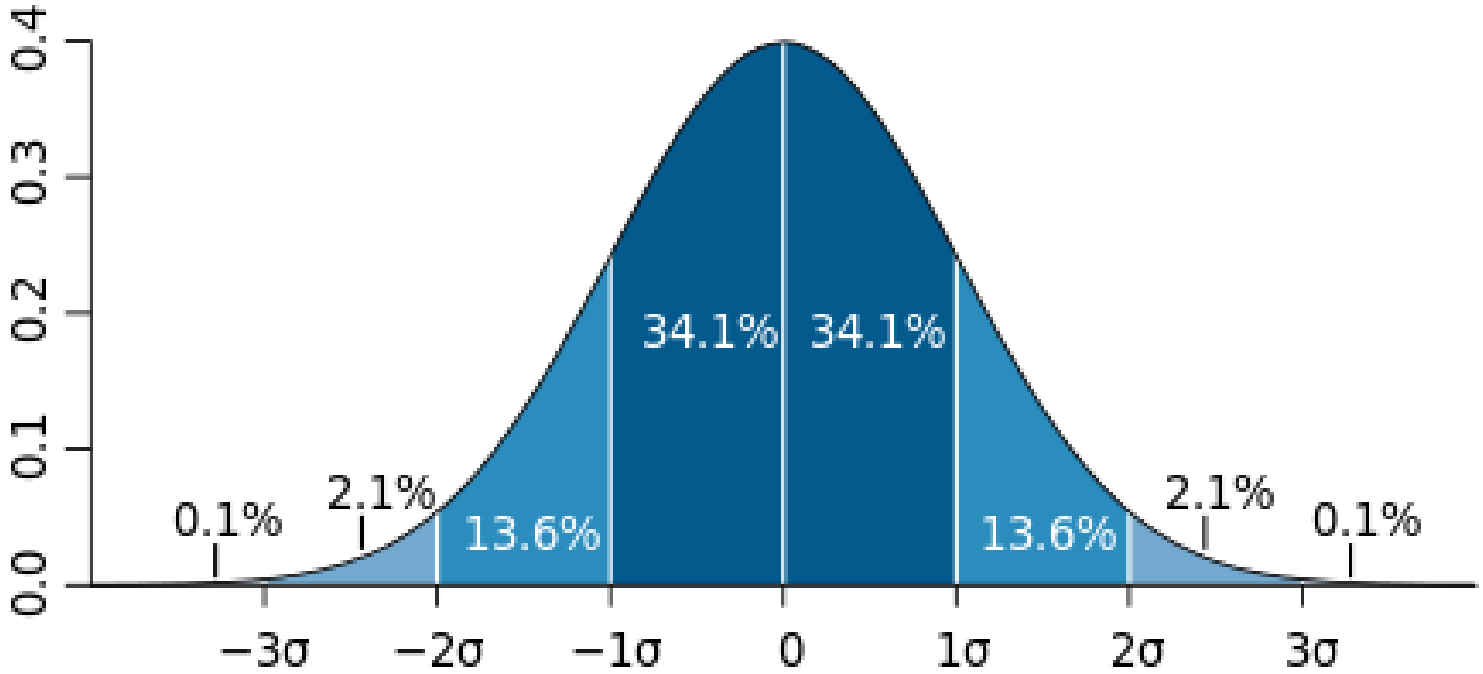
- Đối xứng
- Hình chuông
- Trung bình = Trung vị = Mode



Đánh giá phân phối chuẩn: Số đo mô tả

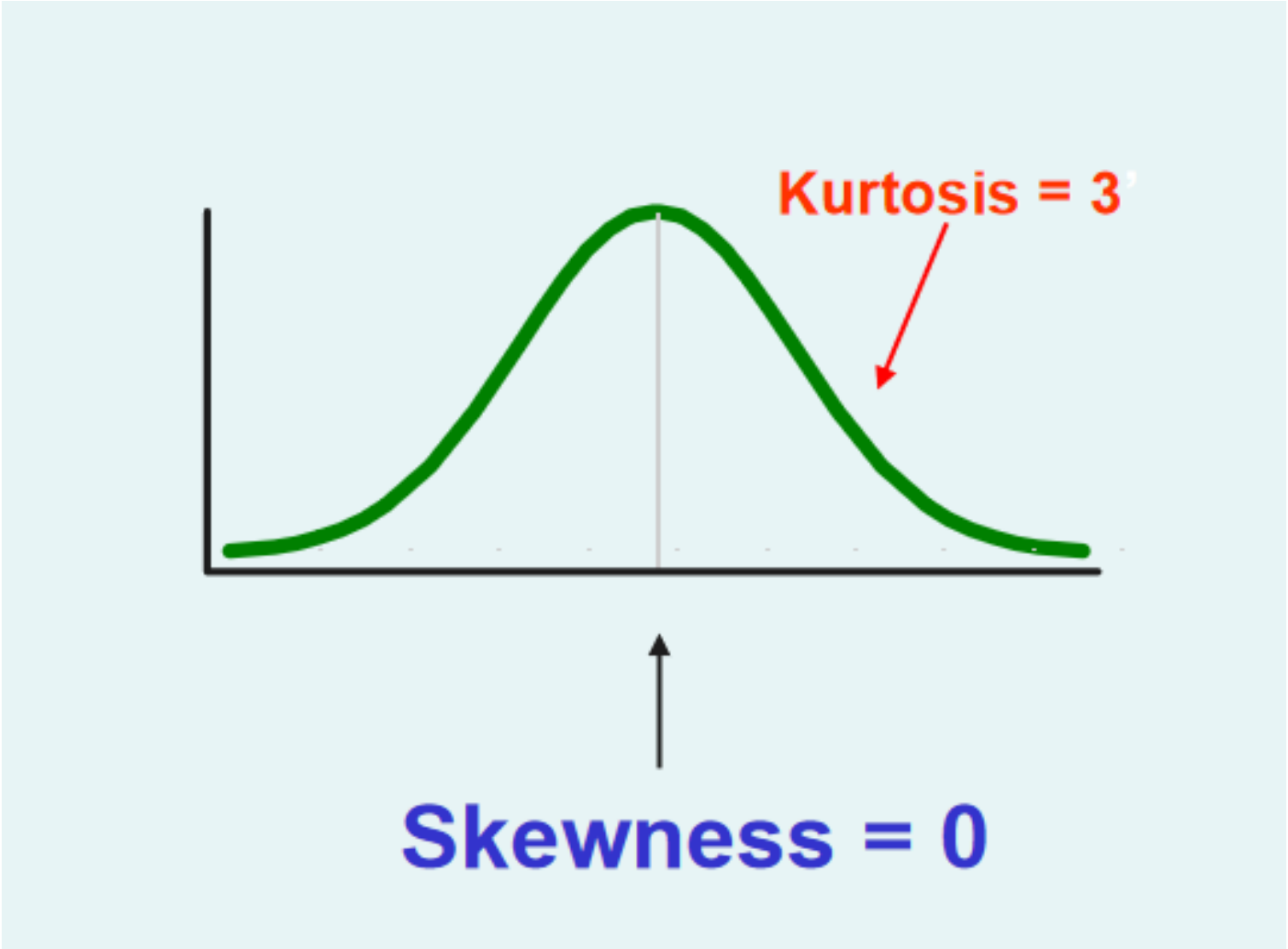
? So sánh trung bình và độ lệch chuẩn

- 2/3 số quan sát trong $[\mu - \sigma, \mu + \sigma]$;
- 95% số quan sát trong $[\mu - 2\sigma, \mu + 2\sigma]$;
- Độ trải giữa bằng 1.33σ;
- Khoảng biến thiên bằng 6σ



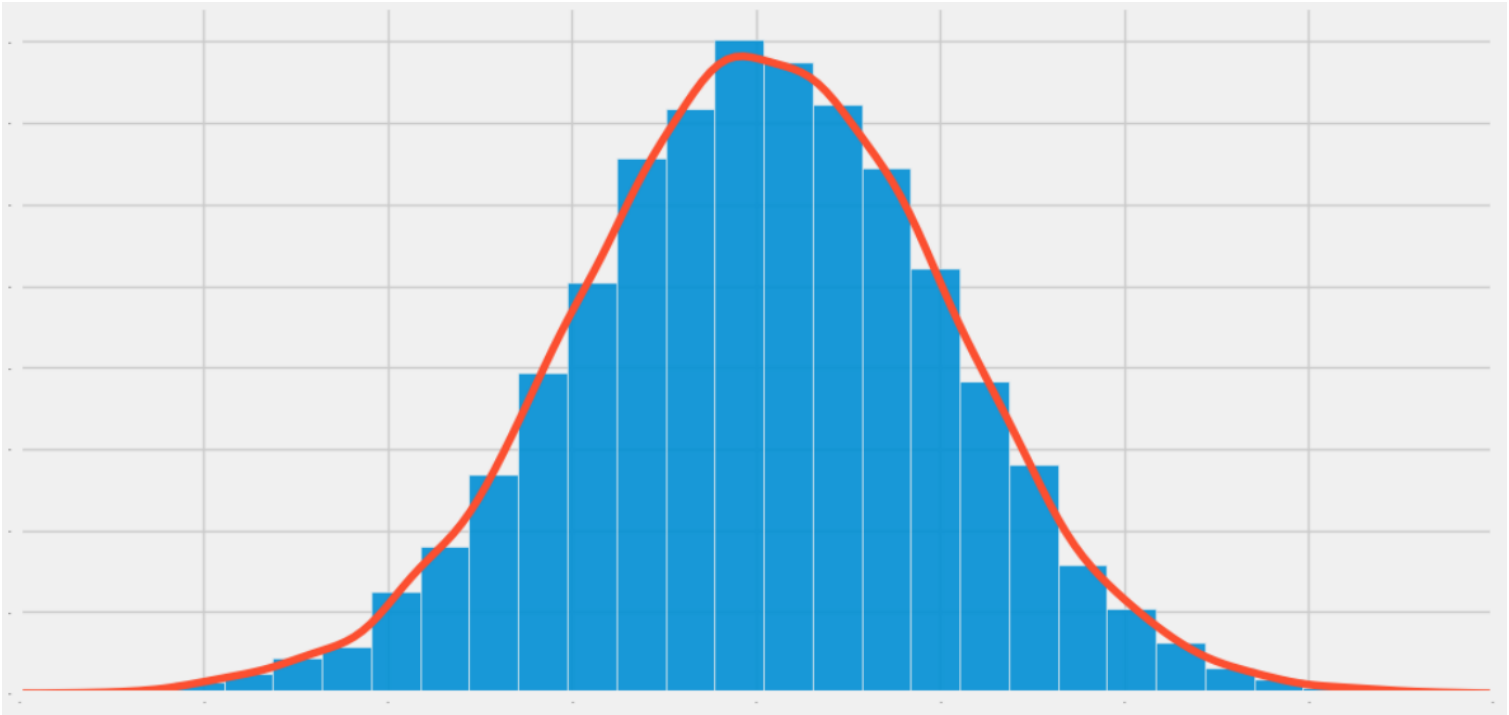
Đánh giá phân phối chuẩn: Số đo mô tả

? Skewness $= 0$ và Kurtosis $= 3$



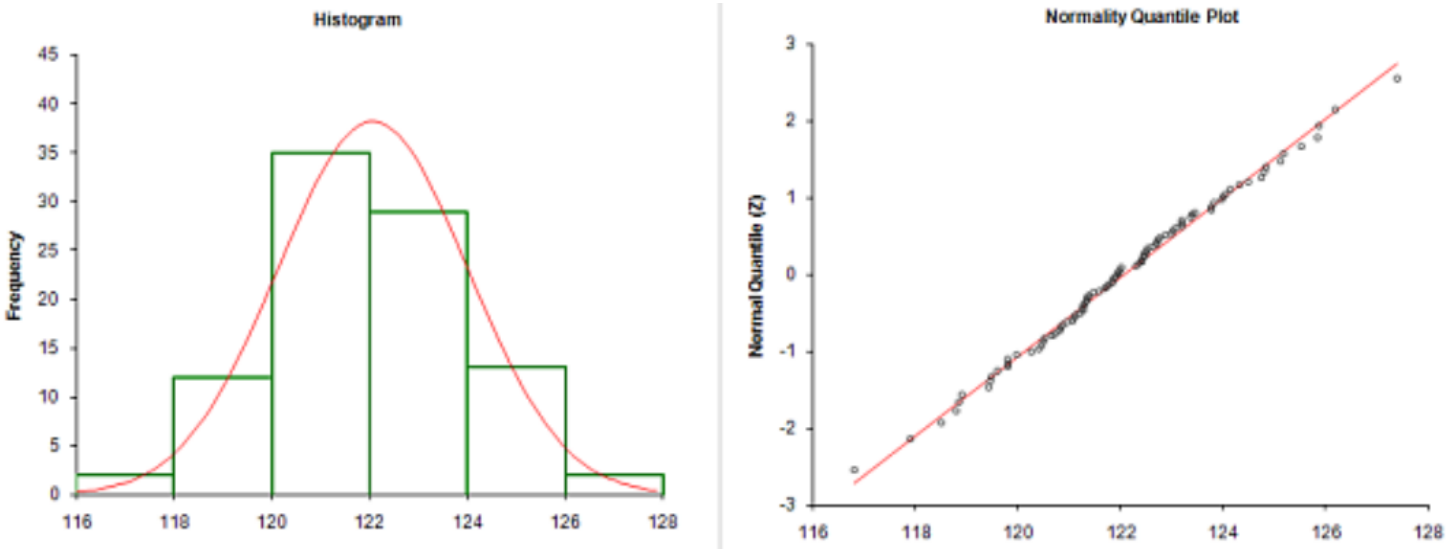
Đánh giá phân phối chuẩn: Biểu đồ

Biểu đồ phân phối tần số, đa giác tần số, biểu đồ hộp, biểu đồ thân và lá:
Kiểm tra tính đối xứng, hình chuông



Đánh giá phân phối chuẩn: Biểu đồ

Kiểm tra biểu đồ Q-Q (quantile–quantile plot/Q-Q plot): Có dạng đường thẳng dốc lên

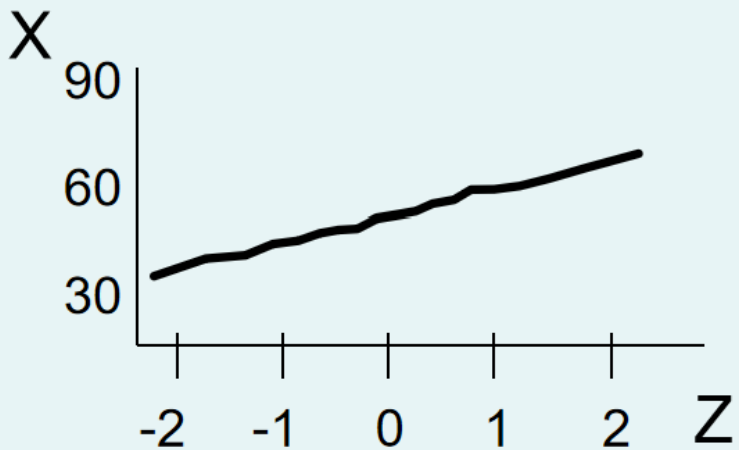


Biểu đồ Q-Q (quantile–quantile plot)

- Biểu đồ Q–Q là một biểu đồ xác suất dùng để so sánh hai phân phối xác suất bằng cách vẽ các phân vị của chúng với nhau.
- Nếu hai phân phối như nhau thì các điểm trên biểu đồ Q-Q sẽ xấp xỉ đường thẳng $y = x$.

Biểu đồ xác suất phân phối chuẩn (normal probability plot) là một trường hợp đặc biệt của biểu đồ Q-Q cho PP chuẩn.

Nếu dữ liệu được chọn từ phân phối chuẩn thì biểu đồ xác suất PP chuẩn sẽ xấp xỉ tuyến tính.



Cách vẽ biểu đồ xác suất PP chuẩn

- Sắp xếp dữ liệu theo thứ tự tăng dần;
- Tìm các giá trị phân vị tương ứng của phân phối chuẩn hóa (Z)[7];
- Vẽ các cặp điểm với giá trị quan sát được (X) trên trục thẳng đứng và giá trị phân vị của phân phối chuẩn hóa (Z) trên trục ngang;
- Đánh giá tính tuyến tính của tập điểm.

---

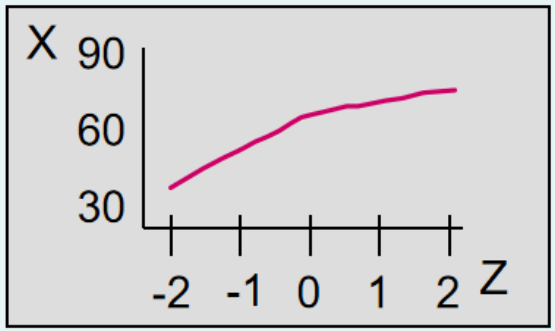
[7] The formula used by the "qqnorm" function in the basic "stats" package in R (programming language) is as follows:

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right),$$

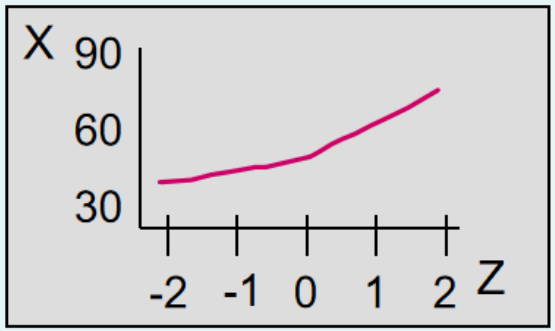
for $i = 1, 2, \ldots, n$, where $a = 3/8$ if $n \leqslant 10$ and $a = 0.5$ for $n > 10$, and Φ^{-1} is the standard normal quantile function.

Biểu đồ Q-Q

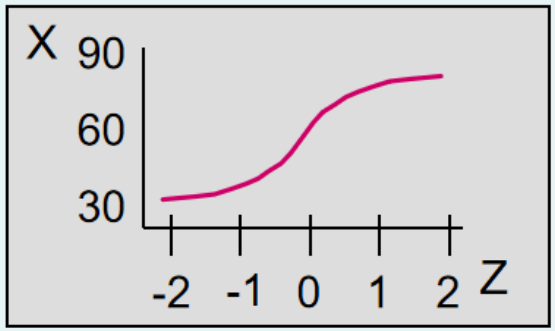
Left-Skewed



Right-Skewed



Rectangular

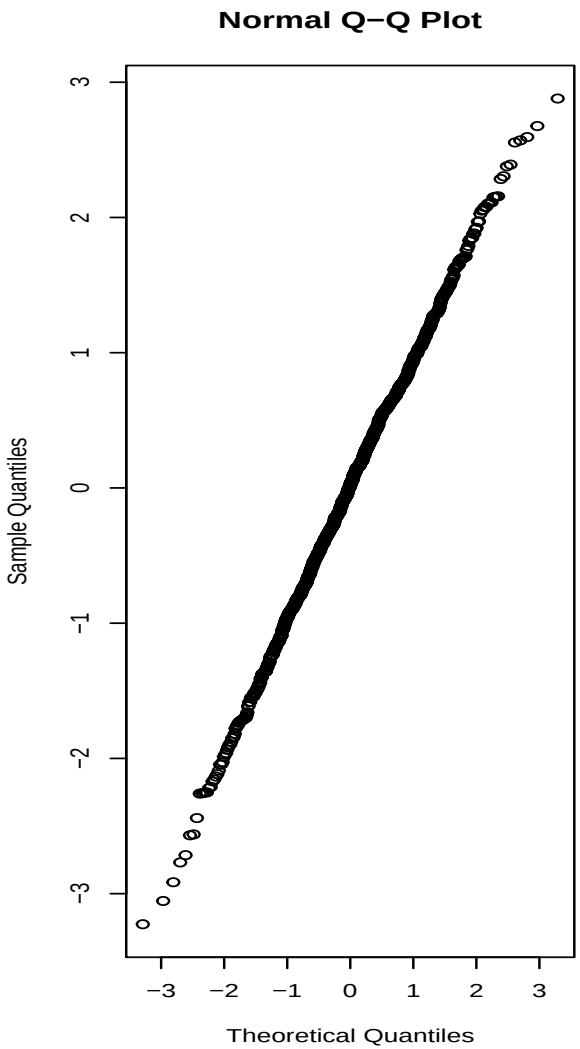


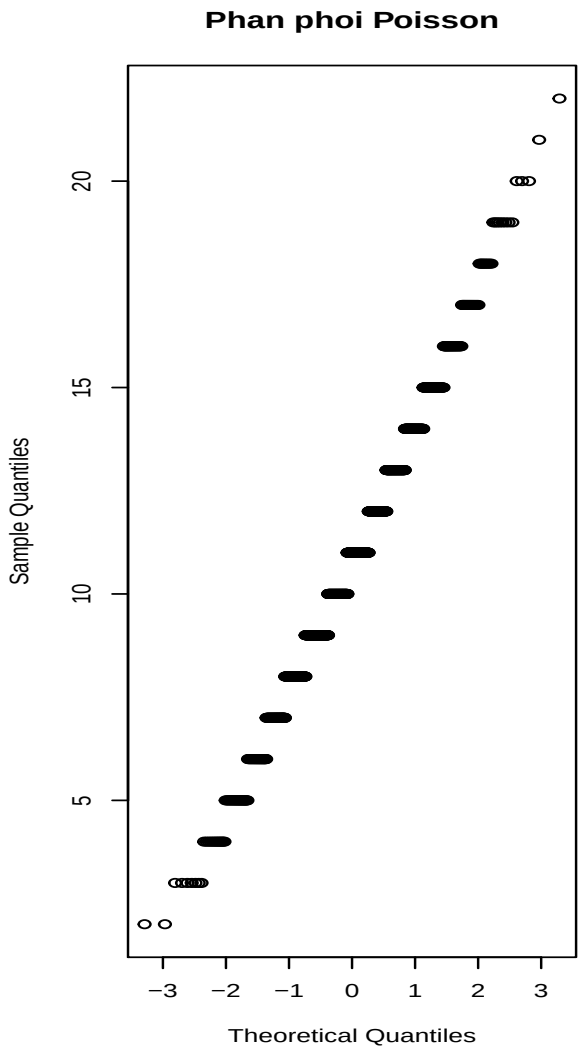
Nonlinear plots indicate a deviation from normality

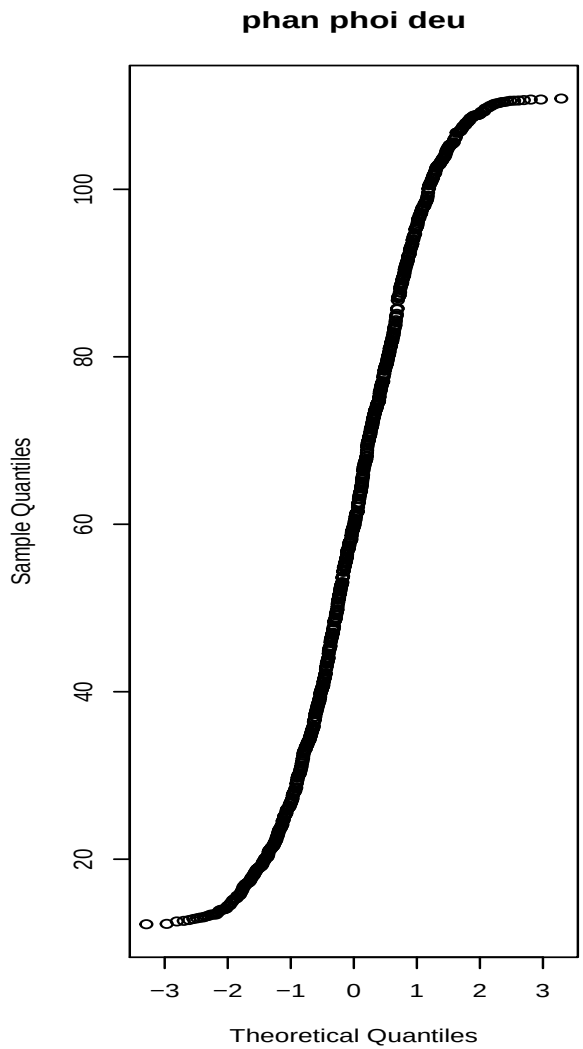
Biểu đồ Q-Q

```
x <- rnorm(1000, 0, 1)
qqnorm(x)
x <- rpois(1000, 11)
qqnorm(x, main = 'Phan phoi Poisson')
x <- runif(1000, 12, 111)
qqnorm(x, main  =  'phan phoi deu')
```

Biểu đồ Q-Q plot







Kiểm chứng PP chuẩn dùng kiểm định Khi-bình phương

- Xếp các phần tử của mẫu vào các khoảng phù hợp, giả sử khoảng thứ i có O_i phần tử của mẫu;
- Ước tính các tham số μ và σ theo mẫu (nếu chưa biết);
- Tính các giá trị xác suất p_i theo H_0 và tần số E_i theo H_0;
- Tính giá trị thống kê:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

- Tại mức ý nghĩa α ta quyết định bác bỏ H_0 nếu $\chi^2 > \chi^2_{k-2-1,\alpha}$ (ước lượng μ, σ) hoặc $\chi^2 > \chi^2_{k-1,\alpha}$ (không cần ước lượng μ, σ).

(8)

[8] Tiêu chuẩn Khi-bình phương chỉ áp dụng khi tần số tương ứng mỗi giá trị hay khoảng giá trị $n_i \geqslant 5$. Khi tần số thực nghiệm nhỏ, ta phải ghép giá trị hoặc khoảng giá trị để tần số tăng thêm.

Kiểm chứng PP chuẩn dùng kiểm định Kolmogorov

Khi sử dụng tiêu chuẩn Kolmogorov để kiểm định phân phối chuẩn, ta phải giả thiết kỳ vọng toán và phương sai của phân phối chuẩn lý thuyết phải được biết trước mà không ước tính qua mẫu. Nếu điều này không thỏa mãn thì việc kiểm định sẽ kém chính xác.

```
IQ <- rnorm(1000, 100, 15)
ks.test(IQ, 'pnorm',  100,  15)
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  IQ
D = 0.016578, p-value = 0.9463
alternative hypothesis: two-sided
```

(9)

[9] Khi các tham số của phân phối chuẩn chưa được biết mà phải ước lượng qua mẫu, kiểm định Kolmogorov được thay bằng kiểm định Liliefors (Thống kê được tính như kiểm định KS nhưng giá trị tới hạn Kolmogorov K_α được thay bằng giá trị tới hạn Liliefors L_α)

Kiểm định Jarque–Bera (Jarque–Bera test)

- Kiểm định Jarque–Bera dựa trên hai đặc trưng của phân phối chuẩn thông qua hai hệ số độ nghiêng (Skewness) và hệ số đo độ nhọn (Kurtosis).
- Kiểm định Jarque–Bera được đặt tên theo hai nhà toán học Carlos Jarque và Anil K. Bera.
- Kiểm định Jarque–Bera thường được sử dùng trong phân tích hồi quy để kiểm định dạng phân phối chuẩn của phần dư.

Kiểm định Jarque–Bera (Jarque–Bera test)

- Đặc trưng thứ nhất của phân phối chuẩn là tính đối xứng qua trung bình. Đặc trưng này thể hiện qua hệ số đo độ nghiêng bằng 0:

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{s^3},$$

trong đó $\overline{x}, s$ là trung bình và độ lệch chuẩn của mẫu.

- Đặc trưng thứ hai của phân phối chuẩn là hệ thức giữa độ phẳng của các phần đuôi của phân phối so với phần trung tâm thể hiện qua hệ số Kurtosis bằng 3:

$$Kurtosis = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{s^4},$$

trong đó $\overline{x}, s$ là trung bình và độ lệch chuẩn của mẫu.

Kiểm định Jarque–Bera

- Giá trị thống kê của phép kiểm định là

$$JB = n\left(\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24}\right).$$

- Nếu H_0 đúng và cỡ mẫu lớn thì JB sẽ có phân phối khi-bình phương với 2 bậc tự do. [10]
- Trong trường hợp cỡ mẫu rất lớn, ta sẽ bác bỏ H_0 nếu $JB > \chi^2_{2,\alpha}$. [11]

[10] Nếu cỡ mẫu nhỏ thì ta sẽ so sánh JB với giá trị tương ứng trong bảng giá trị Bowman-Shelton.

[11] Nếu cỡ mẫu nhỏ ta sẽ bác bỏ H_0 nếu B lớn hơn giá trị tương ứng trong bảng giá trị Bowman-Shelton.

Kiểm định Jarque–Bera trong R bằng hàm `jarque.bera.test()` [12]. Hàm trong gói `tseries`

```
IQ <- rnorm(1000, 100, 15)
library(tseries)
jarque.bera.test(IQ)

        Jarque Bera Test

data:  IQ
X-squared = 0.79445, df = 2, p-value = 0.6722
```

[12] `jarque.bera.test(x)`, trong đó x là véc tơ dữ liệu

Kiểm định Shapiro-Wilk

- Kiểm định Shapiro-Wilk kiểm tra một tập dữ liệu có được chọn từ một phân phối chuẩn được đưa ra bởi Samuel Sanford Shapiro và Martin Wilk năm 1965. [13]
- Kiểm định Shapiro-Wilk trong R ta dùng hàm `shapiro.test(x)`, trong đó x là véc tơ dữ liệu mẫu với cỡ từ 3 đến 5000.

```
IQ <- rnorm(1000, 100, 15)
shapiro.test(IQ)
```

```
        Shapiro-Wilk normality test

data:  IQ
W = 0.99871, p-value = 0.693
```

[13] Monte Carlo simulation has found that Shapiro–Wilk has the best power for a given significance, followed closely by Anderson–Darling when comparing the Shapiro–Wilk, Kolmogorov–Smirnov, and Lilliefors. However, the Shapiro–Wilk test is known not to work well in samples with many identical values.

CẢM ƠN CÁC BẠN ĐÃ LẮNG NGHE!