STA 2002, Summer 2024
Probability and Statistics II

Order Statistics and QQ plot

Why study order statistics

- **Issue 1**: Is there a method for us to check whether our data is statistically similar to a normal distribution?
  - (Yes, we can use QQ plots.)
- **Issue 2**: If the data is not statistically similar to normal distribution, is there a way to do hypothesis testing?
  - (Yes, **non-parametric / distribution-free** confidence interval and hypothesis testing.)

Order Statistics serves as an essential tool

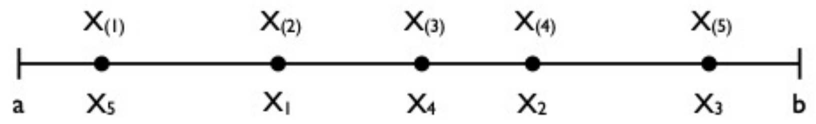
Reading: Chapter 6.3

Order Statistics

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from a common distribution f.

$$X_1, X_2, \ldots, X_n \sim f$$

Definition



Denote the ***ordered*** sample values $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ as the order statistic

For each k, the k-th order statistic is

$$X_{(k)} = k\text{-th smallest of } X_1, X_2, \ldots, X_n$$

Example

$$\begin{aligned} X_{(1)} &= \min\{X_1, X_2, \ldots, X_n\}, \\ X_{(2)} &= \textit{second smallest of } X_1, \ldots, X_n, \\ X_{(n)} &= \max\{X_1, X_2, \ldots, X_n\}. \end{aligned}$$

Density of Order Statistics

Let $X_1, X_2, \dots, X_n$ be i.i.d. random samples from a common distribution with CDF $F(x)$ and PDF $f(x)$. What is the CDF and PDF of any order statistic $X_{(k)}$?

A special case: $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}.$

$$P\left(X_{(n)} \le x\right) = P(\max\{X_1, \dots, X_n\} \le x) = P(X_1 \le x, X_2 \le x, \dots, X_n \le x)$$

$$= P(X_1 \le x) \times P(X_2 \le x) \times \cdots \times P(X_n \le x)$$

$$= F(x)^n$$

CDF: $F_{X_{(n)}}(x) = F(x)^n$

PDF: $f_{X_{(n)}}(x) = nF(x)^{n-1} f(x)$

Density of Order Statistics

Let $X_1, X_2, \dots, X_n$ be i.i.d. random samples from a common distribution with CDF $F(x)$ and PDF $f(x)$. What is the CDF and PDF of any order statistic $X_{(k)}$?

A special case: $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$,

$$P\left(X_{(1)} > x\right) = P(\min\{X_1, \dots, X_n\} > x) = P(X_1 > x, X_2 > x, \dots, X_n > x)$$

$$= P(X_1 > x) \times P(X_2 > x) \times \cdots \times P(X_n > x)$$

$$= (1 - F(x))^n$$

CDF: $F_{X_{(1)}}(x) = 1 - (1 - F(x))^n$

PDF: $f_{X_{(1)}}(x) = n(1 - F(x))^{n-1} f(x)$

Density of Order Statistics

Let $X_1, X_2, \dots, X_n$ be i.i.d. random samples from a common distribution with CDF $F(x)$ and PDF $f(x)$. What is the CDF and PDF of any order statistic $X_{(k)}$?

What is the CDF and PDF of $X_{(k)}$ for any k?

We define $W \sim Bin(n, F(x))$ as the number of X_i that is smaller than x.

$$
\begin{aligned}
F_{X_{(k)}}(x) &= P(X_{(k)} \leq x) \\
&= P(W \geq k) \\
&= \sum_{l=k}^{n} P(W = l) \\
&= \sum_{l=k}^{n} \binom{n}{l} F(x)^l (1 - F(x))^{n-l} \\
&= \sum_{l=k}^{n-1} \binom{n}{l} F(x)^l (1 - F(x))^{n-l} + F(x)^n.
\end{aligned}
$$

Density of Order Statistics

We define $W \sim Bin(n, F(x))$ as the number of X_i that is smaller than x.

$$F_{X_{(k)}}(x) = \sum_{l=k}^{n} \binom{n}{l} F(x)^l (1-F(x))^{n-l} \;\; = \sum_{l=k}^{n-1} \binom{n}{l} F(x)^l (1-F(x))^{n-l} + F(x)^n.$$

$$\begin{aligned}
f_{X_{(k)}}(x) &= \sum_{l=k}^{n-1} \binom{n}{l} l[F(x)]^{l-1} f(x)[1-F(x)]^{n-l} + \sum_{l=k}^{n-1} \binom{n}{l} (n-l)[F(x)]^l(-f(x))[1-F(x)]^{n-l-1} \\
&\quad + nF(x)^{n-1}f(x) \\
&= \sum_{l=k}^{n} \binom{n}{l} l[F(x)]^{l-1} f(x)[1-F(x)]^{n-l} + \sum_{l=k}^{n-1} \binom{n}{l} (n-l)[F(x)]^l(-f(x))[1-F(x)]^{n-l-1} \\
&= \sum_{l=k}^{n} \binom{n}{l} l[F(x)]^{l-1} f(x)[1-F(x)]^{n-l} + \sum_{l'=k+1}^{n} \binom{n}{l'-1} (n-l'+1)[F(x)]^{l'-1}(-f(x))[1-F(x)]^{n-l'} \\
&= \sum_{l=k}^{n} \frac{n!}{(l-1)!(n-l)!}[F(x)]^{l-1} f(x)[1-F(x)]^{n-l} \;\; + \sum_{l'=k+1}^{n} \frac{n!}{(l'-1)!(n-l')!}[F(x)]^{l'-1}(-f(x))[1-F(x)]^{n-l'} \\
&= \frac{n!}{(k-1)!1!(n-k)!}[F(x)]^{k-1} f(x)[1-F(x)]^{n-k}.
\end{aligned}$$

Density of Order Statistics

Let $X_1, X_2, \dots, X_n$ be i.i.d. random samples from a common distribution with CDF $F(x)$ and PDF $f(x)$. What is the CDF and PDF of any order statistic $X_{(k)}$?

Theorem 1 *Suppose that $X_1, \dots, X_n$ are i.i.d. continuous random variables with common pdf $f(x)$ and cdf $F(x)$. For $k = 1, \dots, n$, denote the cdf and pdf of the kth order statistic $X_{(k)}$ to be respectively $F_{X_{(k)}}$ and $f_{X_{(k)}}$. They can be written as*

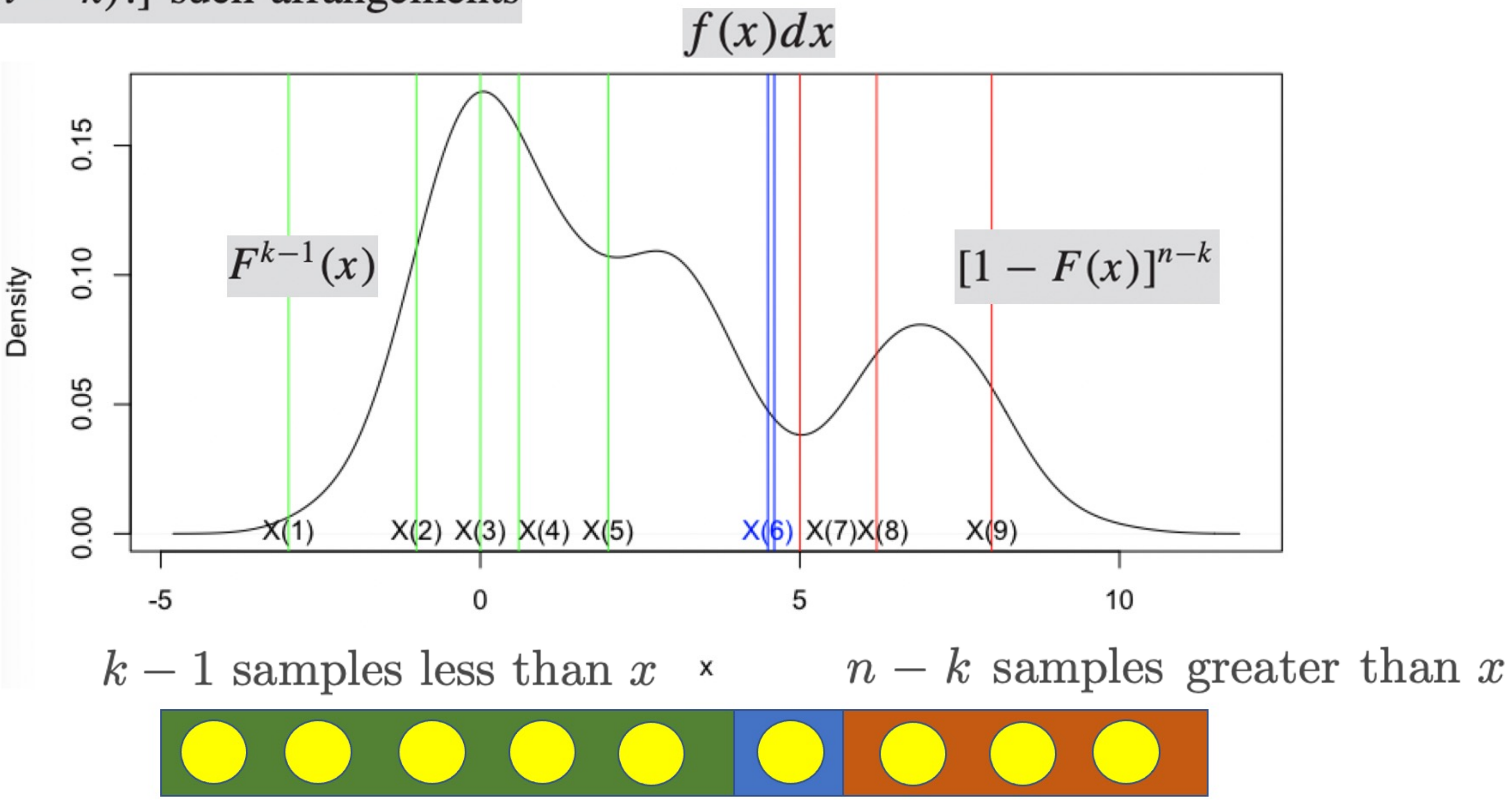
$$F_{X_{(k)}}(x) = \sum_{l=k}^{n} \binom{n}{l} F(x)^l (1 - F(x))^{n-l},$$

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!} [F(x)]^{k-1} f(x) [1 - F(x)]^{n-k}.$$

Density of Order Statistics

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!}[F(x)]^{k-1}f(x)[1-F(x)]^{n-k}.$$

$n!/[(k-1)!1!(n-k)!]$ such arrangements

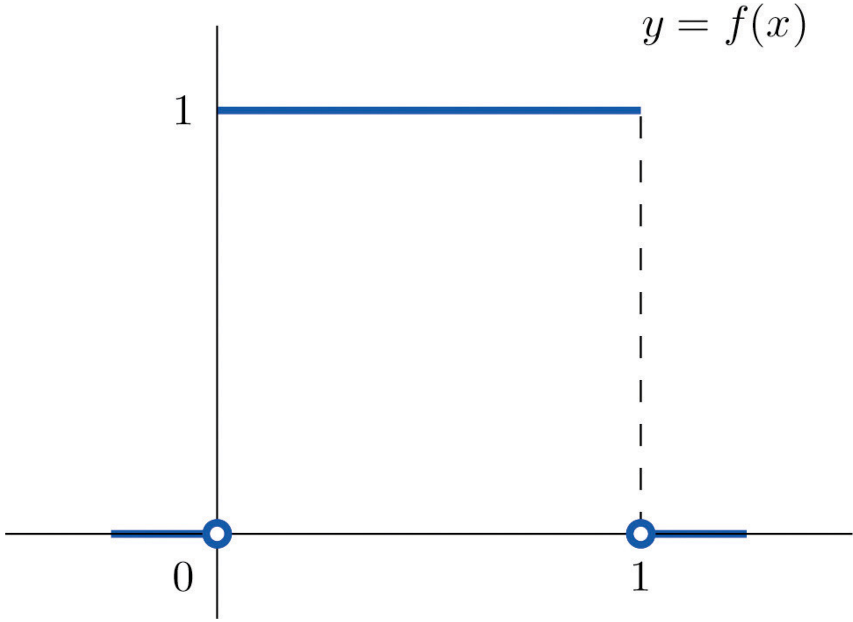


Example – Uniform distribution

Let $X_1, X_2, \dots, X_n$ be i.i.d. samples from the uniform distribution on [0,1].

$$X_1, X_2, \dots, X_n \sim \text{Uniform}[0,1]$$

$$F(x) = x, \quad f(x) = 1, \forall 0 \le x \le 1$$



$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!}[F(x)]^{k-1} f(x) [1 - F(x)]^{n-k}.$$

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1} (1-x)^{n-k}$$

Example – Uniform distribution

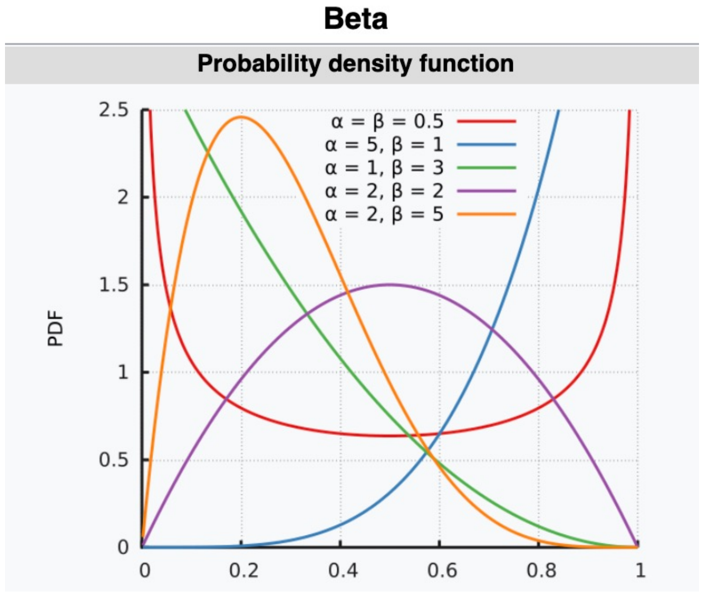
$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1}(1-x)^{n-k}$$

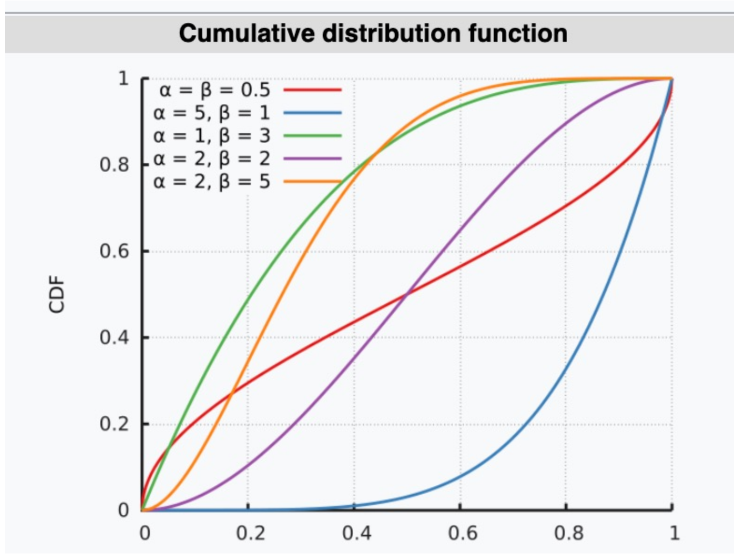
Beta distribution

$X \sim Beta(\alpha, \beta)$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$





Gamma function: $\Gamma(t) = \int_0^\infty y^{t-1} e^{-y}\, dy,$

$$\Gamma(n) = (n-1)!$$

$$\boxed{X_{(k)} \sim Beta(k, n-k+1)}$$

Credit: Wiki

Example – Uniform distribution

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1}(1-x)^{n-k} \qquad X_{(k)} \sim Beta(k, n-k+1)$$

The mean and variance of $X_{(k)}$

$$E\left(X_{(k)}\right) = \int_0^1 x \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1}(1-x)^{n-k} dx = \int_0^1 \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k+1-1}(1-x)^{n+1-(k+1)} dx$$

$$= \frac{n!}{(k-1)!\,1!\,(n-k)!} \times \frac{k!\,1!\,(n+1-(k+1))!}{(n+1)!} = \boxed{\frac{k}{n+1}}$$

$$E\left(X_{(k)}^2\right) = \int_0^1 x^2 \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1}(1-x)^{n-k} dx = \int_0^1 \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k+2-1}(1-x)^{(n+2)-(k+2)} dx$$

$$= \frac{n!}{(k-1)!\,1!\,(n-k)!} \times \frac{(k+1)!\,1!\,(n+2-(k+2))!}{(n+2)!} = \frac{k(k+1)}{(n+1)(n+2)}$$

$$\Rightarrow \quad Var\left(X_{(k)}^2\right) = \frac{k(k+1)}{(n+1)(n+2)} - \left(\frac{k}{n+1}\right)^2 = \boxed{\frac{k(n+1-k)}{(n+1)^2\,(n+2)}}$$

Example – Uniform distribution

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,1!\,(n-k)!} x^{k-1}(1-x)^{n-k}$$

$$X_{(k)} \sim Beta(k, n-k+1)$$

The mean and variance of $X_{(k)}$

Beta distribution

$$X \sim Beta(\alpha, \beta)$$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$E(X) = \frac{\alpha}{\alpha+\beta}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\alpha = k, \beta = n-k+1$$

$$E\left(X_{(k)}\right) = \frac{\alpha}{\alpha+\beta} = \frac{k}{n+1}$$

$$Var\left(X_{(k)}\right) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{k(n+1-k)}{(n+1)^2(n+2)}$$

Expectation of Order Statistics

- For uniform distribution

$$X_1, X_2, \dots, X_n \sim \mathrm{Uniform}[0,1]$$

- The expectation

$$\boxed{E\left(X_{(k)}\right) = \frac{\alpha}{\alpha + \beta} = \frac{k}{n+1}}$$

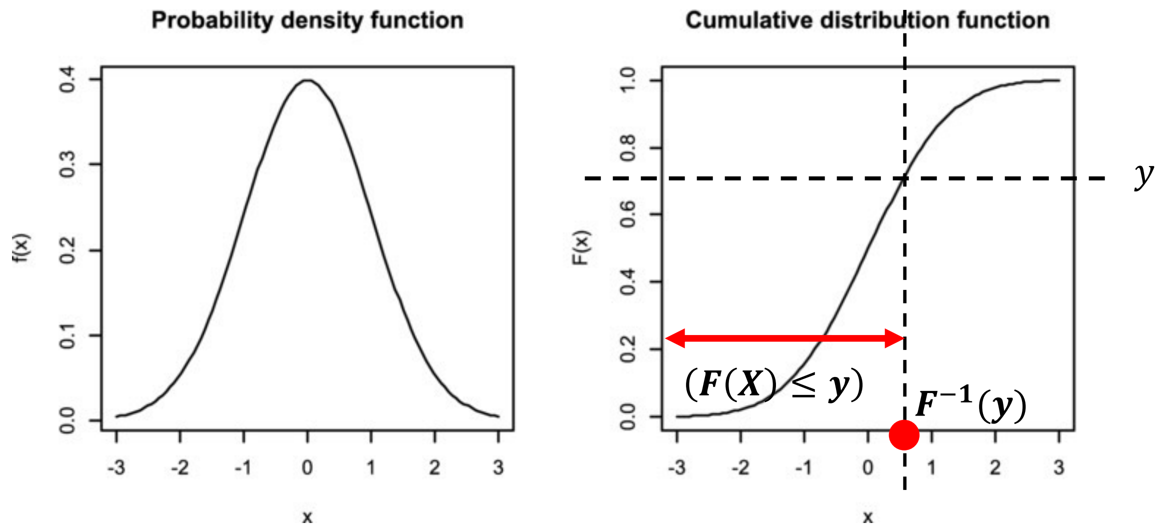
- How about other distributions (non-uniform)? Is this true for any other distributions? No.

Expectation of $F(X_{(k)})$

- Let X be a random variable with PDF $f(x)$ and *strictly increasing* CDF $F(x)$.
- Let $Y = F(X)$ be a new random variable. $0 \leq Y \leq 1$.
- The CDF of Y:

$$P(Y \leq y) = P(F(X) \leq y) = P\left(X \leq F^{-1}(y)\right) = F\left(F^{-1}(y)\right) = y$$

- Therefore, $Y \sim \text{Uniform}[0,1]$



Expectation of $F(X_{(k)})$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from a common distribution with PDF $f(x)$ and *strictly increasing* CDF $F(x)$.

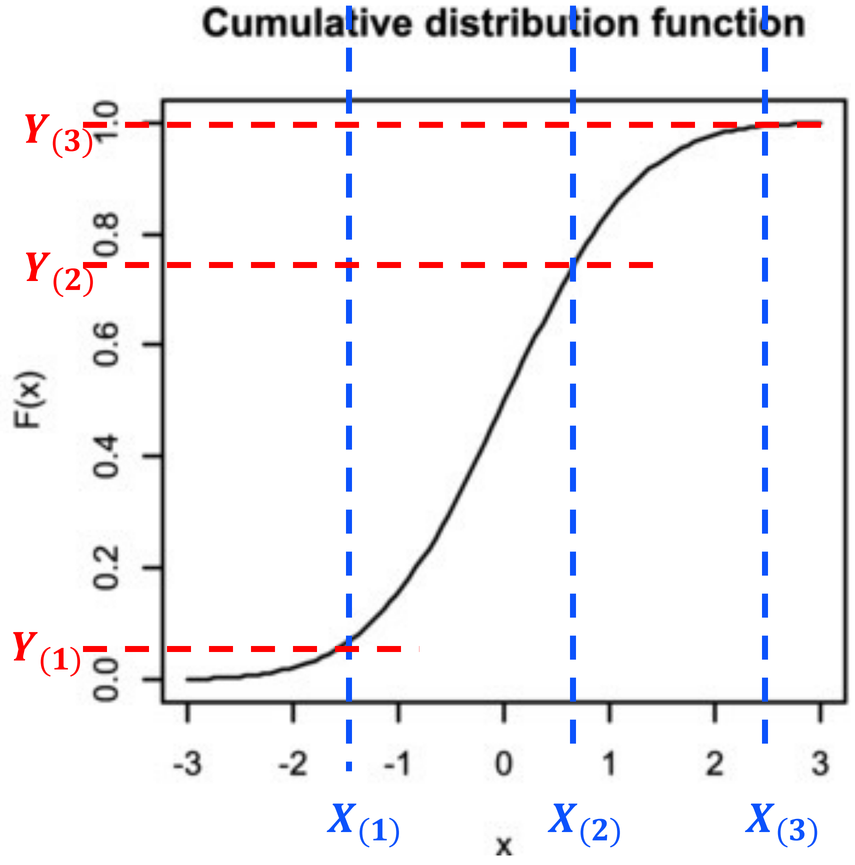
$$X_1, X_2, \ldots, X_n \sim f$$

Let $Y_1 = F(X_1), Y_2 = F(X_2), \ldots, Y_n = f(X_n)$

$$Y_1, Y_2, \ldots, Y_n \sim \mathrm{Uniform}[0,1]$$

$Y_{(k)} = F\left(X_{(k)}\right)$ is the k-th order statistic for uniform distributions:

$$E\left(Y_{(k)}\right) = E\left(F\left(X_{(k)}\right)\right) = \frac{k}{n+1}$$



Q-Q plot

Theoretical vs Sample Quantiles

- The p-th **theoretical quantile** of the distribution F is $\boldsymbol{\pi_p}$:

$$\boxed{F(\pi_p) = p} \qquad\qquad \pi_p = F^{-1}(p)$$

- For example, when $p = 0.5, \pi_{0.5}$ is called the <u>median</u> of the distribution F.
- For standard normal distribution, we have $\pi_p = Z_{1-p}$

How to estimate π_p from samples $X_1, X_2, \dots, X_n \sim F$?

$$\boxed{E\left(F\left(X_{(k)}\right)\right) = \frac{k}{n+1}}$$

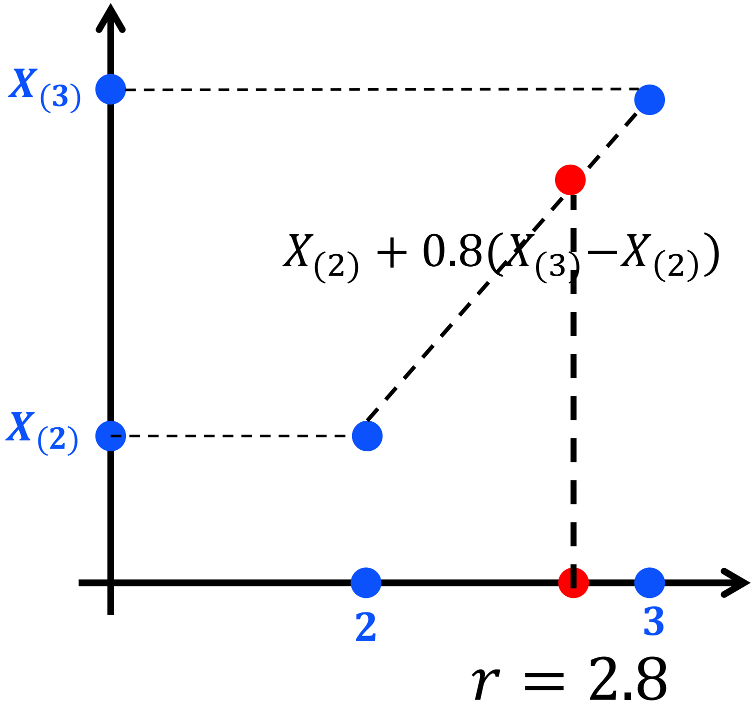
- For $p = \frac{k}{n+1}$, $X_{(k)}$ is an estimate for π_p.
- $X_{(k)}, k = (n+1)p$ is the p-th **sample quantile** $\boldsymbol{\hat{\pi}_p}$.

Theoretical vs Sample Quantiles

Calculation of the p-th sample quantile

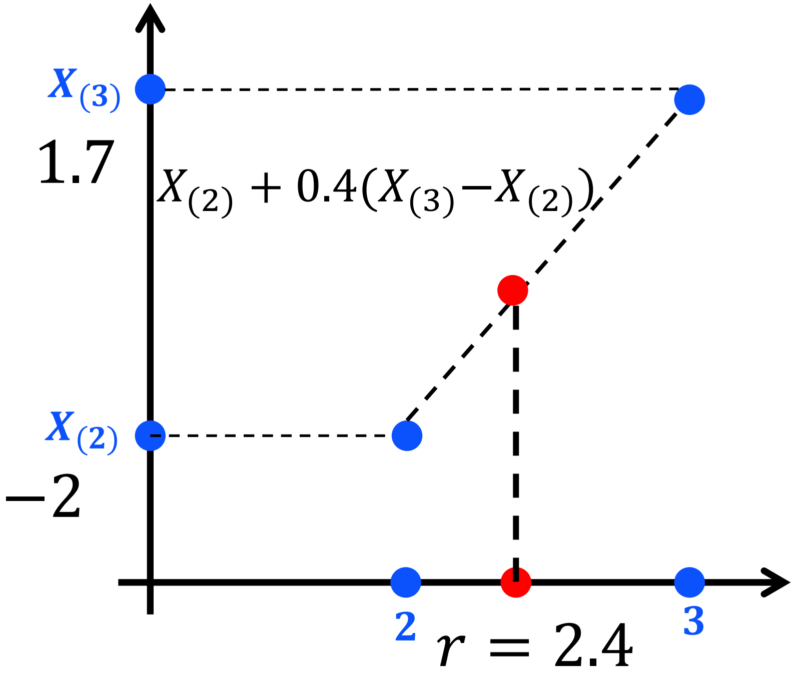
- Order the samples $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$
- $r = (n+1)p$
- If r is an integer, the sample quantile is $\hat{\pi}_p = X_{(r)}$
- If r is not an integer, $r = \lfloor r \rfloor + (r - \lfloor r \rfloor)$, the sample quantile is

$$\hat{\pi}_p = X_{(\lfloor r \rfloor)} + (r - \lfloor r \rfloor)(X_{(\lfloor r \rfloor + 1)} - X_{(\lfloor r \rfloor)})$$



Example

i	$x_{(i)}$	$i/(n+1)$
1	-3.9	1/6
2	-2.0	2/6
3	1.7	3/6
4	7.3	4/6
5	11.7	5/6



- $p = \frac{1}{2}$, the sample median is $X_{(3)} = 1.7$
- $p = 0.4 \Rightarrow r = (n+1)p = 2.4$

$$\hat{\pi}_{0.4} = X_{(2)} + 0.4(X_{(3)} - X_{(2)}) = -0.52$$

Quantile-Quantile (Q-Q) plot

Given data $x_1, x_2, \ldots, x_n$, we suspect that they are coming from a common distribution F, say normal distribution or exponential distribution or gamma distribution. We can then compute the theoretical quantiles $\pi_{i/(n+1)} = F^{-1}\left(\frac{i}{n+1}\right)$ for $i = 1,2, \ldots, \mathrm{n}$. If the data is indeed statistically similar to the distribution, then we expect that

$$\boxed{\underbrace{\widehat{\pi}_{i/(n+1)} = x_{(i)}}_{\text{sample quantiles}} \approx \underbrace{\pi_{i/(n+1)}}_{\text{theoretical quantiles}}}$$

If we plot a scatterplot of the pairs $(\pi_{i/(n+1)}, x_{(i)})$, it should be close to the line y = x.

This is what we call **quantiles-quantiles plot or QQ plots**, as we plot the theoretical quantiles against the sample quantiles.

Q-Q plot Examples

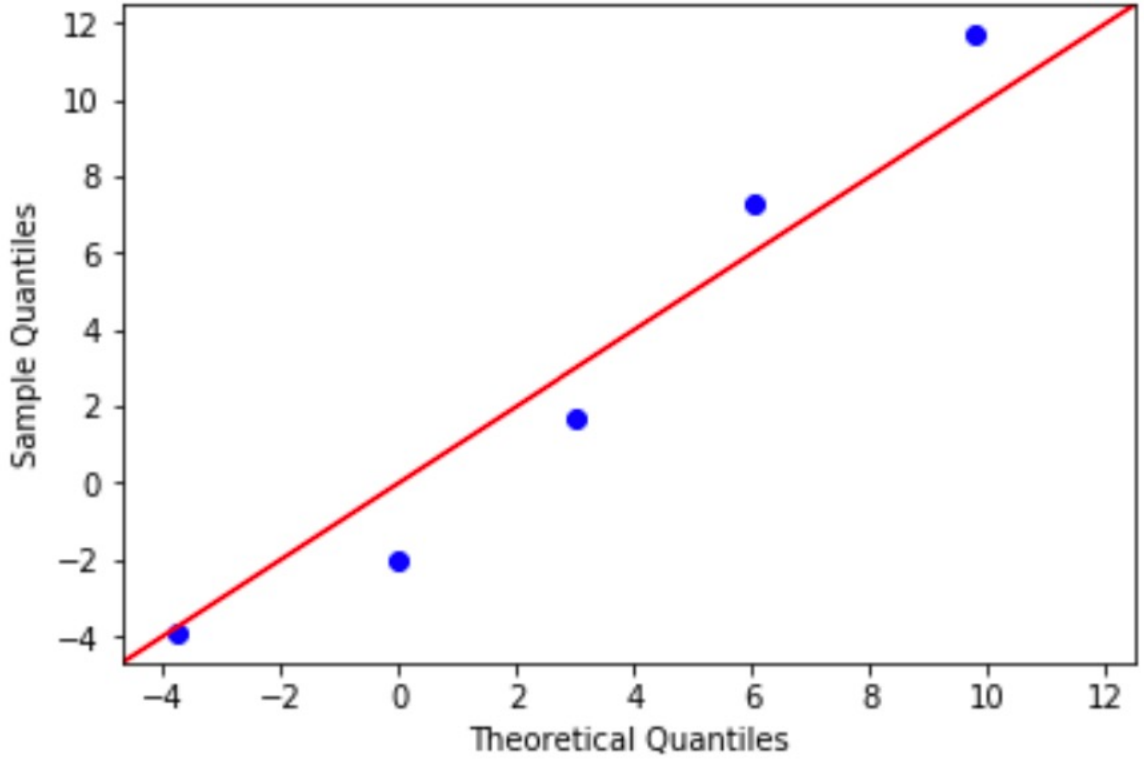
i	$x_{(i)}$	$i/(n+1)$
1	-3.9	1/6
2	-2.0	2/6
3	1.7	3/6
4	7.3	4/6
5	11.7	5/6

We suspect that the data is coming from $N(3, 7^2)$, and we would like to draw a QQ plot.

Q-Q plot Examples

i	$x_{(i)}$	$i/(n+1)$	$\pi_{i/(n+1)}$
1	-3.9	1/6	-3.77
2	-2.0	2/6	-0.02
3	1.7	3/6	3
4	7.3	4/6	6.02
5	11.7	5/6	9.77

Sample quantiles: $x_{(i)}$; Theoretical quantiles: $\pi_{i/(n+1)}$



Theoretical quantiles of $N(3,7^2)$

$$X \sim N(3,7^2)$$

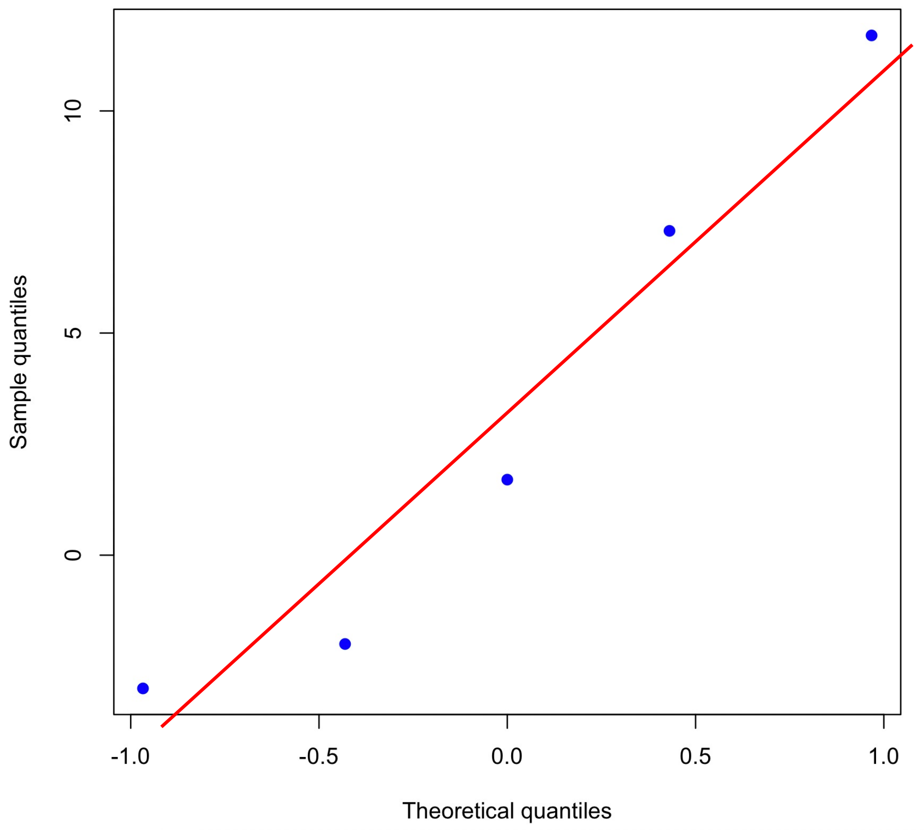
$$P(X < \pi_p) = P\left(\frac{X-3}{7} \le \frac{\pi_p - 3}{7}\right) = p$$

$$\Rightarrow \frac{\pi_p - 3}{7} = Z_{1-p} \Rightarrow \pi_p = 3 + 7Z_{1-p}$$

$$\pi_{i/(n+1)} = 3 + 7 \times Z_{1-i/(n+1)}$$

Q-Q plot Examples

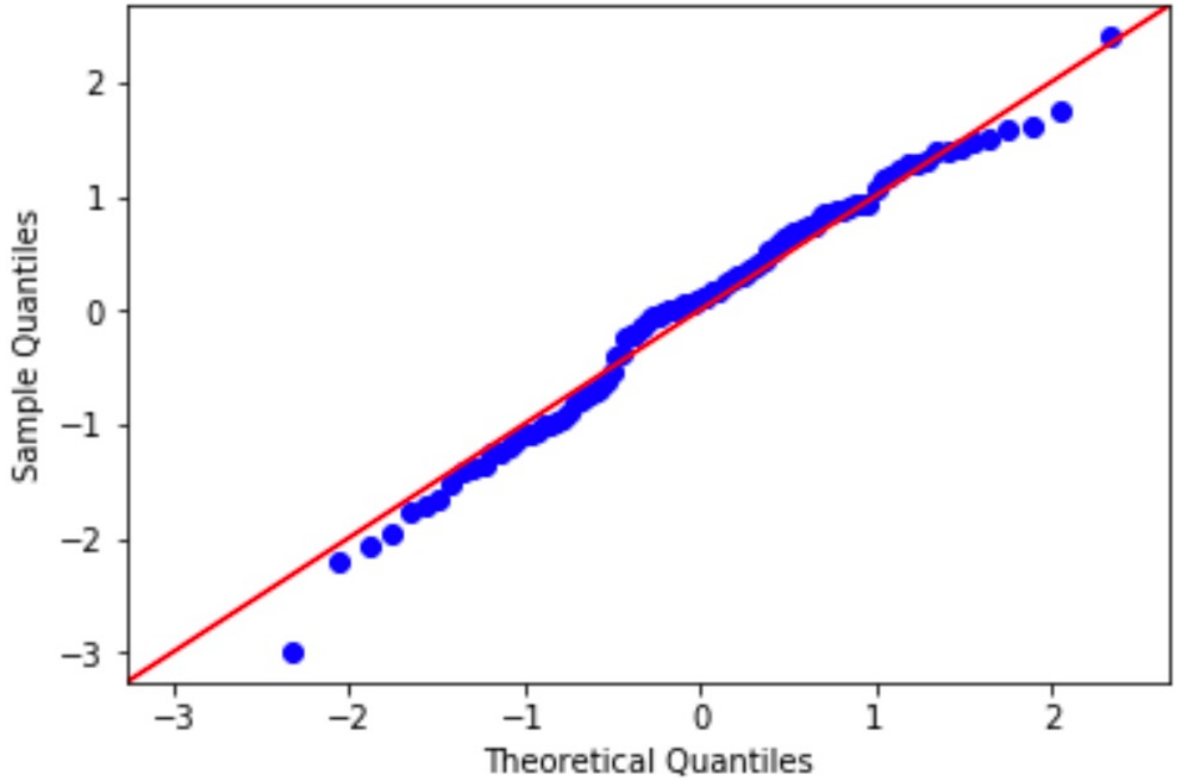
Sample quantiles			Theoretical quantiles of N(0,1)
i	$x_{(i)}$	$i/(n+1)$	$\pi_{i/(n+1)}$
1	-3.9	1/6	-0.97
2	-2.0	2/6	-0.43
3	1.7	3/6	0
4	7.3	4/6	0.43
5	11.7	5/6	0.97



The distributions can be *shifted* and *stretched*.
As long as the shapes match, the Q-Q plot is close to a straight line.

Example: A good fit

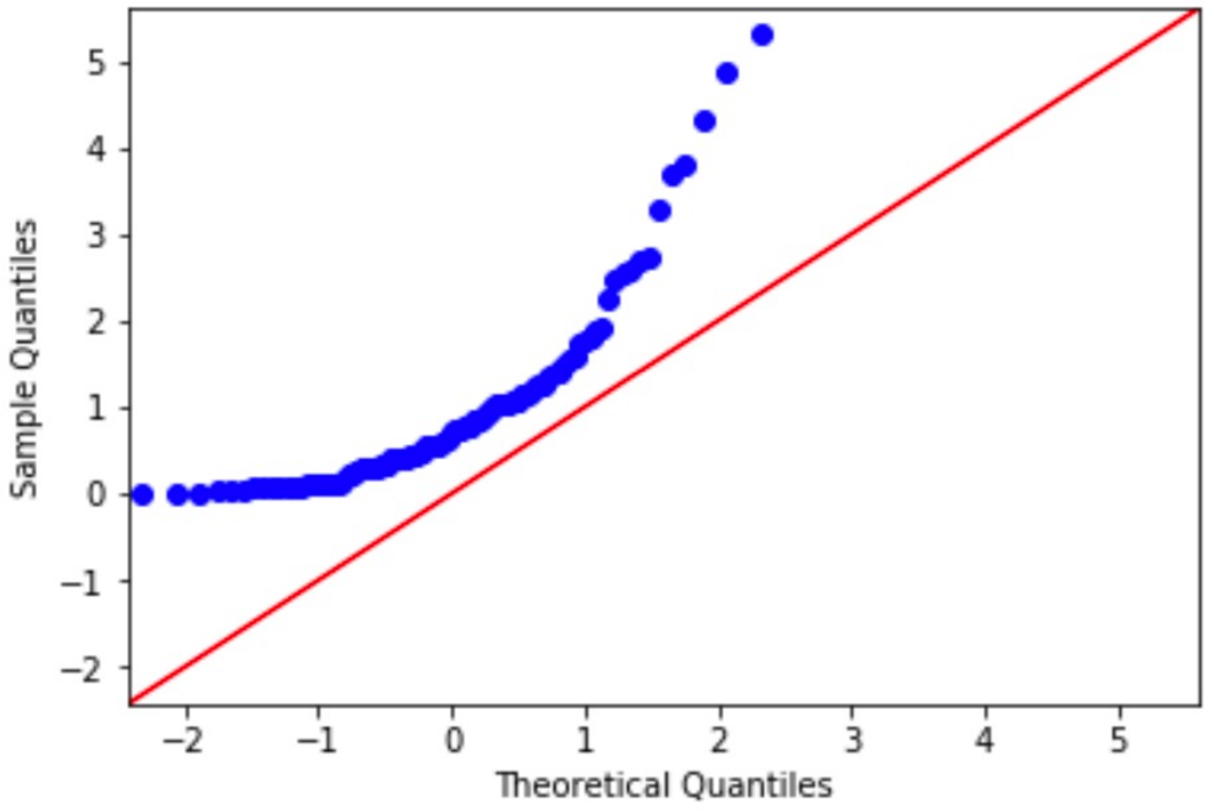
Hypothesis: the data is coming from standard normal.



This dataset looks like a good fit.

Example: A poor fit

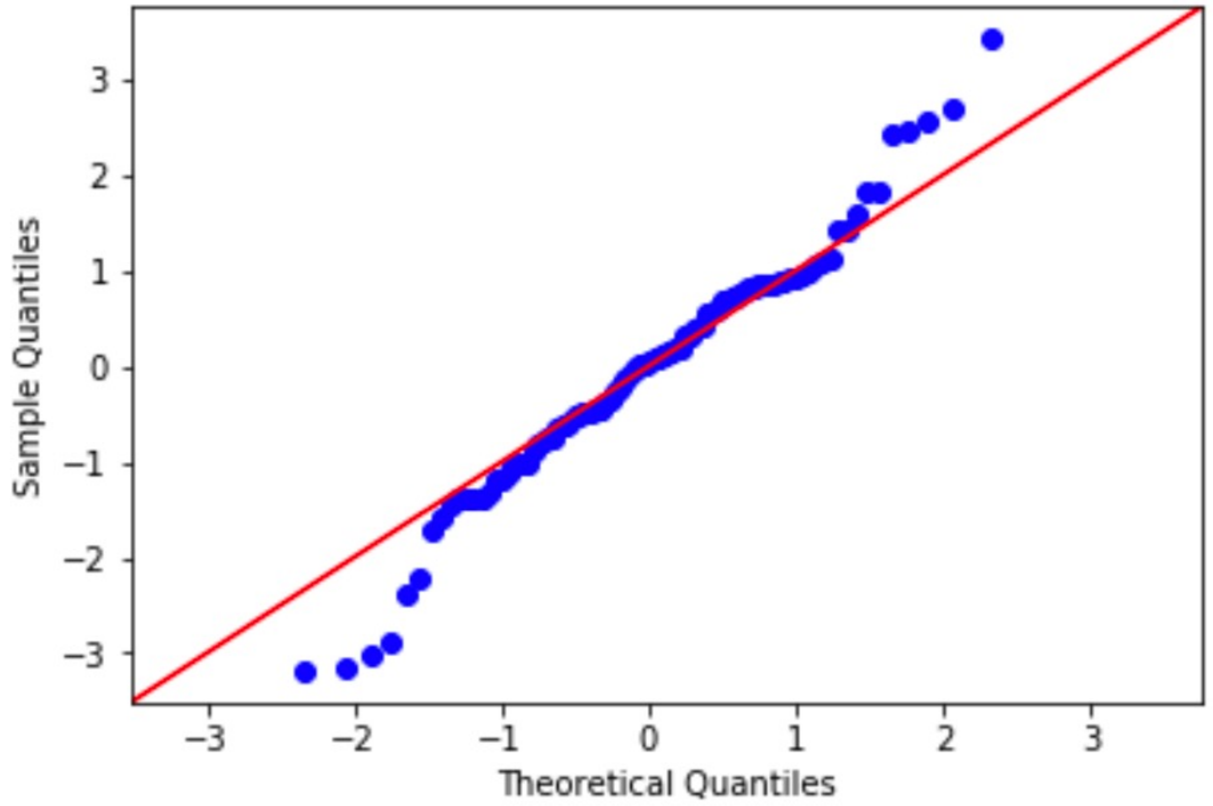
Hypothesis: the data is coming from standard normal.



This is a poor fit, since the data is clearly non-negative, and is more likely to take on larger values than normal distribution (larger "right tail").

Example: A poor fit

Hypothesis: the data is coming from standard normal.



This is a poor fit, since the data is more likely to take on extremely large or extremely small values than standard normal (a heavy-tailed distribution)