AUA CS 108, Statistics, Fall 2019

Lecture 11

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

18 Sep 2019

Contents

Last Lecture ReCap

- What is a Sample α-Quantile?

Last Lecture ReCap

- What is a Sample α-Quantile?
- What is a Theoretical α-Quantile?

Last Lecture ReCap

- What is a Sample α-Quantile?
- What is a Theoretical α-Quantile?
- When do we will write our next Quiz?

Last Lecture ReCap

- What is a Sample α-Quantile?
- What is a Theoretical α-Quantile?
- When do we will write our next Quiz?
- What is the 35% Quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 5x^4, & x \in [0,1] \\ 0, & \textit{otherwise} \end{cases} \quad ?$$

Last Lecture ReCap

- What is a Sample α-Quantile?
- What is a Theoretical α-Quantile?
- When do we will write our next Quiz?
- What is the 35% Quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 5x^4, & x \in [0,1] \\ 0, & \textit{otherwise} \end{cases} \quad ?$$

- What is a Q-Q Plot?

Last Lecture ReCap

- What is a Sample α-Quantile?
- What is a Theoretical α-Quantile?
- When do we will write our next Quiz?
- What is the 35% Quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 5x^4, & x \in [0,1] \\ 0, & \textit{otherwise} \end{cases} \quad ?$$

- What is a Q-Q Plot?
- What is it for?

Q-Q Plots, Data vs Data

Problem: we have two Datasets, not necessarily of the same size:

$$x: \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \; y_1, y_2, ..., y_m,$$

and we want to see if x and y are coming from the same Distribution

Q-Q Plots, Data vs Data

Problem: we have two Datasets, not necessarily of the same size:

$$x: \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \; y_1, y_2, ..., y_m,$$

and we want to see if x and y are coming from the same Distribution

To draw a **Q-Q Plot** (Quantile-Quantile Plot), we take some levels of quantiles, say, for some p,

$$\alpha = \frac{1}{p}, \frac{2}{p}, ..., \frac{p-1}{p}$$

and then draw the points (q_α^x, q_α^y).

Q-Q Plots, Data vs Data

Problem: we have two Datasets, not necessarily of the same size:

$$x: \; x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \; y_1, y_2, ..., y_m,$$

and we want to see if x and y are coming from the same Distribution

To draw a **Q-Q Plot** (Quantile-Quantile Plot), we take some levels of quantiles, say, for some p,

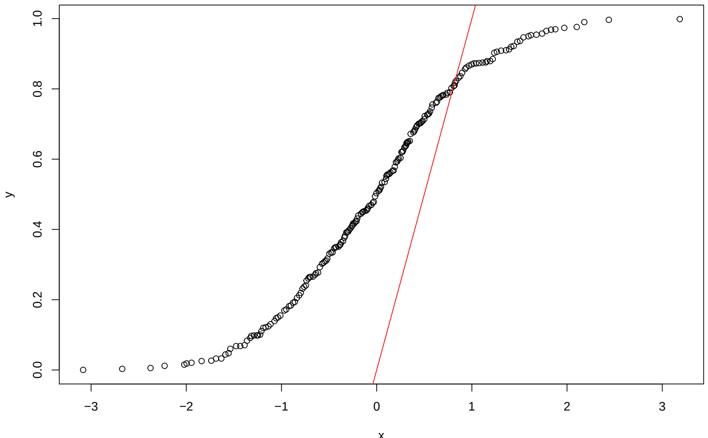
$$\alpha = \frac{1}{p}, \frac{2}{p}, ..., \frac{p-1}{p}$$

and then draw the points (q_α^x, q_α^y).

Idea: If x and y are coming from the same Distribution, then the Quantiles of x and y need to be approximately the same, $q_\alpha^x \approx q_\alpha^y$, so geometrically, the points (q_α^x, q_α^y) need to be close to the bisector line.

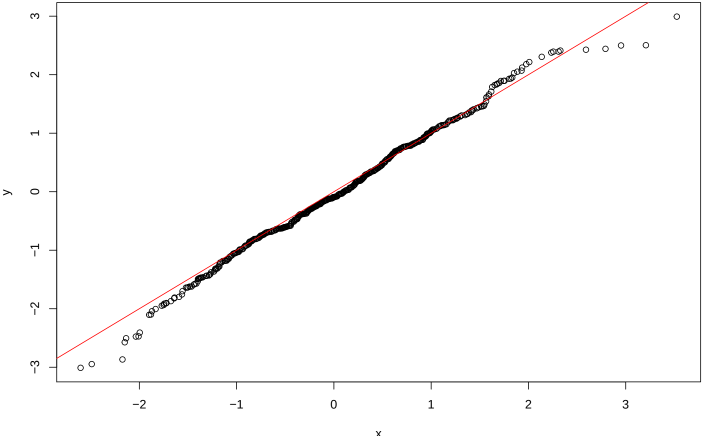
Example, Q-Q Plots, Data vs Data

```
x <- rnorm(1000)
y <- runif(200)
qqplot(x,y)
abline(0,1, col="red")
```



Example, Q-Q Plots, Data vs Data

```
x <- rnorm(1000)
y <- rnorm(500)
qqplot(x,y)
abline(0,1, col="red")
```



Example, Q-Q Plot by Hands, Data vs Data

Example: Assume

$$x:\ -1, 2, 1, 2, 3, 2, 1 \qquad y:\ 0, 3, 4, 1, 1, 1, 1, 2$$

Draw the Q-Q Plot for x and y.

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f).

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f). The Problem is to estimate visually if the Dataset comes from that Distribution.

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f). The Problem is to estimate visually if the Dataset comes from that Distribution.

Example: Say, is the following Dataset

```
##  [1] -0.83 -0.84 -0.61 -0.40  0.56 -0.96  0.58  0.41 -0.
## [13]  0.45 -0.63  0.94  0.22 -0.54 -0.28 -0.12 -0.87
```

from a Normal Distribution?

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f). The Problem is to estimate visually if the Dataset comes from that Distribution.

Example: Say, is the following Dataset

```
##  [1] -0.83 -0.84 -0.61 -0.40  0.56 -0.96  0.58  0.41 -0.
## [13]  0.45 -0.63  0.94  0.22 -0.54 -0.28 -0.12 -0.87
```

from a Normal Distribution?

To answer this question, we again take some levels of quantiles, say, for some p,

$$\alpha = \frac{1}{p}, \frac{2}{p}, ..., \frac{p-1}{p}$$

and then draw the points (q_α^F, q_α^x), where q_α^F is the α-quantile of the Theoretical Distribution, and q_α^x is the α-quantile of x.

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f). The Problem is to estimate visually if the Dataset comes from that Distribution.

Example: Say, is the following Dataset

```
##  [1] -0.83 -0.84 -0.61 -0.40  0.56 -0.96  0.58  0.41 -0.
## [13]  0.45 -0.63  0.94  0.22 -0.54 -0.28 -0.12 -0.87
```

from a Normal Distribution?

To answer this question, we again take some levels of quantiles, say, for some p,

$$\alpha = \frac{1}{p}, \frac{2}{p}, ..., \frac{p-1}{p}$$

and then draw the points (q^F_α, q^x_α), where q^F_α is the α-quantile of the Theoretical Distribution, and q^x_α is the α-quantile of x.

Idea: If x is from the Distribution given by F, then we need to have $q^F_\alpha \approx q^x_\alpha$, so, graphically, the point will be close to the bisector.

Normal Q-Q Plot

In **R**, we have a function `qqnorm` which plots the Q-Q Plot for the Dataset x vs the Normal Distribution.

Normal Q-Q Plot

In **R**, we have a function `qqnorm` which plots the Q-Q Plot for the Dataset x vs the Normal Distribution. Unfortunately, we do not have this kind of function for other standard distributions, say, Uniform.

[1] or one can write his/her own function `qqunif` or `qqexp`, say

Normal Q-Q Plot

In **R**, we have a function `qqnorm` which plots the Q-Q Plot for the Dataset x vs the Normal Distribution. Unfortunately, we do not have this kind of function for other standard distributions, say, Uniform. But one can use the `qqplot(x,y)` command, by generating y from the given Distribution[1].

[1] or one can write his/her own function `qqunif` or `qqexp`, say

Normal Q-Q Plot

In **R**, we have a function `qqnorm` which plots the Q-Q Plot for the Dataset x vs the Normal Distribution. Unfortunately, we do not have this kind of function for other standard distributions, say, Uniform. But one can use the `qqplot(x,y)` command, by generating y from the given Distribution[1].

Another **R** command is `qqline` which adds a line passing (by default) through the first and third Quartiles,

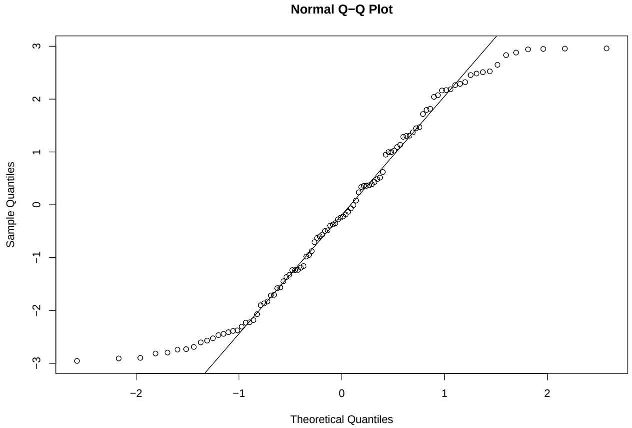
$$(q^F_{0.25}, q^x_{0.25}) \qquad \textit{and} \qquad (q^F_{0.75}, q^x_{0.75}).$$

[1] or one can write his/her own function `qqunif` or `qqexp`, say

Some Experiments

Here are some experiments with `qqnorm`

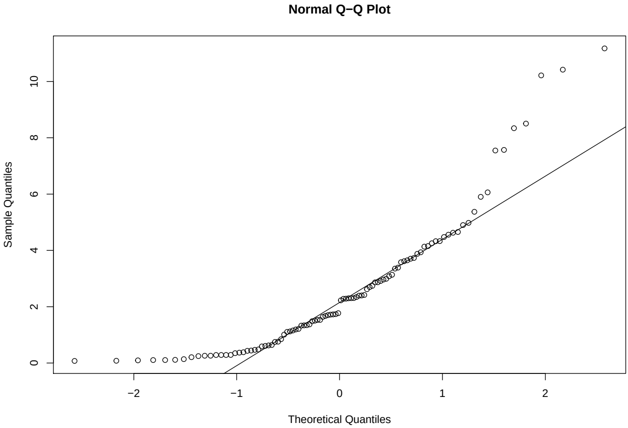
```
x <- runif(100,-3,3)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with `qqnorm`

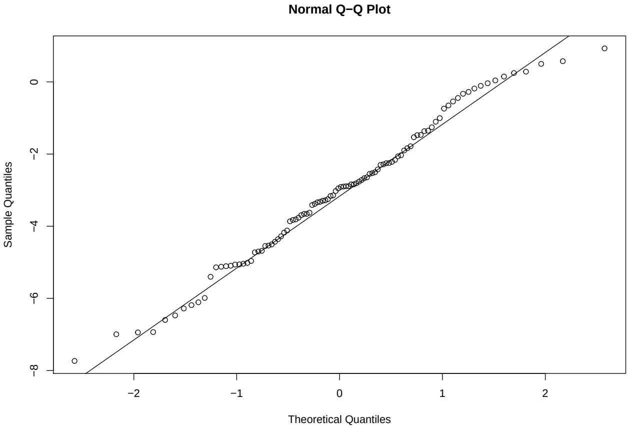
```
x <- rexp(100,0.4)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with qqnorm

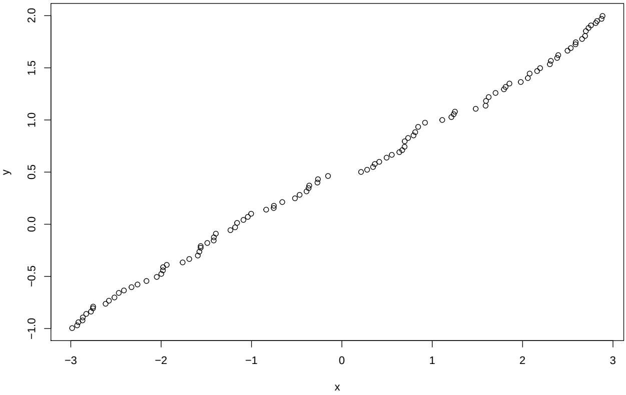
```
x <- rnorm(100, mean = -3, sd = 2)
qqnorm(x)
qqline(x)
```



Some Experiments

Now, assume we want to see if our Dataset x is from $Unif[-1, 2]$:

```
x <- runif(100,-3,3)
y <- runif(1000,-1,2)
qqplot(x,y)
```



Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*.

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting.

[2] Can you state rigorously and prove this?

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

---

[2] Can you state rigorously and prove this?

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

So if, say, x is a sample from $\mathcal{N}(2, 3^2)$, then

- when doing a Q-Q Plot of x vs $\mathcal{N}(2, 3^2)$, the Quantiles will be

[2] Can you state rigorously and prove this?

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

So if, say, x is a sample from $\mathcal{N}(2, 3^2)$, then

- when doing a Q-Q Plot of x vs $\mathcal{N}(2, 3^2)$, the Quantiles will be on the bisector;

[2] Can you state rigorously and prove this?

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

So if, say, x is a sample from $\mathcal{N}(2, 3^2)$, then

- when doing a Q-Q Plot of x vs $\mathcal{N}(2, 3^2)$, the Quantiles will be on the bisector;
- when doing a Q-Q Plot of x vs $\mathcal{N}(0, 1)$, the Quantiles will be

[2] Can you state rigorously and prove this?

Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

So if, say, x is a sample from $\mathcal{N}(2, 3^2)$, then

- when doing a Q-Q Plot of x vs $\mathcal{N}(2, 3^2)$, the Quantiles will be on the bisector;
- when doing a Q-Q Plot of x vs $\mathcal{N}(0, 1)$, the Quantiles will be on some line (can you find the line equation?);

[2] Can you state rigorously and prove this?

Important Note

So if `qqnorm` shows that the quantiles are close to a line, that means that the Dataset is possibly from a Normal Distribution.

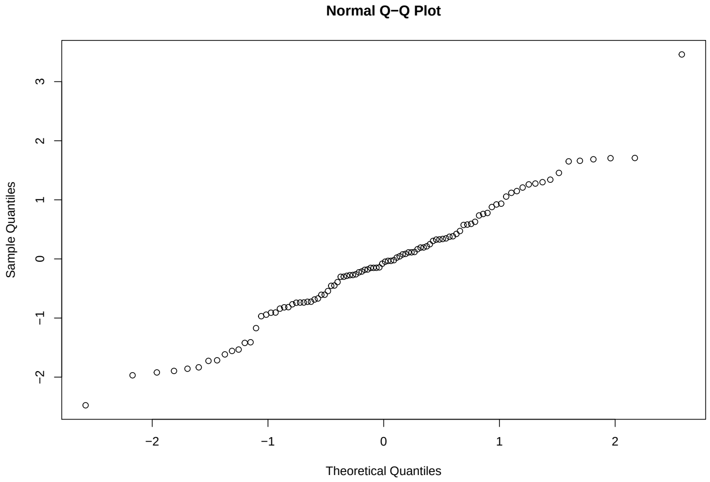
Important Note

So if `qqnorm` shows that the quantiles are close to a line, that means that the Dataset is possibly from a Normal Distribution.

And if `qqnorm` shows that the quantiles are close to the bisector, that means that the Dataset is possibly from the Standard Normal Distribution.

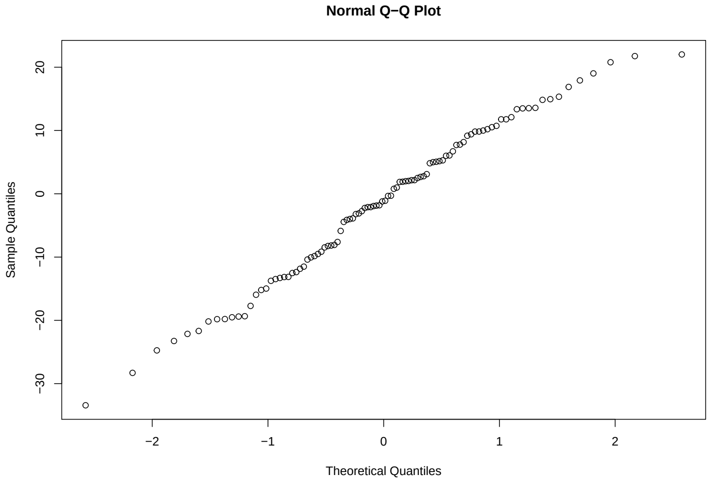
Some Experiments

```
x <- rnorm(100, mean=0, sd=1)
qqnorm(x)
```



Some Experiments

```r
x <- rnorm(100, mean=2, sd=12)
qqnorm(x)
```



Exercise: Express the Quantiles of $\mathcal{N}(\mu, \sigma^2)$ in terms of the quantiles of $\mathcal{N}(0, 1)$.

Important Note, v2

The above important note works also for the Uniform Distribution. This is again because all Uniform Distributions are the scaled-translated versions of the Standard Uniform $Unif\,[0,1]$.

Important Note, v2

The above important note works also for the Uniform Distribution. This is again because all Uniform Distributions are the scaled-translated versions of the Standard Uniform $Unif[0,1]$.

So if you will compare your Dataset with $Unif[0,1]$, and Q-Q Plot will show that the Quantiles are close to a line, that means that probably your Dataset is from a Uniform Distribution, with some parameters.

Exercise: Express the Quantiles of $Unif[a,b]$ in terms of the quantiles of $Unif[0,1]$.