YSU ASDS, Statistics, Fall 2019

Lecture 07-08

Michael Poghosyan

17 Sep 2019

Descriptive Statistics

Contents

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD
- What are the Quartiles?

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD
- What are the Quartiles?
- What is a BoxPlot?

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD
- What are the Quartiles?
- What is a BoxPlot?
- What is it for?

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD
- What are the Quartiles?
- What is a BoxPlot?
- What is it for?
- What is an Outlier?

BoxPlot, Common Error

Here is a common error when Plotting the BoxPlot:

BoxPlot, Common Error

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**!

BoxPlot, Common Error

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**! W_1 and W_2 need to be from our Dataset!

BoxPlot, Common Error

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**! W_1 and W_2 need to be from our Dataset!

Take as W_1 and W_2 the smallest and largest **Datapoints**, respectively, in

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

Additions/Variations:

Some Variations:

- Variable Width BoxPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot
- BeanPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot
- BeanPlot

See, for Example, this page.

Boxplot, Why we use it

We use BoxPlots to:

Boxplot, Why we use it

We use BoxPlots to:

- Visualize the distribution of the Dataset

Boxplot, Why we use it

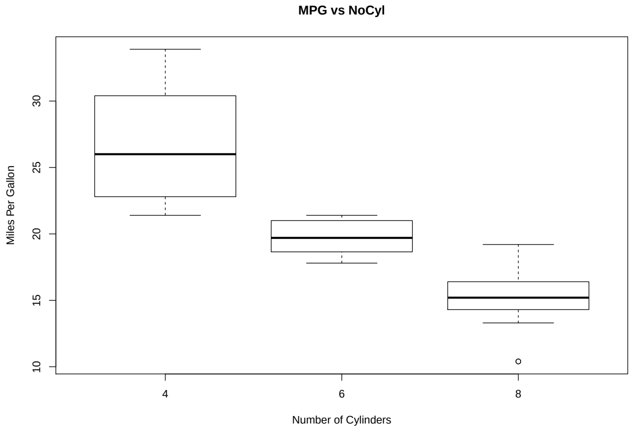
We use BoxPlots to:

- Visualize the distribution of the Dataset
- To compare two or more Datasets

Example

Here we use the `mtcars` Dataset:

```
boxplot( mpg~cyl, data=mtcars, main="MPG vs NoCyl",
   xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

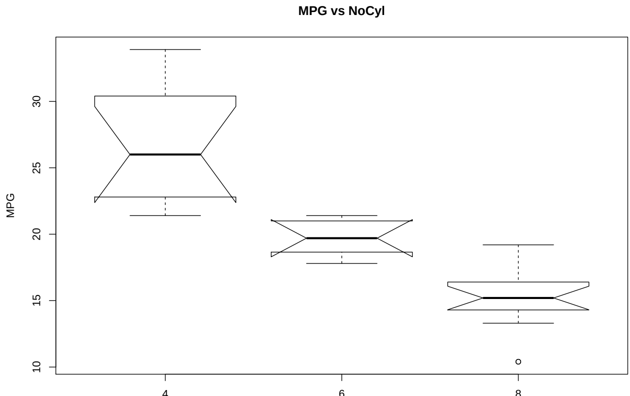


Example

Again,

```
boxplot( mpg~cyl, data=mtcars, notch = T,
         main="MPG vs NoCyl", xlab="Number of Cylinders", y
```

```
## Warning in bxp(list(stats = structure(c(21.4, 22.8, 26,
## some notches went outside hinges ('box'): maybe set notc
```



Note

Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2} IQR,\ Q_3 + \frac{3}{2} IQR \right].$$

Note

Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles.
Recall that:

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$. We want to find a real number q_α dividing our (sorted) Dataset into the proportion $100\alpha\% - 100(1-\alpha)\%$, i.e., q_α is a point such that the α-portion of the Datapoints are to the left to q_α, and others are to the right.

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics.

Sample Quantiles

Let $x: \; x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0,1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics.

Note: There are different definitions of the α-quantile in the literature and in software implementations. Say, **R** has 9 methods to calculate quantiles.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%).

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%). By our definition, *Quantile is a Datapoint*, but a Quartile is not necessarily a Datapoint.

Sample Quantiles

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: In the case when $[\alpha \cdot n] = 0$, we take $x_{(0)} = x_{(1)}$.

Note: Quartiles are not always Quantiles (in the sense of our definitions). Say, Q_1 is not always the 25% Quantile (despite their idea is to split the Dataset into the proportion 25%-75%). By our definition, *Quantile is a Datapoint*, but a Quartile is not necessarily a Datapoint.

Note: Sometimes Quantiles are called Percentiles.

Example

Example: Find the 20% and 60% quantiles of

$$x : -2, 3, 5, 7, 8, -3, 4, 5, 2$$

Solution: OTB

Example

Now, let us calculate Quantiles in **R**:

```
x <- 1:15
quantile(x,0.21)
```

```
##  21%
## 3.94
```

```
quantile(x, c(0.1,0.3,0.7))
```

```
##  10%  30%  70%
##  2.4  5.2 10.8
```

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$.

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0, 1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0,1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

If F is stricly increasing and continuous, then we can define

$$F(q_\alpha) = \alpha, \qquad i.e., \qquad q_\alpha = F^{-1}(\alpha).$$

Theoretical Quantiles

Now assume X is a r.v. with CDF $F(x)$ and PDF $f(x)$. For $\alpha \in (0, 1)$, we define the α-quantile q_α to be the real number satisfying:

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

If F is stricly increasing and continuous, then we can define

$$F(q_\alpha) = \alpha, \qquad i.e., \qquad q_\alpha = F^{-1}(\alpha).$$

If F has a Density, $f(x)$, then q_α can be calculated from

$$\int_{-\infty}^{q_\alpha} f(x)dx = \alpha.$$

Theoretical Quantiles, Geometrically, by CDF

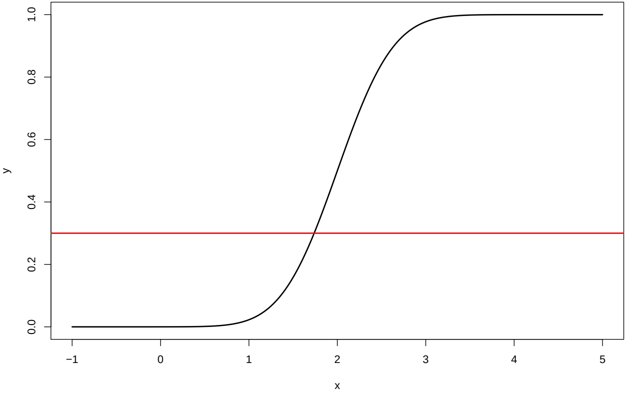
First we draw the CDF $y = F(x)$ graph, then draw the line $y = \alpha$.

Theoretical Quantiles, Geometrically, by CDF

First we draw the CDF $y = F(x)$ graph, then draw the line $y = \alpha$. Now, we keep the portion of the graph of $y = F(x)$ above (or on) the line $y = \alpha$. Then we take the leftmost point of the remaining part, and the x-coordinate of that point will be q_α.

Theoretical Quantiles, Geometrically, by CDF

```
alpha <- 0.3
x <- seq(-1,5, by = 0.01)
y <- pnorm(x, mean = 2, sd = 0.5)
plot(x,y, type = "l", xlim = c(-1,5), lwd = 2)
abline(h = alpha, lwd = 2, col = "red")
```



Theoretical Quantiles, Geometrically, by PDF

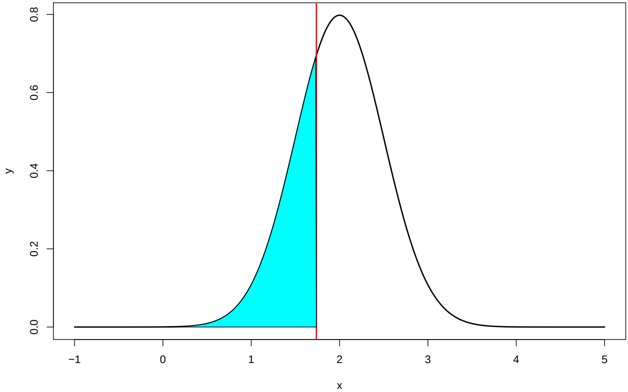
Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$.

Theoretical Quantiles, Geometrically, by PDF

Now, assume our Distribution is continuous. We plot the graph of the PDF $y = f(x)$. We take q_α to be the smallest point such that the area under the PDF curve **left to** q_α is exactly α.

Theoretical Quantiles, Geometrically, by PDF

```
alpha <- 0.3; q.alpha <- qnorm(alpha, mean = 2, sd = 0.5)
x <- seq(-1,5, by = 0.01)
y <- dnorm(x, mean = 2, sd = 0.5)
plot(x,y, type = "l", xlim = c(-1,5), lwd = 2)
abline(v = q.alpha, lwd = 2, col = "red")
polygon(c(x[x<=q.alpha], q.alpha),c(y[x<=q.alpha],0),col="cyan")
```



Examples

Example: Find the 30% quantile of $Unif\,[3, 10]$

Solution: OTB

Examples

Example: Find the 30% quantile of $Unif[3,10]$

Solution: OTB

Example: Find the 70% quantile of the Distribution with the PDF

$$f(x) = \begin{cases} 3x^2, & x \in [0,1] \\ 0, & otherwise \end{cases}$$

Solution: OTB

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \le q_\alpha) \ge \alpha \qquad \text{and} \qquad \mathbb{P}(X \ge q_\alpha) \ge 1 - \alpha.$$

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim Bernoulli(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim Bernoulli(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Note: If $\alpha = 0.5$, we call $q_\alpha = q_{0.5}$ to be the **Median of the Distribution**.

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \le q_\alpha) \ge \alpha \qquad \text{and} \qquad \mathbb{P}(X \ge q_\alpha) \ge 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \le q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim \textit{Bernoulli}(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \le q_\alpha) = \alpha$.

Note: If $\alpha = 0.5$, we call $q_\alpha = q_{0.5}$ to be the **Median of the Distribution**. So if we consider a Continuous r.v. and draw the PDF of that r.v., then the Median is the (leftmost) point dividing the area under the PDF curve into 50%-50% portions.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0, 1)$.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0,1)$.

Say, we will take $\alpha \in (0,1)$ and find two points $a, b \in \mathbb{R}$ such that for $X \sim \mathcal{N}(0,1)$

$$\mathbb{P}(X \leq a) = \mathbb{P}(X \geq b) = \frac{\alpha}{2}.$$

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0, 1)$.

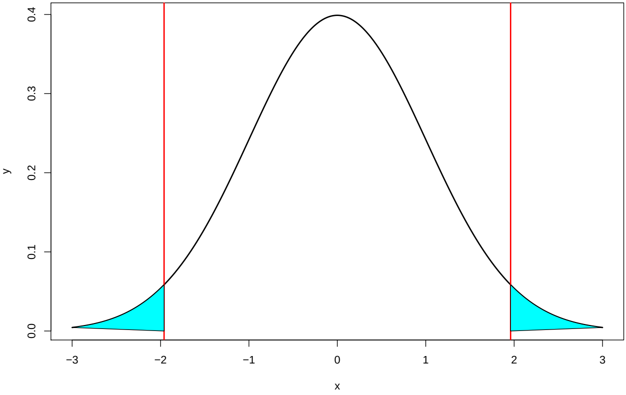
Say, we will take $\alpha \in (0, 1)$ and find two points $a, b \in \mathbb{R}$ such that for $X \sim \mathcal{N}(0, 1)$

$$\mathbb{P}(X \leq a) = \mathbb{P}(X \geq b) = \frac{\alpha}{2}.$$

The idea is to find a symmetric (in fact, the smallest length) interval $[a, b]$ such that for a Standard Normal r.v. X, the chances of $X \notin [a, b]$ are small, are exactly α.

Graphically

```
alpha <- 0.05; z.alpha <- qnorm(alpha/2, mean = 0, sd = 1)
x <- seq(-3,3, by = 0.01)
y <- dnorm(x, mean = 0, sd = 1)
plot(x,y, type = "l", xlim = c(-3,3), lwd = 2)
abline(v = z.alpha, lwd = 2, col = "red")
abline(v = -z.alpha, lwd = 2, col = "red")
polygon(c(x[x<=z.alpha], z.alpha),c(y[x<=z.alpha],0),col="cyan")
polygon(c(x[x>=-z.alpha], -z.alpha),c(y[x>=-z.alpha],0),col="cyan")
```



Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

So

$$b = -z_{\alpha/2} = z_{1-\alpha/2}$$

Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

So

$$b = -z_{\alpha/2} = z_{1-\alpha/2}$$

Note: Please be careful when using Normal Tables. Usually, there is a picture above the table, on which you can find the explanation of the process. Just search “Normal tables” in Google Images.

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;
- a given Dataset comes from a given Distribution;

Q-Q Plots

Next, we consider three important statistical problems: Check visually if

- two given Datasets (possibly, of different sizes) are from the same Distribution;
- a given Dataset comes from a given Distribution;
- given two theoretical Distributions, check if one of them is a shifted-scaled version of the other one, or check if one has *fatter tails* than the other one

Q-Q Plots, Data vs Data

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

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

Q-Q Plots, Data vs Data

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

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

Question: Are x and y coming from the same Distribution?

Q-Q Plots, Data vs Data

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

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

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually.

Q-Q Plots, Data vs Data

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

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

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually. To draw the Q-Q Plot for Datasets, we 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^x, q_α^y).

Q-Q Plots, Data vs Data

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

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

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually. To draw the Q-Q Plot for Datasets, we take some levels of quantiles, say, for some n,

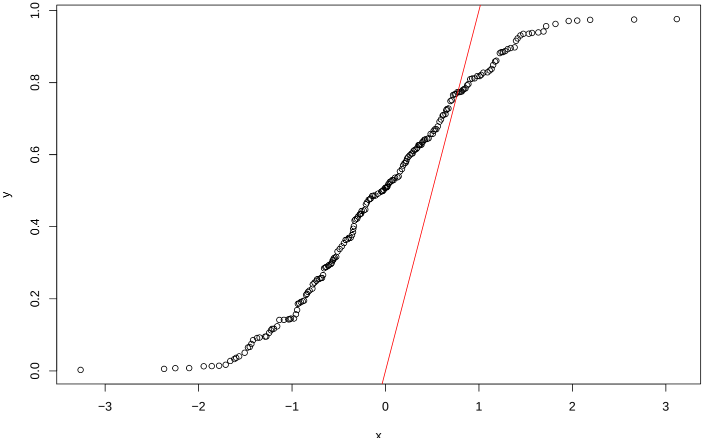
$$\alpha = \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}$$

and then draw the points (q_α^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.

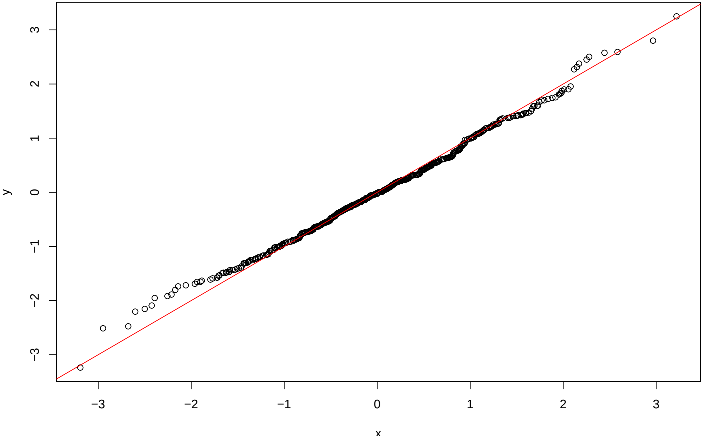
Example, Q-Q Plots, Data vs Data

```
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] -0.49  0.96 -0.84  0.15  0.28 -0.83  0.59 -0.38  0.
## [12] -0.81 -0.85  0.56  0.70  0.36  0.43 -0.23 -0.68 -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] -0.49  0.96 -0.84  0.15  0.28 -0.83  0.59 -0.38  0.
## [12] -0.81 -0.85  0.56  0.70  0.36  0.43 -0.23 -0.68 -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] -0.49  0.96 -0.84  0.15  0.28 -0.83  0.59 -0.38  0.
## [12] -0.81 -0.85  0.56  0.70  0.36  0.43 -0.23 -0.68 -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.

Idea: If x is from the Distribution given by F, then we need to have $q_\alpha^F \approx q_\alpha^x$, 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. Unformtunately, 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. Unformtunately, 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. Unformtunately, 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,

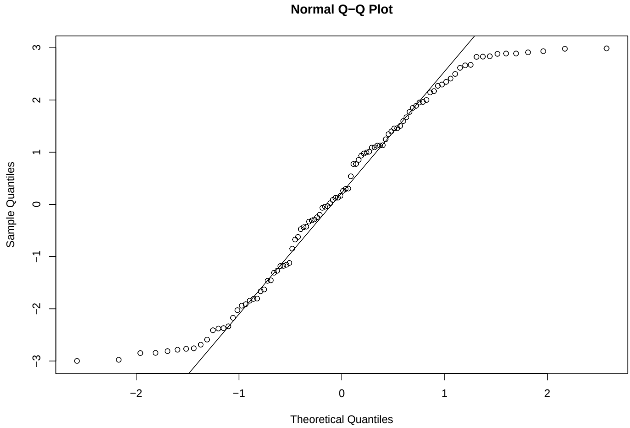
$$(q^F_{0.25}, q^x_{0.25}) \qquad and \qquad (q^F_{0.75}, q^x_{0.75}).$$

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

Some Experiments

Here are some experiments with `qqnorm`

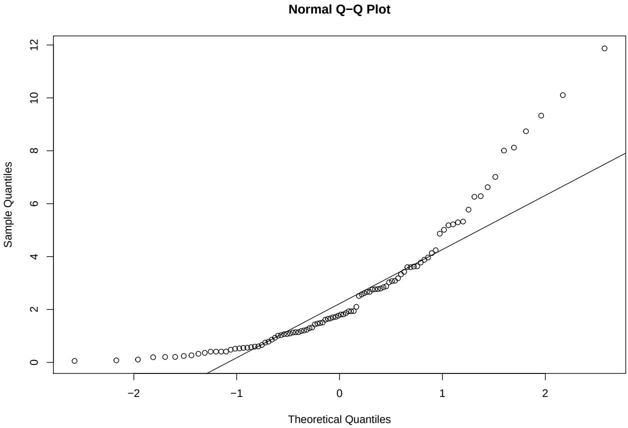
```
x <- runif(100,-3,3)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with `qqnorm`

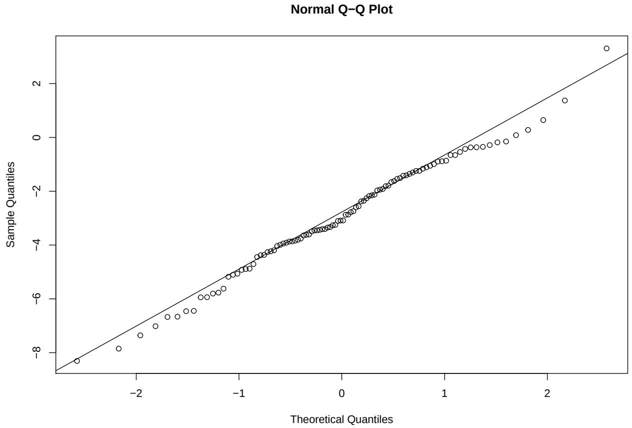
```
x <- rexp(100,0.4)
qqnorm(x)
qqline(x)
```



Some Experiments

Here are some experiments with qqnorm

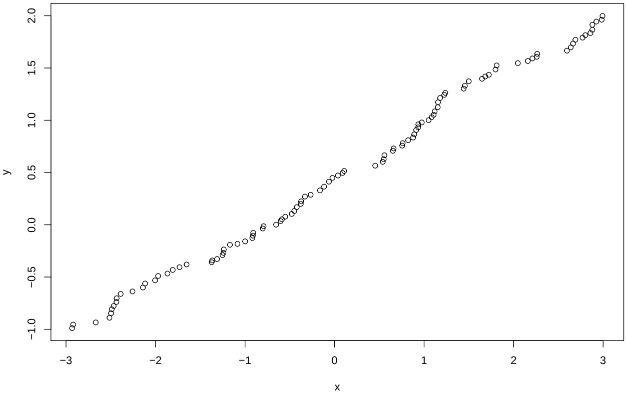
```
x <- rnorm(100, mean = -3, sd = 2)
qqnorm(x)
qqline(x)
```



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.

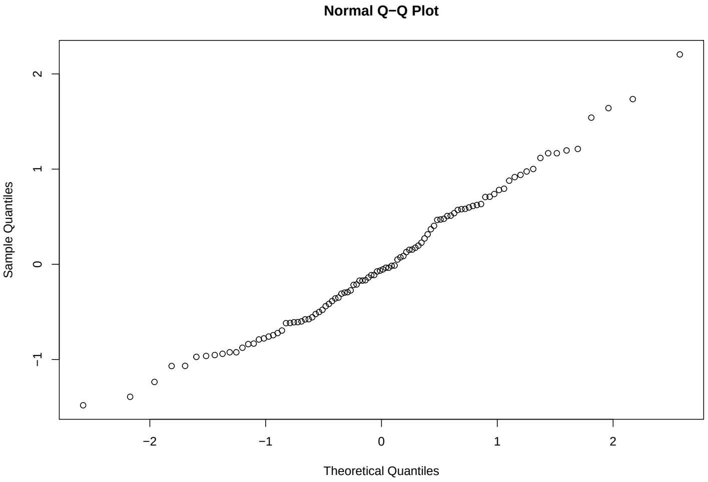
Important Note

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.

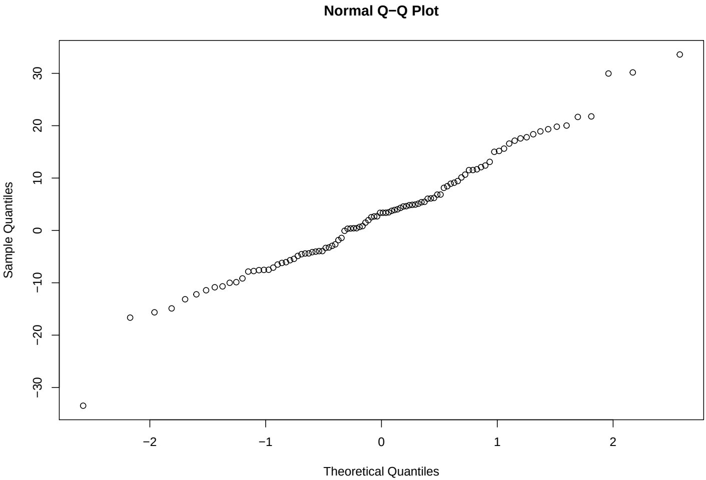
Some Experiments

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



Some Experiments

```
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.