Một Vài Chủ Đề Thực Tiễn Xử Lý Dữ Liệu

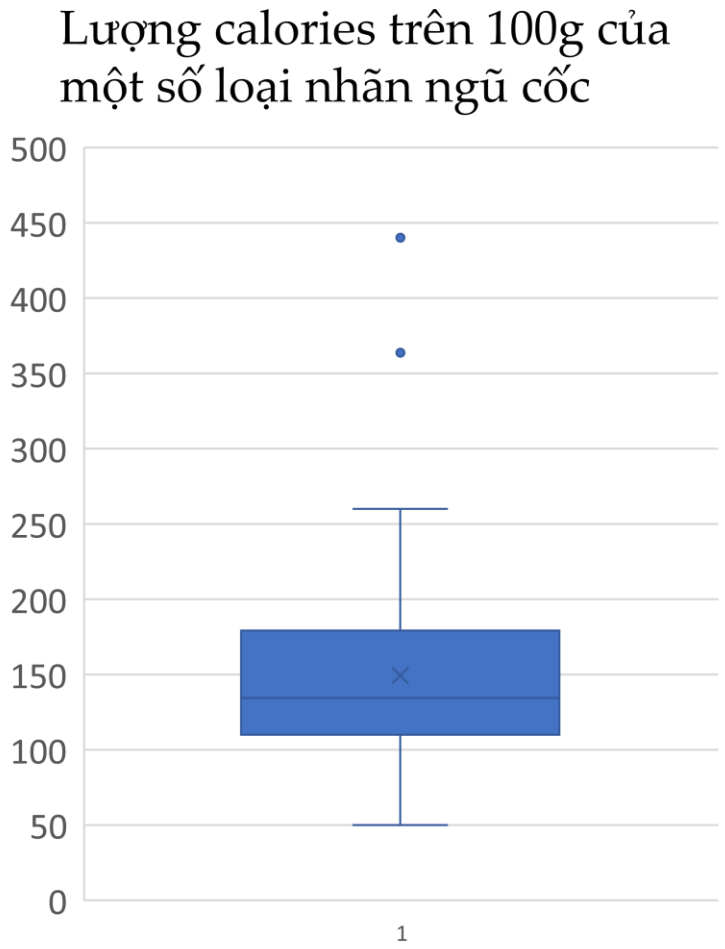
Khái Quát

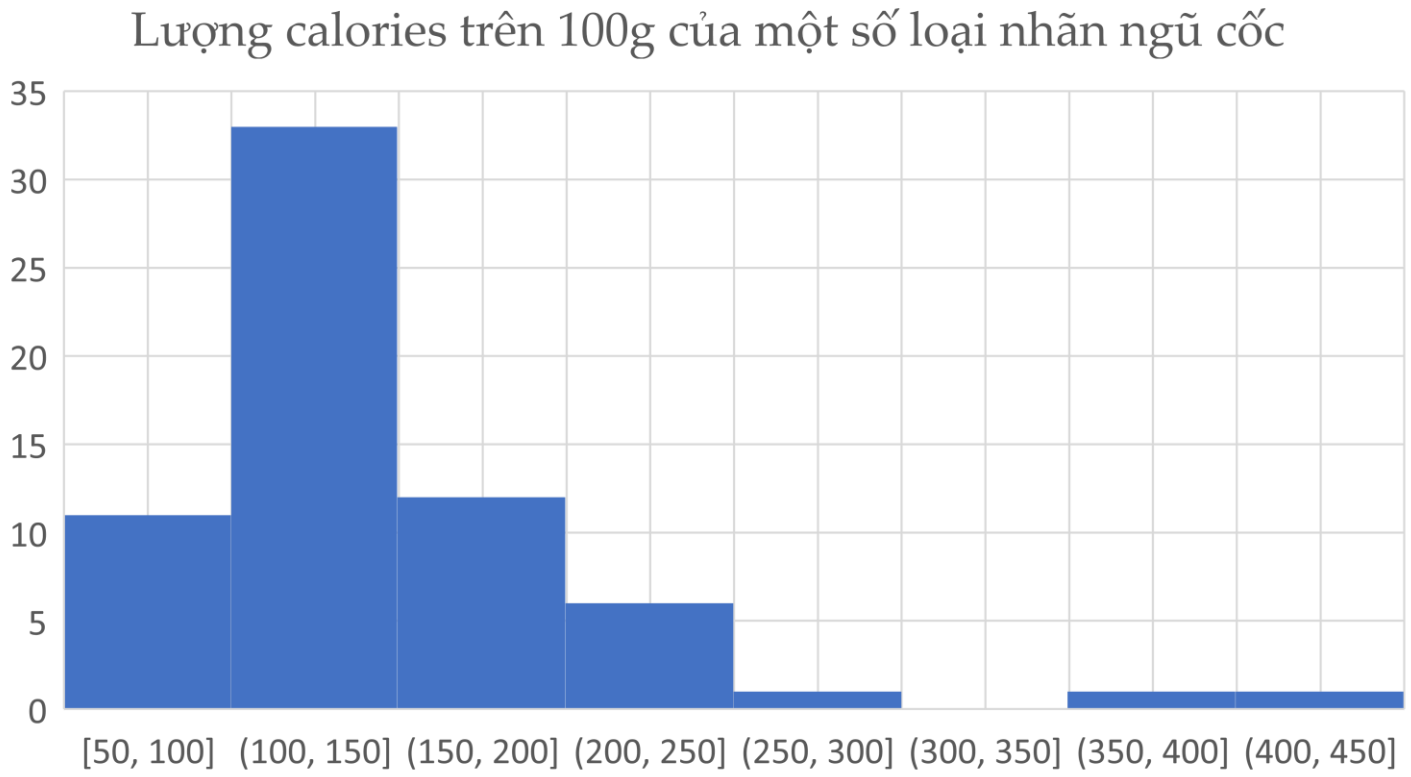
- Dữ liệu bất thường (outliers).
- Missing values.
- Kiểm định phân phối chuẩn.

Điểm Bất Thường (Outliers)

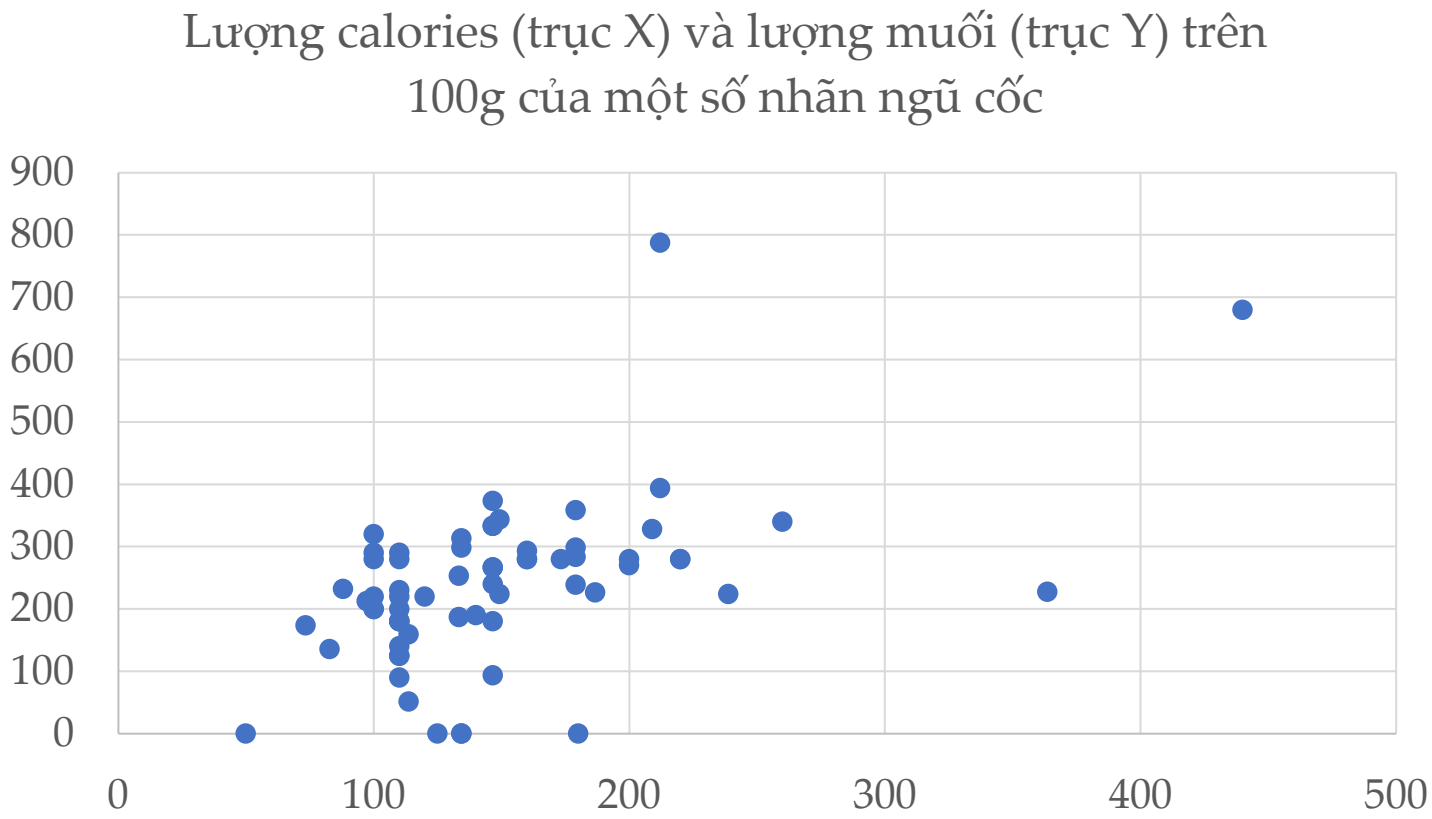
- Điểm bất thường/ ngoại lai/ dị biệt (outliers) được định nghĩa là điểm khác biệt đáng kể so với các điểm còn lại.
- Điểm bất thường thường được gây ra bởi nguyên nhân sai sót nhập liệu hoặc xử lý số liệu, hoặc do một cơ chế mới xuất hiện tạo ra dữ liệu.
- Cách phát hiện điểm khác thường: sử dụng đồ thị hoặc tiêu chuẩn thống kê.

Ví Dụ 1:


Lượng calories trên 100g của một số loại nhãn ngũ cốc
500
450
400
350
300
250
200
150
100
50
0
1

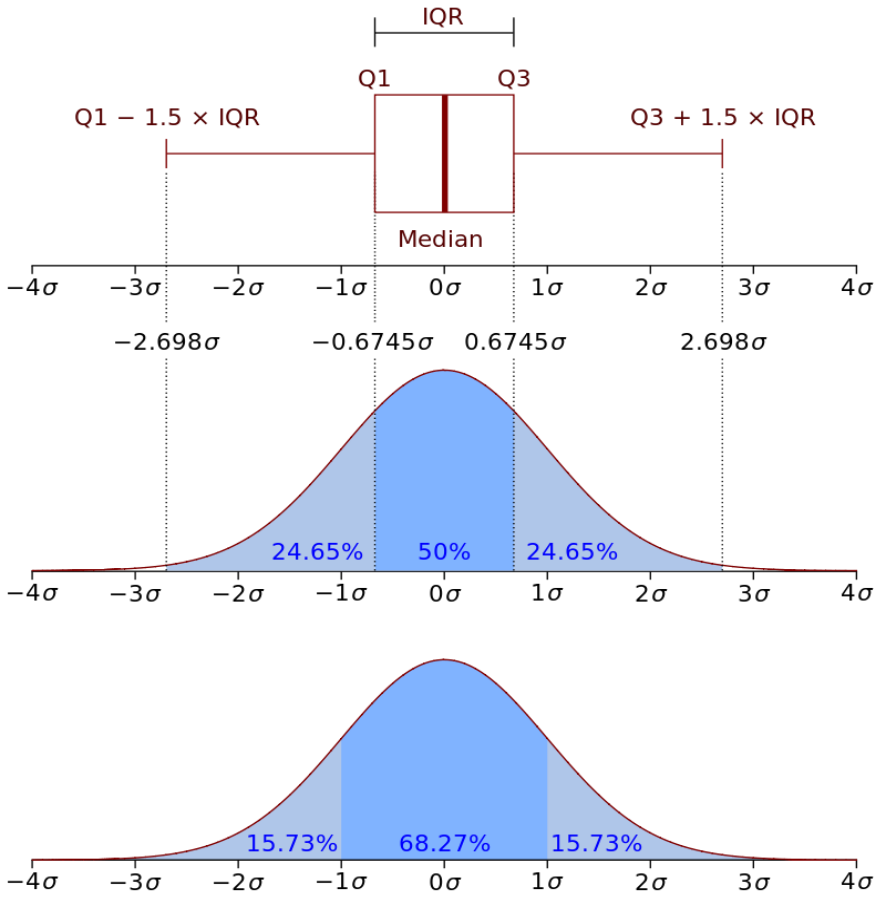

Lượng calories trên 100g của một số loại nhãn ngũ cốc
35
30
25
20
15
10
5
0
[50, 100]
(100, 150]
(150, 200]
(200, 250]
(250, 300]
(300, 350]
(350, 400]
(400, 450]

Ví Dụ 2:


Lượng calories (trục X) và lượng muối (trục Y) trên 100g của một số nhãn ngũ cốc
900
800
700
600
500
400
300
200
100
0
0
100
200
300
400
500

Sử Dụng Tiêu Chuẩn IQR

- Nếu dữ liệu của chúng ta có phân phối chuẩn
- Nếu điểm dữ liệu nằm ngoài 1.5 lần IQR của khu vực 50%: tương đương với tỷ lệ xuất hiện 7/1,000.
- Nếu điểm dữ liệu nằm ngoài 3 lần IQR của khu vực 50%: tương đương với tỷ lệ xuất hiện 1/500,000.
- Có thể sử dụng tiêu chuẩn 1.5 hay 3 IQR.



Dữ Liệu Trống/ Khuyết

- Trong quá trình thu thập dữ liệu, chúng ta có thể bắt gặp một hay nhiều ô dữ liệu không có giá trị. Đó là hiện tượng **dữ liệu trống (missing values).**
- Dữ liệu trống có thể xảy ra:
  - Ngẫu nhiên (hoàn toàn hoặc không hoàn toàn): quá trình khuyết dữ liệu xảy ra ngẫu nhiên, không phụ thuộc vào giá trị hay tính chất của biến bị khuyết dữ liệu.
  - Không ngẫu nhiên: quá trình thiếu của dữ liệu xảy ra phụ thuộc vào giá trị của dữ liệu. Ví dụ: thu thập dữ liệu VHLSS các hộ rất nghèo hoặc thu nhập rất cao thường không có kết quả phản hồi.

Xử Lý Dữ Liệu Trống

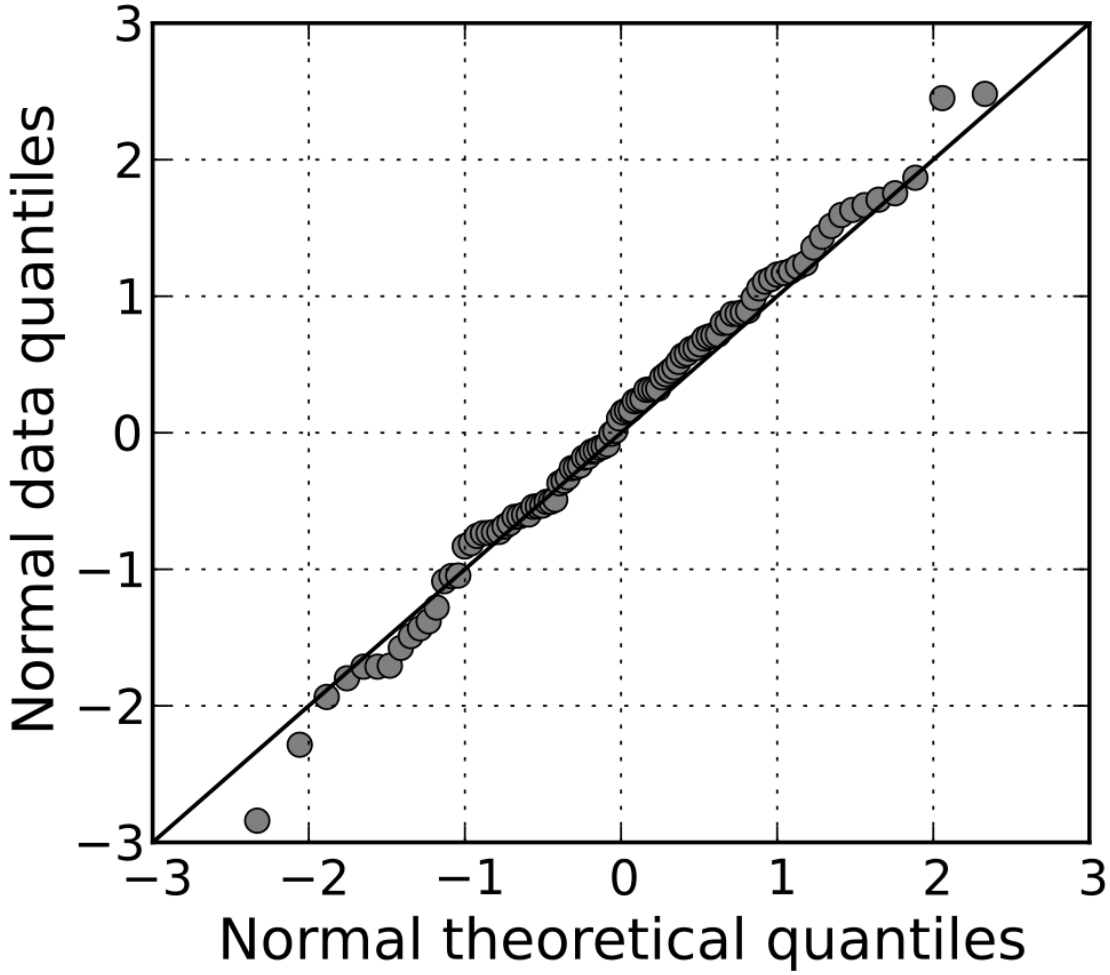
- Loại bỏ dữ liệu: xóa những dữ liệu khỏi mẫu đang có.
- Thay thấy bằng một dữ liệu khác. Một số biện pháp chủ yếu: dùng giá trị trung bình của toàn mẫu, tìm quan sát gần như tương tự hoặc sử dụng phương pháp hồi quy.
- Xử lý dữ liệu trống là phần quan trọng của quá trình xử lý dữ liệu (data cleaning).

Các Kiểm Định Phân Phối Chuẩn (1)

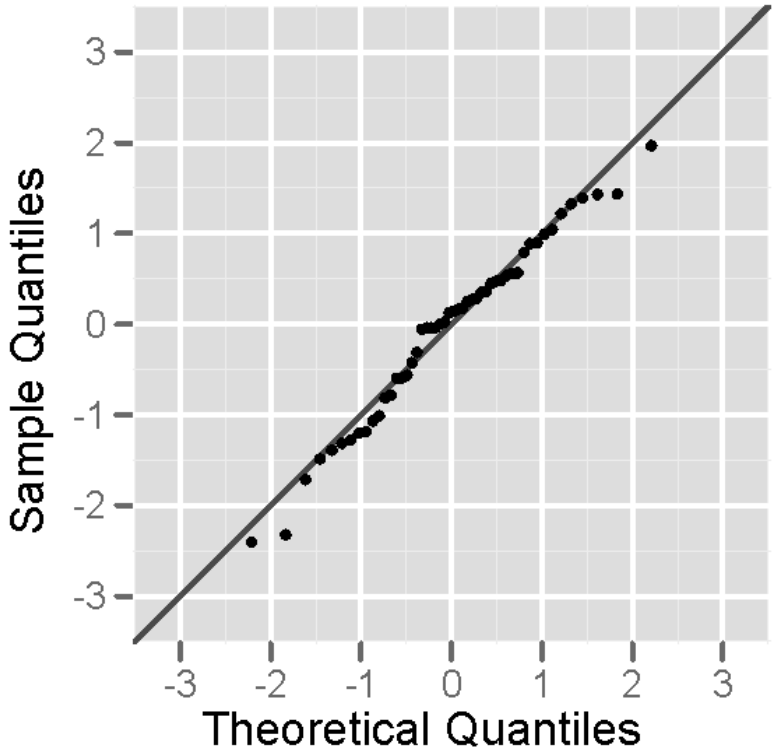
- Các kiểm định phân phối (normality test) chuẩn thường gặp:
  - Sử dụng histogram hoặc QQ plot.
  - Jarque-Bera test.
  - D'Agostino's K^2 test.
  - Shapio-Wilk test.
- Kiểm định Jarque-Bera và D'Agostino's kiểm tra độ nhọn và độ xiên có xấp xỉ phân phối chuẩn không. Kiểm định Shapio-Wilk so sánh hàm c.d.f. của phân phối chuẩn với xấp xỉ c.d.f. của dữ liệu thực tế.

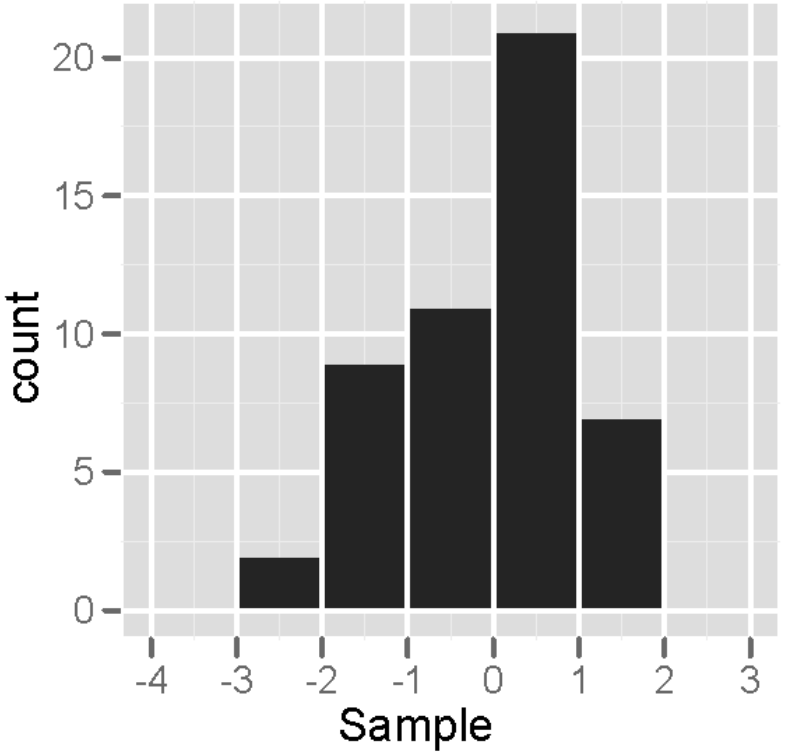
QQ plot
(Quantile-Quantile Plot)

- **Biểu đồ QQ** thể hiện sự so sánh quantile giữa hai loại dữ liệu: dữ liệu theo phân phối chuẩn lý thuyết và dữ liệu quan sát chúng ta có.
- Quantile là chia nhỏ phân phối thành từng khoảng có xác suất đều nhau. Giống như quartile là chia 4 khoảng, quantile có số khoảng tùy chọn.
- Mỗi chấm là giá trị của quantile của hai phân phối.
- Các chấm gần theo sát đường 45 nghĩa là phân phối của dữ liệu gần sát với phân phối chuẩn.

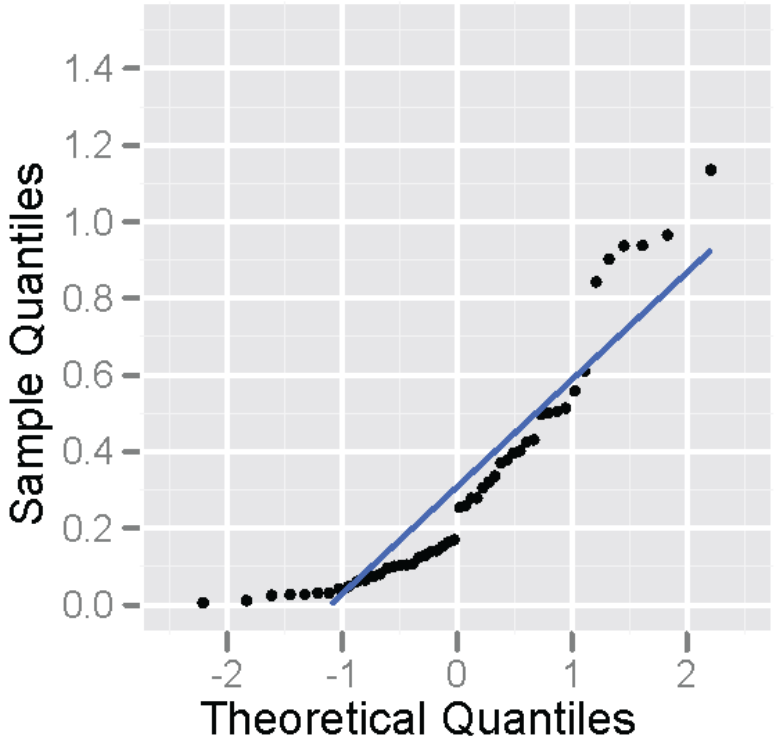


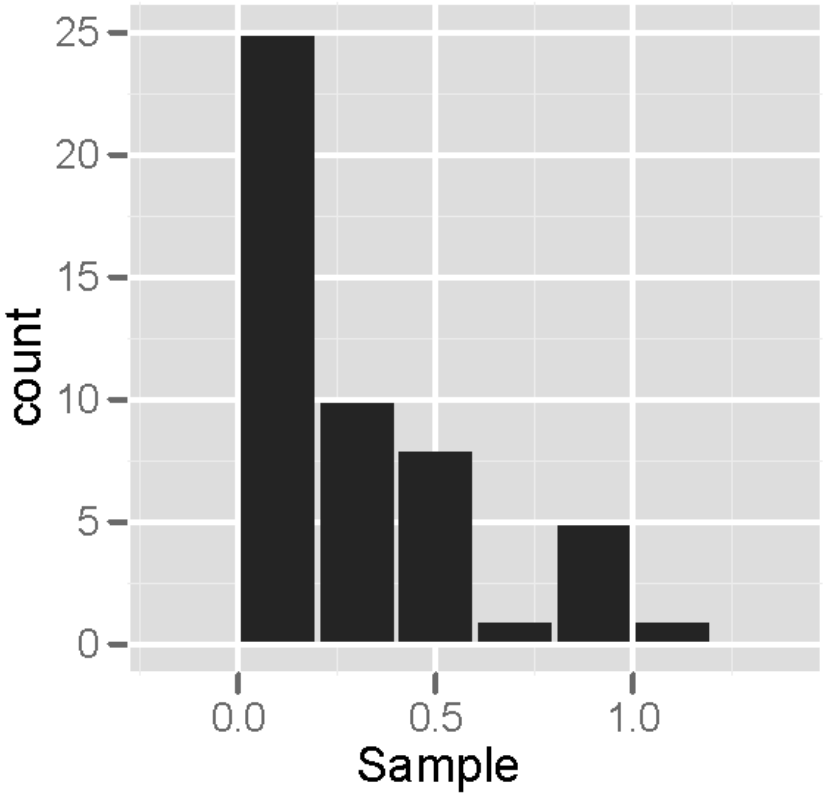
Ví dụ 1: mẫu 50


Sample Quantiles
Theoretical Quantiles
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


count
Sample
20
15
10
5
0
-4
-3
-2
-1
0
1
2
3

Ví dụ 2: mẫu 50


Sample Quantiles
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Theoretical Quantiles


count
25
20
15
10
5
0
0.0
0.5
1.0
Sample

Các Kiểm Định Phân Phối Chuẩn (2)

- Các lưu ý khi sử dụng các kiểm định.
  - Xác định H_0 là gì? H_0 là phân phối chuẩn hay không phải phân phối chuẩn.
  - Phần mềm trả về giá trị p-value. Với giá trị này, chúng ta từ chối hay hay bác bỏ H_0.
  - So sánh kết quả kiểm định với biểu đồ các bạn vẽ.
- Ví dụ: kiểm định D'Agostino's K^2 có

$$H_0: \text{normality is true}$$

$$H_1: H_0 \text{ is not true}$$

Giá trị p-value = 0.63. Kết luận của bạn về tính chuẩn của phân phối dữ liệu?