AUA CS108, Statistics, Fall 2020

Lecture 14

Michael Poghosyan

28 Sep 2020

Contents

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]  0.033  0.136  0.887  0.764 -0.749  0.987  0.347  0
## [11] -0.405 -0.645  0.612  0.401  0.233 -0.920 -0.133  0
```

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]  0.033  0.136  0.887  0.764 -0.749  0.987  0.347  0
## [11] -0.405 -0.645  0.612  0.401  0.233 -0.920 -0.133  0
```

from a Normal Distribution?

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

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

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]  0.033  0.136  0.887  0.764 -0.749  0.987  0.347  0
## [11] -0.405 -0.645  0.612  0.401  0.233 -0.920 -0.133  0
```

from a Normal Distribution?

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

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

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,

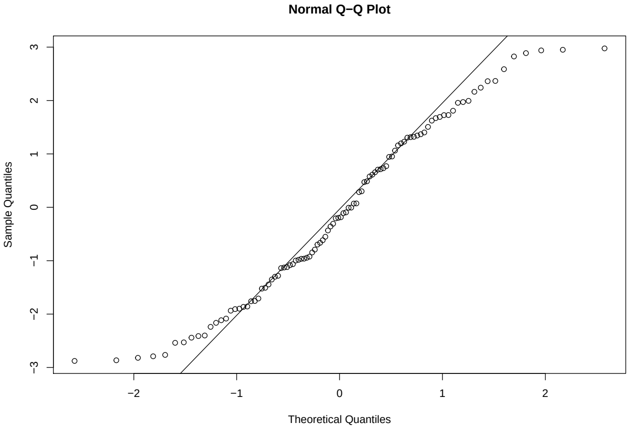
$$(q^F_{0.25}, q^x_{0.25}) \qquad and \qquad (q^F_{0.75}, q^x_{0.75}).$$

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

Some Experiments

Here are some experiments with `qqnorm`

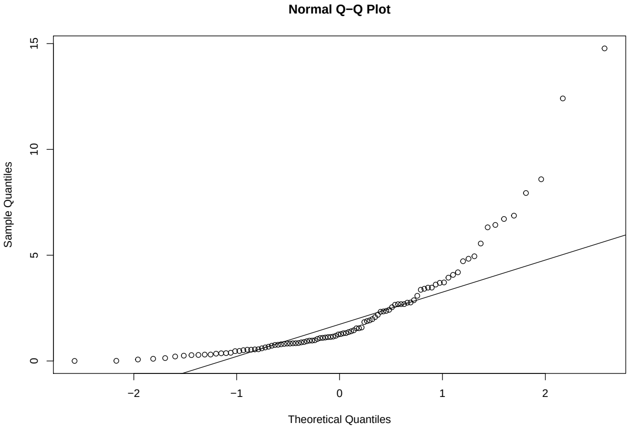
```r
x <- runif(100,-3,3)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with `qqnorm`

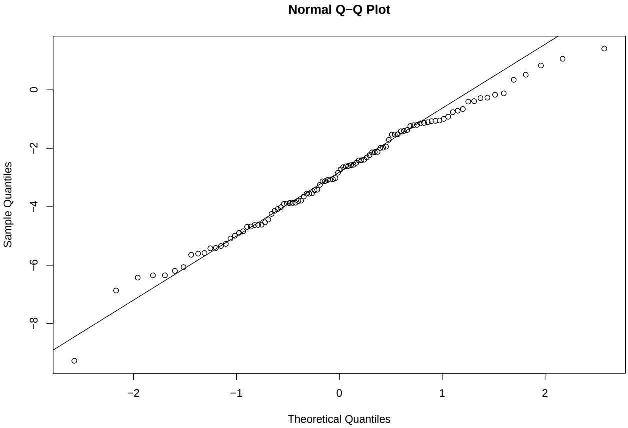
```
x <- rexp(100,0.4)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with `qqnorm`

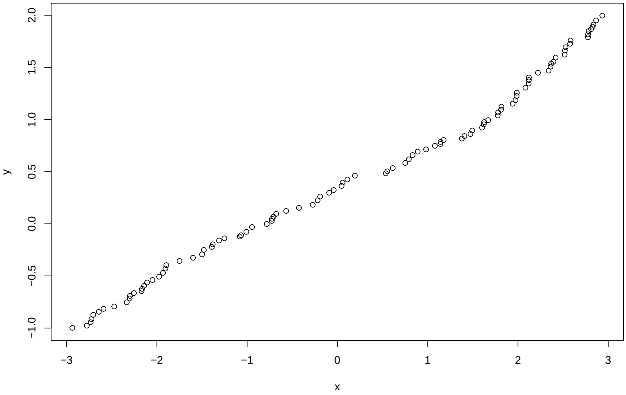
```
x <- rnorm(100, mean = -3, sd = 2)
qqnorm(x)
qqline(x)
```



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.

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.

Note: The theoretical justification of the above is in the following: if z_α is the quantile of order α of $\mathcal{N}(0, 1)$, and if q_α is the same order quantile of $\mathcal{N}(\mu, \sigma^2)$, then there is a linear relationship between q_α and z_α.

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.

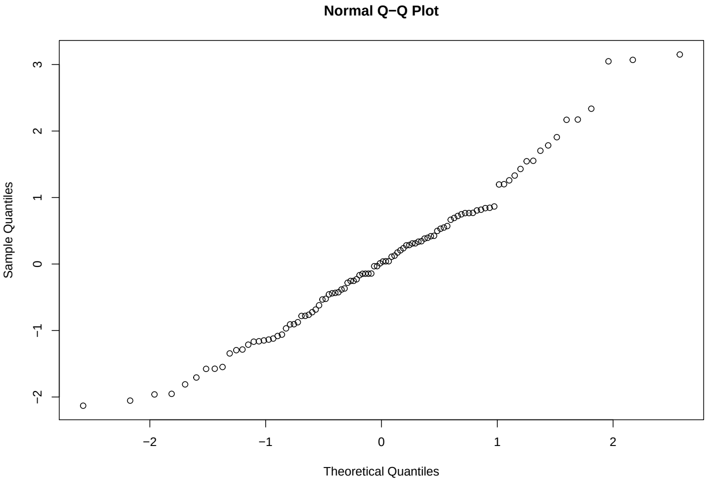
And if `qqnorm` shows that the quantiles are close to the bisector, that means that the Dataset is possibly from the Standard Normal Distribution.

Note: The theoretical justification of the above is in the following: if z_α is the quantile of order α of $\mathcal{N}(0,1)$, and if q_α is the same order quantile of $\mathcal{N}(\mu,\sigma^2)$, then there is a linear relationship between q_α and z_α.

Exercise: Find that relationship in terms of μ and σ.

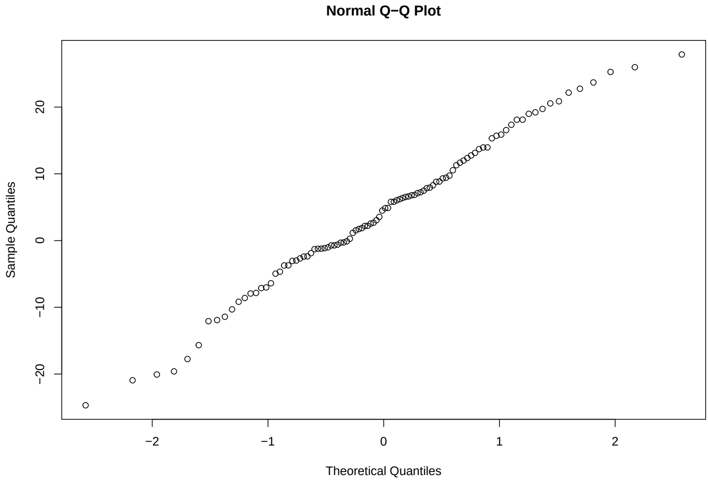
Some Experiments

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



Some Experiments

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



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.

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: Find a relationship between the quantiles of $Unif[a,b]$ and $Unif[0,1]$.

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G).

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

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

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

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