WEEK 5

Exploratory Data Analysis for Bivariate Data: Quantiles and Q-Q Plots

Here we define the notion of a quantile, and give the Q-Q Plots ideas.

The idea of quantile is a generalization of the Median idea. The idea of the Sample Median was to give a number dividing the dataset into to parts, such that half of the data is to the left (or equal to) of that number, and half of the data - to the right (or equal to).

The idea of a Sample Quantile is a straightforward generalization of the Median idea: if we want to define the α-order quantile, or the α-quantile, for $\alpha \in (0,1)$, then we want to find a number that will divide our dataset into the proportions α and $1-\alpha$, i.e., we want to find a number such that $100\alpha\%$ of our datapoints will be to the left of that number (or equal to), and the rest, i.e., $100(1-\alpha)\%$ of the datapoints will be to the right (or equal to) of that number.

Analogously, quantiles can be defined also for theoretical distributions. Here, in this section, we define the quantiles (or percentiles) for a dataset and for a distribution, and then compare the quantiles using the Q-Q Plot.

5.1 Theoretical Quantiles, Quantiles for a Distribution

Assume we have a CDF $F(x)$ for some distribution.

Definition 5.1. *For $\alpha \in (0,1)$, the α-th quantile q_α of that distribution is defined by*

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geqslant \alpha\}. \tag{5.1}$$

If F is strictly increasing and continuous, then the α-th quantile q_α is defined to be the unique solution to

$$F(q_\alpha) = \alpha,$$

or, in terms of the PDF $f(x)$, we will have

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

In other words, if q_α is the α-th quantile for F, which is continuous and strictly increasing, and if X is a r.v. with CDF $F(x)$, then

$$\mathbb{P}(X \leqslant q_\alpha) = \alpha \qquad \text{and} \qquad \mathbb{P}(X \geqslant q_\alpha) = 1-\alpha.$$

So for the α-th quantile q_α, and for r.v. X with CDF $F(x)$, we will have that with probability α the values of X are to the left than or equal to q_α, and with the probability $1-\alpha$, the values of X are larger than q_α. Or, which is the same, the area under the PDF of that distribution in the region $(-\infty, q_\alpha]$ is equal to α, i.e., the line $x = q_\alpha$ divides the area under the graph of PDF into the portions α (left portion) and $1-\alpha$ (right portion), see Fig. 5.1.

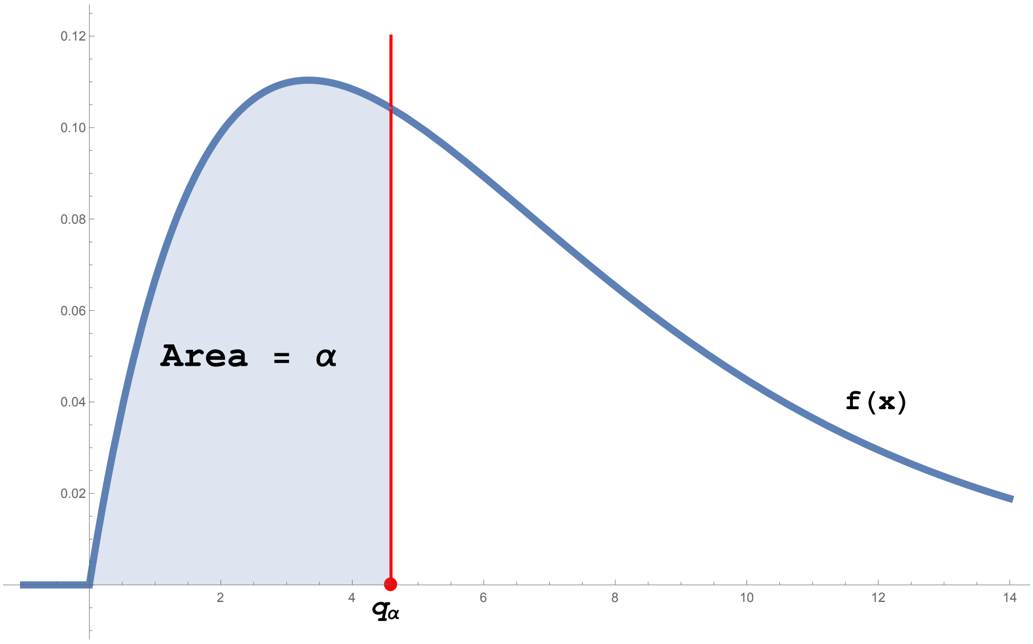


Fig. 5.1: Theoretical distribution's α-quantile, q_α, using the PDF graph. The area under the PDF left to the vertical line passing through the point q_α is exactly α

To explain the notion of the α-th quantile geometrically on the CDF graph - q_α is the leftmost point on the x-axis, where the graph of our CDF $F(x)$ crosses or jumps over[1] the value α, see Fig. 5.2.

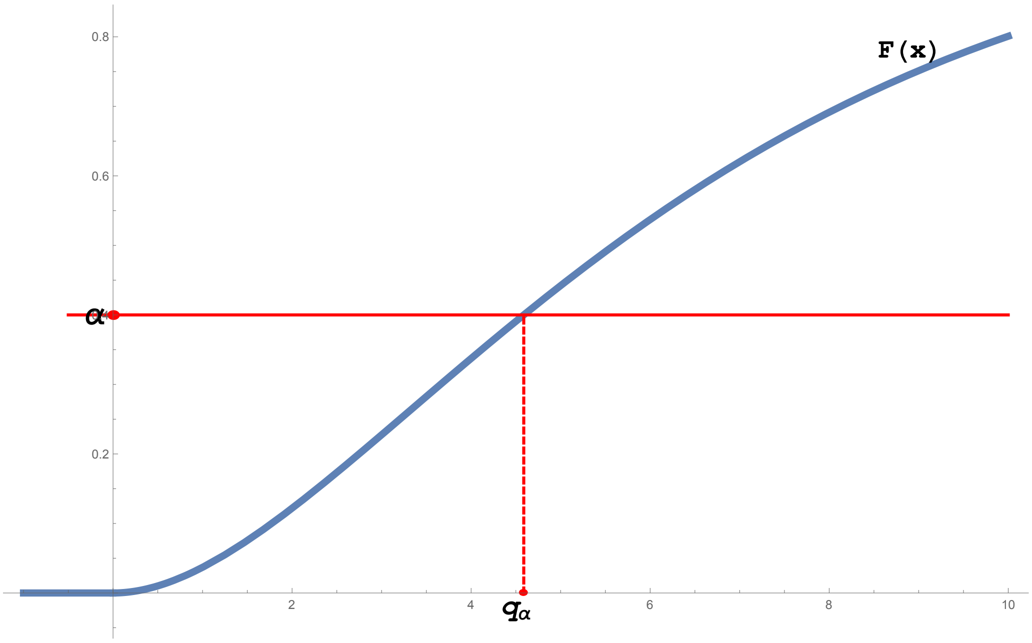


Fig. 5.2: Theoretical distribution's α-quantile, q_α, using the CDF graph. If we will consider the line $y = \alpha$ (solid red), then the (leftmost) intersection point with the CDF is $x = q_\alpha$

[1] For the continuous distribution case, it will cross for sure, but not for the discrete case, in general

Example, Quantiles for a Theoretical Distribution, Continuous case: Consider the distribution (called a Pareto Distribution, with some parameters) with the following PDF:

$$f(x) = \begin{cases} \dfrac{2}{x^3}, & x \geqslant 1 \\ 0, & x < 1 \end{cases}$$

It is easy to check that f is a PDF, i.e., $\int_{-\infty}^{\infty} f(x)dx = 1$, and $f(x) \geqslant 0$ for any $x \in \mathbb{R}$. The graph of f is given in Fig. 5.3.

Now, let us calculate the 40% quantile for this distribution, i.e., calculate $q_{0.4}$. To that end, we need to find a point on the x-axis such that the vertical line passing through that point will divide the total area under the PDF graph into the portions 0.4 (to the left to that line) and 0.6 (to the right). So we need to have that the area under the PDF $f(x)$ from $-\infty$ to the quantile $q_{0.4}$ needs to be equal to 0.4, i.e.

$$\text{Area under the PDF curve from } -\infty \text{ to } q_{0.4} = \int_{-\infty}^{q_{0.4}} f(x)dx = 0.4$$

In other words, we need to solve

$$\int_{-\infty}^{q_{0.4}} f(x)dx = 0.4$$

to find $q_{0.4}$. Of course, $q_{0.4}$ cannot be less than 1, otherwise the integral will give 0. So $q_{0.4} > 1$. In that case,

$$0.4 = \int_{-\infty}^{q_{0.4}} f(x)dx = \int_{1}^{q_{0.4}} f(x)dx = \int_{1}^{q_{0.4}} \frac{2}{x^3}dx \overset{\text{Show this!}}{=\!=\!=} 1 - \frac{1}{(q_{0.4})^2}.$$

This gives that $\dfrac{1}{(q_{0.4})^2} = 0.6$, hence, recalling that $q_{0.4} > 1$, we will get

$$q_{0.4} = \sqrt{\frac{5}{3}},$$

see Fig. 5.3.

Now, let us solve the general problem of calculation of all α-quantiles for any $\alpha \in (0,1)$. As above, we need to solve the equation

$$\text{Area under the PDF curve from } -\infty \text{ to } q_\alpha = \int_{-\infty}^{q_\alpha} f(x)dx = \alpha$$

to calculate the α-quantile q_α. So we need to solve

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

for q_α. Again, q_α cannot be less than 1, since our PDF is 0 for any $x < 1$. So $q_\alpha > 1$. Then,

$$\alpha = \int_{-\infty}^{q_\alpha} f(x)dx = \int_{1}^{q_\alpha} f(x)dx = \int_{1}^{q_\alpha} \frac{2}{x^3}dx = 1 - \frac{1}{(q_\alpha)^2}.$$

This gives $\dfrac{1}{(q_\alpha)^2} = 1 - \alpha$, implying that

$$q_\alpha = \pm\sqrt{\frac{1}{1-\alpha}}.$$

Using the condition $q_\alpha > 1$ (and hence, $q_\alpha > 0$), we will get that

$$q_\alpha = \sqrt{\frac{1}{1-\alpha}}.$$

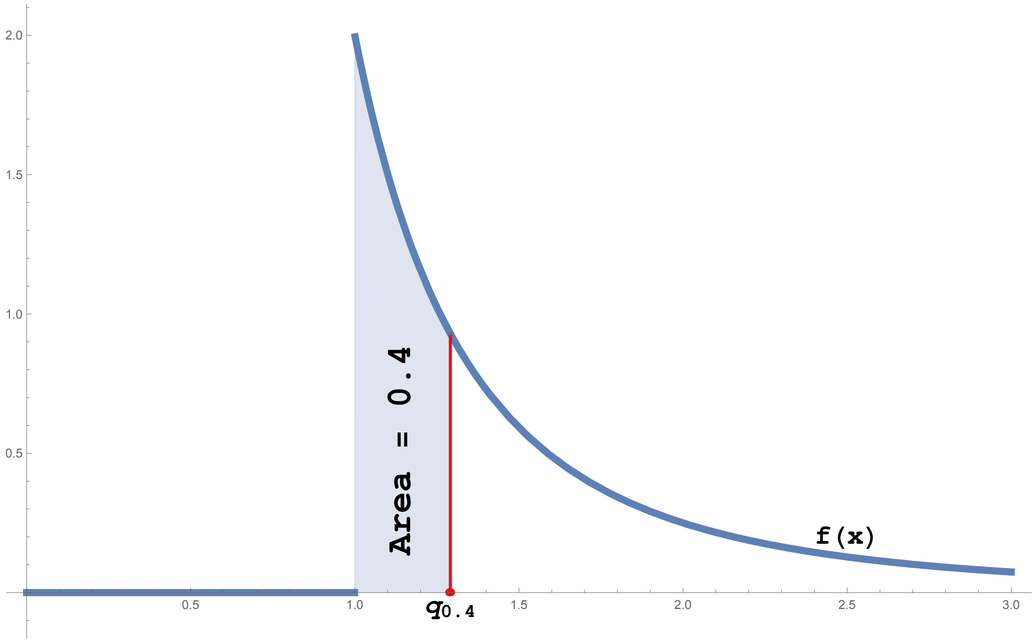


Fig. 5.3: Pareto Distribution PDF and its 40% quantile.

Example, Quantiles for a Theoretical Distribution, Continuous case: Here let us give another example of quantile calculation using the CDF. The aim of this example is to show why we use the infimum in the quantile definition (5.1).

Consider the distribution given by its CDF $F(x)$ as in Fig. 5.4, and let $\alpha = 0.6$, so we want to find the 60% quantile of this distribution. Clearly,

$$\{x \in \mathbb{R} : \ F(x) \geqslant \alpha\} = [2, +\infty),$$

and $F(q) = \alpha$ has infinitely many solutions, namely, any number $q \in [2, 3]$ will satisfy. So we cannot find a unique point q with $F(q) = \alpha$. Of course, any such point q will divide the range of our distribution into two pieces, and the probability that the values of a r.v. with this distribution will be less or equal to q will be α:

$$\mathbb{P}(X \leqslant q) = \alpha,$$

where X is a r.v. with CDF $F(x)$. So, in theory, this q can serve as a quantile. But usually people take the minimal such q (in fact, the infimum of such q-s). So in our case, $q_\alpha = 2$.

Example, Theoretical Quantiles, Discrete case:

Now let us consider the following r.v.:

X	1	3	5
$\mathbb{P}(X = x)$	0.3	0.5	0.2

We want to calculate the $\alpha = 0.6$-th quantile for the distribution of X.

The CDF of X, $F(x)$ is given in Fig. 5.5. From the graph it is clear that

$$\{x \in \mathbb{R} : F(x) \geqslant \alpha\} = \{x \in \mathbb{R} : F(x) \geqslant 0.6\} = [3, +\infty),$$

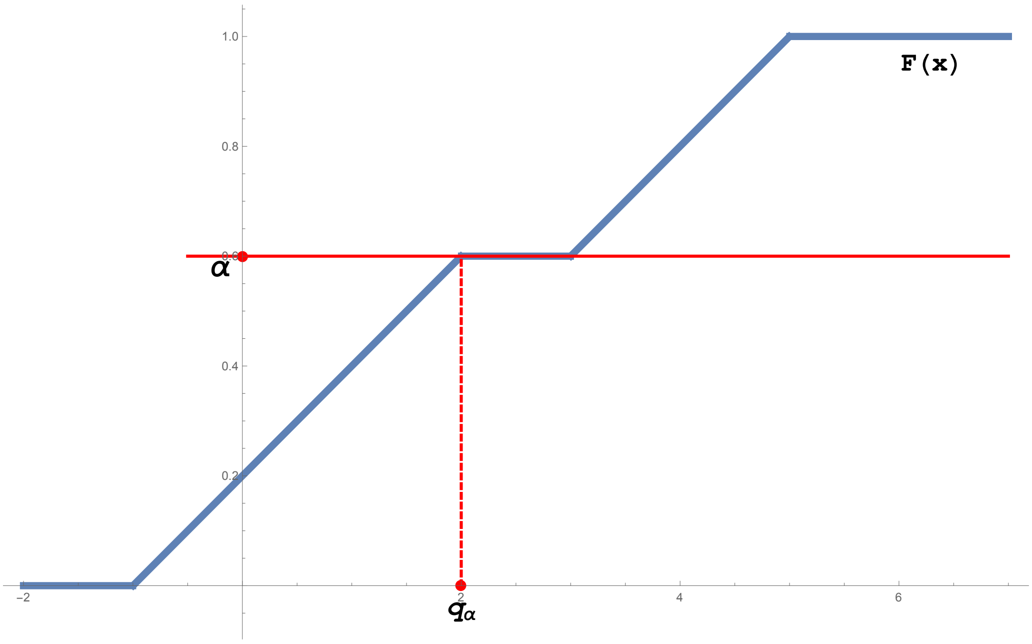


Fig. 5.4: CDF for some distribution and its $\alpha = 0.6$ quantile.

so $q_{0.6} = q_\alpha = 3$.

Similarly, for any $\alpha \in (0.3, 0.8]$, $q_\alpha = 3$. But, say, for $\alpha = 0.81$, $q_\alpha = 5$.

Note that for the Continuous Disctribution case, for any $\alpha \in (0,1)$, we will always have a $q \in \mathbb{R}$ (unique or not) with $F(q) = \alpha$. But in the Discrete Distribution case, not for all α we will have such q-s. Say, for our example, no q exists with $F(q) = 0.6$. That's why in the definition of the quantiles (5.1) we are not using $\inf\{x \in \mathbb{R} : F(x) = \alpha\}$, but $\inf\{x \in \mathbb{R} : F(x) \geqslant \alpha\}$, with inequality sign.

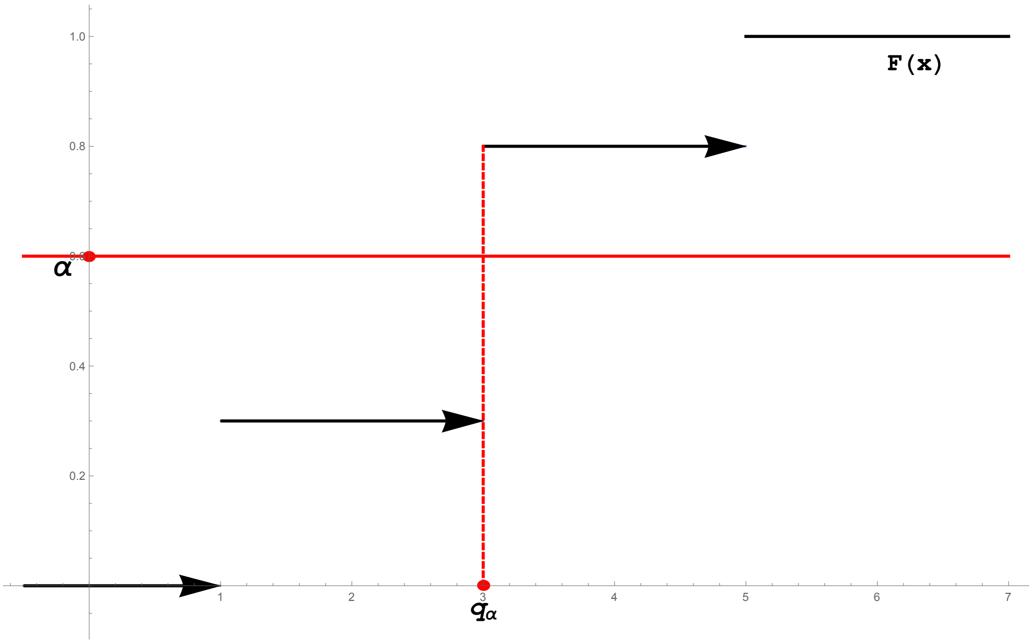


Fig. 5.5: CDF for a discrete distribution and its $\alpha = 0.6$ quantile.

R CODE, QUANITILES IN R: In **R**, any distribution has 4 standard functions ($[name]$ is the name of the distribution in **R**):

r[name] - generates random numbers from the distribution $[name]$. Say, $rnorm(100)$ will generate 100 random (pseudo-random) numbers from the Standard Normal distribution, and $rpois(100, lambda = 3)$ will generate 100 random numbers from $Pois(3)$ distribution.

p[name] - gives the CDF of the distribution $[name]$. For example, $punif(0.2)$ will give the value of the CDF at the point 0.2 for the Standard Uniform distribution, $Unif[0,1]$. Also, $punif(0.2, min = 2, max = 5)$ will give the value of the CDF at the point 0.2 for the Uniform distribution $Unif[2,5]$ (can you guess the value?).

d[name] - gives the PDF of the distribution $[name]$. For example, $curve(dexp, 0, 4)$ will draw the $Exp(1)$ distribution's PDF in $[0,4]$, and $dexp(1, rate = 2)$ will return the value of the $Exp(2)$ distribution's PDF at $x = 1$.

q[name] - gives the quantiles of the distribution $[name]$. For example, $qcauchy(0.3)$ will give the 30% quantile for the Cauchy Distribution.

Now, the command

```
qnorm(0.2)
```

will return -0.8416212, so the 20% quantile of the Standard Normal Distribution $\mathcal{N}(0,1)$ is

$$q_{0.2}^{\mathcal{N}} = -0.8416212,$$

(well, after doing some rounding, since the actual number will have infinitely many digits after the period - you can find more digits in the "Environment" tab in RStudio). This means that for a r.v. $X \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(X \leqslant -0.8416212) = 0.2$$

We can check this in **R** by running the command

```
pnorm(-0.8416212)
```

or, we can even run

```
pnorm(qnorm(0.2))
```

To show the result graphically, we can run the following code:

```
#Quantile, Geometrically, on the CDF
alpha <- 0.2 #alpha = 0.2
qalpha <- qnorm(alpha) #20% quantile for the Standard Normal Distrib
plot(pnorm, xlim = c(-5,5), lwd = 2)  #The graph of the Standard Normal Distrib's CDF
abline(h = alpha, xlim = c(-5,5), lwd = 2, col="red") #horizontal line y=alpha
abline(v = qalpha, lwd = 2, lty = 2, col = "red") #vertical line through the quantile
qalpha # the value of the quantile
```

We can do similar thing with the PDF:

```
#Quantile, Geometrically, on the PDF
alpha <- 0.2 #alpha = 0.2
qalpha <- qnorm(alpha) #20% quantile for the Standard Normal Distrib
plot(dnorm, xlim = c(-5,5), lwd = 2)  #The graph of the Standard Normal Distrib's CDF
abline(v = qalpha, lwd = 2, lty = 2, col = "red") #vertical line through the quantile
integrate(dnorm,-Inf, qalpha) #The area under the PDF left to the quantile,
                              #the integral of the PDF over (-Infinity, qalpha]
qalpha # the value of the quantile
```

R code, Quantiles in R: Let us calculate the quantiles of orders $0.1, 0.25, 0.5, 0.75, 0.9$ for the distribution $\mathcal{N}(-2, 5^2)$:

```
#Quantiles for N(-2,5^2)
alpha <- c(0.1, 0.25, 0.5, 0.75, 0.9) #the vector of quantile orders
qnorm(alpha, mean = -2, sd = 5)
```

The result is

```
[1] -8.407758 -5.372449 -2.000000  1.372449  4.407758
```

Here the idea is that you can calculate several quantiles for the same distribution simultaneously, by passing the vector of the quantile orders to the quantile function of that distribution.

Another example,

```
#Quantiles for Exp(1)
qexp(c(0.2, 0.5, 0.7))
```

will return the quantiles of orders $0.2, 0.5$ and 0.7 for the distribution $\mathtt{Exp}(1)$, and the result will be

```
[1] 0.2231436 0.6931472 1.2039728
```

5.2 Quantiles for a Dataset

Now, if we have a dataset x, then the α-th quantile is the number for which approximately $100 \cdot \alpha\%$ of data is below that number, and the rest are above it: say, if $\alpha = 0.3$, then the 0.3-quantile is the "point" below which we will have 30% of our observations[2], and above which will be 70% of all observations.

Now, to define for any $\alpha \in (0, 1)$ the α-th quantile of a dataset (or the order α quantile), we will use the following definition[3].

Definition 5.2. *For a dataset x and $\alpha \in (0, 1)$, the quantile of order α is defined by*

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}. \tag{5.2}$$

[2] Approximately 30%, since for 7 point dataset, we need to have 2.1 points are to the left or equal to the quantile $q_{0.3}$, but what it means "2.1 points"?

[3] Please note that there are different definitions of a sample quantile, and they give slightly different values. For example, you can read the help file of R package to find the description of 9 types of quantiles[4]

Example, Dataset quantiles:

... Here you can put your ad ...

Remark, Quantile Definition: Another possible definition is (from Wesserman's book):

Definition 5.3. *For a dataset x and $\alpha \in (0,1)$, the quantile of order α is defined by*

$$q_\alpha = q_\alpha^x = \inf\{x : ECDF(x) \geqslant \alpha\}.$$

Here the idea is one of the standard methods in Statistics: we have the definition of the theoretical quantile given in (5.1):

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

Now, for a dataset x, instead of theoretical CDF $F(x)$ we take the Empirical CDF $ECDF(x)$! Nice and clear! If you will have some theoretical notion defined in terms of CDF, for a dataset replace CDF by ECDF, and that's it $\because\smile$

Btw, is this definition giving the same values for quantiles as our one?

Remark, Quantiles and Quartiles: Using our quantile definition above, one will not obtain that $q_{\frac{1}{2}}$ is always the Median, or $q_{1/4}$ and $q_{3/4}$ are the quartiles Q_1 and Q_3, respectively, for any datasets. This is because, say, by our definition, quantiles are always datapoints, are from our dataset, but Median or Quartiles can be midpoints of datapoints, not elements from our dataset.

We have that Median divides our sorted list of observations into two equal-length parts, so we could define the 0.5-quantile to be our Median. And we could take as $\frac{1}{4}$ and $\frac{3}{4}$-quantiles our first and third quartiles Q_1 and Q_3, as we have talked before that the 25% of observations are to the left of Q_1 and the rest are to the right of Q_1. And in some textbooks this is the case. But here, for the sake of simple formula for quantiles, we will use the definition (5.2) above, which can produce the described effect where 0.25, 0.5 and 0.75 quantiles are not the quartiles.

Hopefully, this will not cause much problems.

5.3 Quantile-Quantile (Q-Q) Plots

One of the standard and important problems of the Statistics is to check if the given data comes from some fixed distribution, say, from the Normal Distribution[5]. Or, another problem is to check if two datasets are generated from the same distribution.

In our course, we will consider two methods to check this kind of things. The first one is the graphical one, where we will get the answer visually. And the second method will be to do some test to check that - and we will talk about this later, when considering hypotheses testing topics.

Here, in this section we will describe one of the non-parametric ways to solve the described problems - the graphical method, Q-Q Plot method.

We will consider the following problems:

[5] Later, in the Inferential Statistics part of our course, we will deal with the Parametric Statistics. That is, we will assume that our data comes from some parametric family of distributions, and our aim will be to estimate that parameters. But where from we can guess that our data comes from that parametric family? - Here the Q-Q Plot method can be used!

Problem 1: We have 2 theoretical distributions (say, $\mathcal{N}(3,4)$ and $\mathtt{Exp}(2)$). We want to compare these distributions to see which one has fatter tails. Another problem is to see if the distributions are close to each other (say, $t(50)$ is close to $\mathcal{N}(0,1)$).

Problem 2: We have a theoretical distribution and a dataset. We want to see if the dataset is generated from the theoretical distribution.

Problem 3: We have two datasets, possibly, of different sizes. We want to see if the datasets are generated from the same distribution.

To approach the above problems, we will use the Q-Q Plot graphical method.

5.3.1 Q-Q Plot for two Distributions

Assume we are given two distributions, by their CDF-s $F(x)$ and $G(x)$.

Definition 5.4. *The **Q-Q Plot** for $F(x)$ and $G(x)$ is the plot of all points (q^F_α, q^G_α), where $\alpha \in (0,1)$, and q^F_α and q^G_α are the α-th quantiles of F and G, respectively.*

It is clear that in the case when F coincides with G, $F \equiv G$, the quantiles will be the same, so all points (q^F_α, q^G_α), $\alpha \in (0,1)$ will give the portion of the bisector $y = x$ on the graph.

Let us see what will happen if we will do the Q-Q Plot for the same family distributions.

Example, Q-Q Plot for Uniform vs Uniform: Let us do the Q-Q Plot for $\mathcal{A} = \mathtt{Unif}[0,1]$ and $\mathcal{B} = \mathtt{Unif}[0,5]$.

So we fix $\alpha \in (0,1)$, and calculate $q^{\mathcal{A}}_\alpha$ and $q^{\mathcal{B}}_\alpha$.

The quantile of order α of the $\mathcal{A} = \mathtt{Unif}[0,1]$ is the point that divides the area under the PDF curve into α and $1-\alpha$ portions (to the left and right, respectively). The PDF of the $\mathtt{Unif}[0,1]$ is the function $f_{\mathcal{A}}(x) = 1$ for $x \in [0,1]$ and $f_{\mathcal{A}}(x) = 0$ for $x \notin [0,1]$. So the α quantile will be $q^{\mathcal{A}}_\alpha = \alpha$. This can be seen also geometrically, using the graph of the PDF.

Another way to see this is to use the CDF of the $\mathtt{Unif}[0,1]$: the CDF has the form

$$F_{\mathcal{A}}(x) = \begin{cases} 0, & x < 0 \\ x, & x \in [0,1] \\ 1, & x > 1. \end{cases}$$

Now, the quantile of order α is the leftmost point q with $F_{\mathcal{A}}(q) = \alpha$, i.e., the intersection point of the line $y = \alpha$ and the graph of CDF $y = F_{\mathcal{A}}(x)$. Obviously, the intersection point will be unique: $q = \alpha$, since $F_{\mathcal{A}}(\alpha) = \alpha$ (as $F_{\mathcal{A}}(x) = x$ for $x \in [0,1]$).

Summarizing, $q^{\mathcal{A}}_\alpha = \alpha$.

Now, let us calculate the α quantile for the distribution $\mathcal{B} = \mathtt{Unif}[0,5]$.

Again we will use geometric ideas. We first find the PDF of $\mathtt{Unif}[0,5]$, which is

$$f_{\mathcal{B}}(x) = \begin{cases} \dfrac{1}{5-0} = \dfrac{1}{5}, & x \in [0,5] \\ 0, & x \notin [0,5]. \end{cases}$$

We want to find the point q such that the area left to the line $x = q$ under the PDF curve will be α. Clearly, $q \in [0,5]$ (think why?). Then the area to the left to $x = q$ will be

$$\text{Area} = \text{height} \times \text{width} = \frac{1}{5} \cdot (q - 0) = \alpha,$$

so $q = 5\alpha$. Hence, $q_\alpha^{\mathcal{B}} = 5\alpha$.

This means that for the Q-Q Plot we need to draw the points $(q_\alpha^{\mathcal{A}}, q_\alpha^{\mathcal{B}}) = (\alpha, 5\alpha)$ for $\alpha \in (0,1)$, which will give the line $y = 5x$ on the graph, see Fig. 5.6.

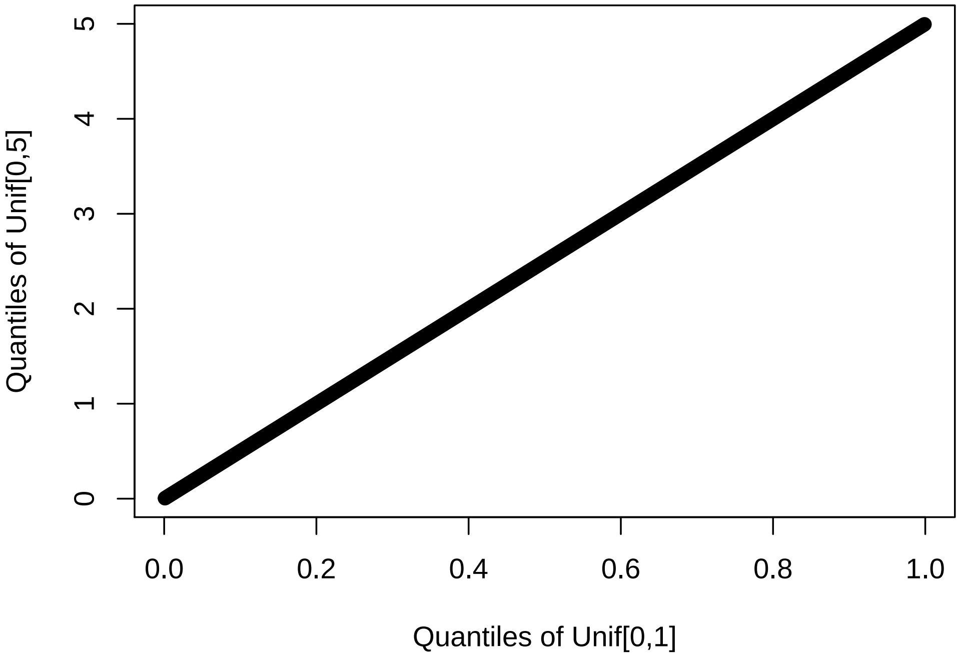


Fig. 5.6: Q-Q Plot for $\mathrm{Unif}[0,5]$ vs $\mathrm{Unif}[0,1]$

R code, Q-Q Plot for the $\mathrm{Unif}[0,5]$ vs $\mathrm{Unif}[0,1]$:

```
#Q-Q plot, Theoretical vs Theoretical
alpha <- seq(from = 0.001, to = 0.999, by = 0.001)
              #alpha is running from 0.001 to 0.999 with stepsize 0.001
x <- qunif(alpha, min = 0, max = 1)
              #quantiles of orders alpha for the above alpha-s for Unif[0,1]
y <- qunif(alpha, min = 0, max = 5)
              #quantiles of orders alpha for the above alpha-s for Unif[0,5]
plot(x,y, pch = 19, xlab = "Quantiles of Unif[0,1]", ylab = "Quantiles of Unif[0,5]")
```

Example, Q-Q Plot for Uniform vs Uniform, v2: Now, let us do the Q-Q Plot for the $\mathcal{A} = \mathrm{Unif}[1,4]$ and $\mathcal{B} = \mathrm{Unif}[3,9]$.

We again take $\alpha \in (0,1)$. Now, one can easily check (do the calculations!), using the ideas above, that

$$q_\alpha^{\mathcal{A}} = 1 + 3\alpha \qquad \text{and} \qquad q_\alpha^{\mathcal{B}} = 3 + 6\alpha.$$

Now, the Q-Q Plot will consists of all points

$$(q_\alpha^{\mathcal{A}}, q_\alpha^{\mathcal{B}}) = (1+3\alpha, 3+6\alpha), \qquad \alpha \in (0,1).$$

To describe this parametric graph, let us denote by $x = 1+3\alpha$ and $y = 3+6\alpha$. Then $\alpha = \frac{1}{3}\cdot(x-1)$, so $y = 3+6\alpha = 3+6\cdot\frac{1}{3}\cdot(x-1) = 2x+1$. So the graph will be some portion of the line (can you guess where the coefficients 2 and 1 come from in the line $y = 2x+1$? Note that $4-1=3$ and $9-3=6$, so $9-3 = 2\cdot(4-1)$ $\ddot\smile$)

$$y = 2x+1,$$

for $x \in (1,4)$ (since $\alpha \in (0,1)$ and $x = 1+3\alpha$, then $x \in (1,4)$). See Fig. 5.7.

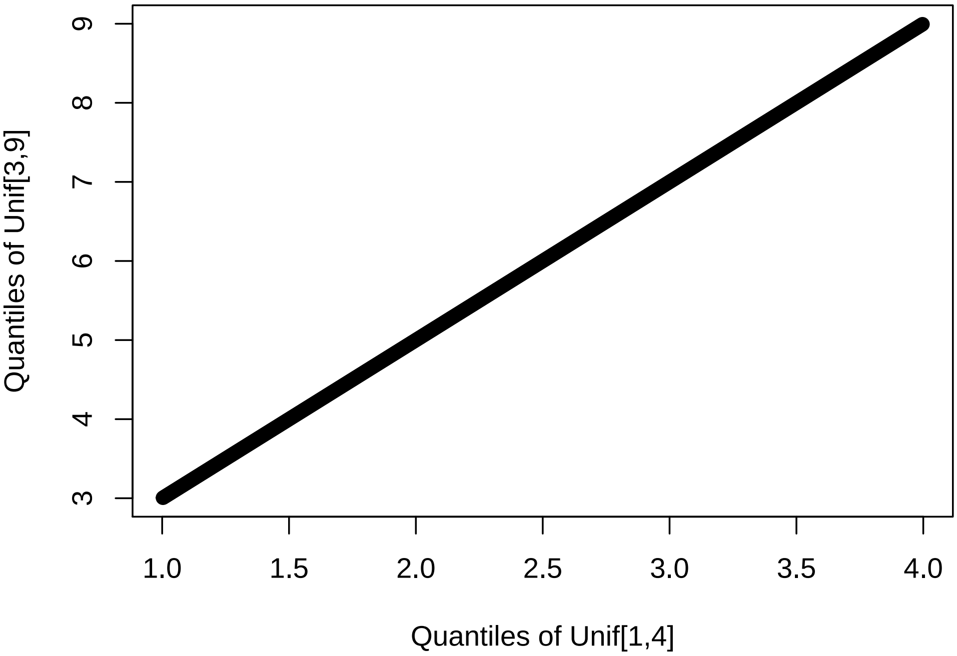


Fig. 5.7: Q-Q Plot for Unif[1,4] vs Unif[3,9]

R code, Q-Q Plot, for Unif[1,4] vs Unif[3,9]:

```
#Q-Q plot, Theoretical vs Theoretical
alpha <- seq(from = 0.001, to = 0.999, by = 0.001)
    #alpha is running from 0.001 to 0.999 with stepsize 0.001
x <- qunif(alpha, min = 1, max = 4)
    #quantiles of orders alpha for the above alpha-s for Unif[0,1]
```

```
y <- qunif(alpha, min = 3, max = 9)
    #quantiles of orders alpha for the above alpha-s for Unif[0,5]
plot(x,y, pch = 19, xlab = "Quantiles of Unif[1,4]", ylab = "Quantiles of Unif[3,9]")
```

Now, to give some idea about the interpretation of the Q-Q Plot, let us consider the example of Standard Normal Distribution $\mathcal{N}(0,1)$ and Cauchy Distribution[6] with the parameters $(0,1)$, $Caucy(0,1)$. First we draw the PDFs of these distribution on the same figure, see Fig. 5.8. The code is:

R code, PDFs of $\mathcal{N}(0,1)$ and $Cauchy(0,1)$:

```
# PDFs for N(0,1) and Cauchy(0,1)
curve(dcauchy, xlim = c(-5,5), ylim = c(0,0.4), lwd = 2, col = "red",
      ylab = "PDFs of N(0,1) and Cauchy(0,1)")
par(new = TRUE)
curve(dnorm, xlim = c(-5,5),  ylim = c(0,0.4), lwd = 2, col = "blue",
      ylab = "PDFs of N(0,1) and Cauchy(0,1)")
legend(1.8, 0.38, c("N(0,1)", "Cauchy(0,1)"), lty = c(2,2), lwd = c(2,2),
       col = c("blue", "red"))
```

Clearly,

- Both distributions are symmetric around 0;
- The Cauchy Distribution PDF has fatter tails, meaning that it tends to 0 as $x \to \pm\infty$ much slower that the PDF of Standard Normal[7]

Now, let us calculate the $0.5, 0.7, 0.8$ and 0.9 quantiles for both distributions and plot the pairs of quantiles $(q_\alpha^{\mathcal{N}}, q_\alpha^{C})$ for $\alpha = 0.5, 0.7, 0.8, 0.9$. The **R** code is here:

R code, Quantiles for $\mathcal{N}(0,1)$ vs $Cauchy(0,1)$:

```
alpha = c(0.5, 0.7, 0.8, 0.9)
xx <- qnorm(alpha)
yy <- qcauchy(alpha)
xx
yy
plot(xx,yy, pch = 19, cex = 1.2, xlim = c(0,3.5), ylim = c(0,3.5),
    xlab = "Quantiles of N(0,1)", ylab = "Quantiles of C(0,1)")
abline(0,1, col = "green", lwd = 2)
```

The last command adds the line $y = x$ (in green) to the graph. The result is in Fig. 5.9.

[6] See https://en.wikipedia.org/wiki/Cauchy_distribution

[7] In fact, Cauchy distribution's PDF is given by $f_{Cauchy}(x) = \dfrac{1}{\pi(1+x^2)}$, which tends to 0 like $\dfrac{1}{x^2}$, as $x \to \pm\infty$, and the Standard Normal Distribution's PDF is given by $f_{StdNormal}(x) = \dfrac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$, which tends to 0 exponentially fast.

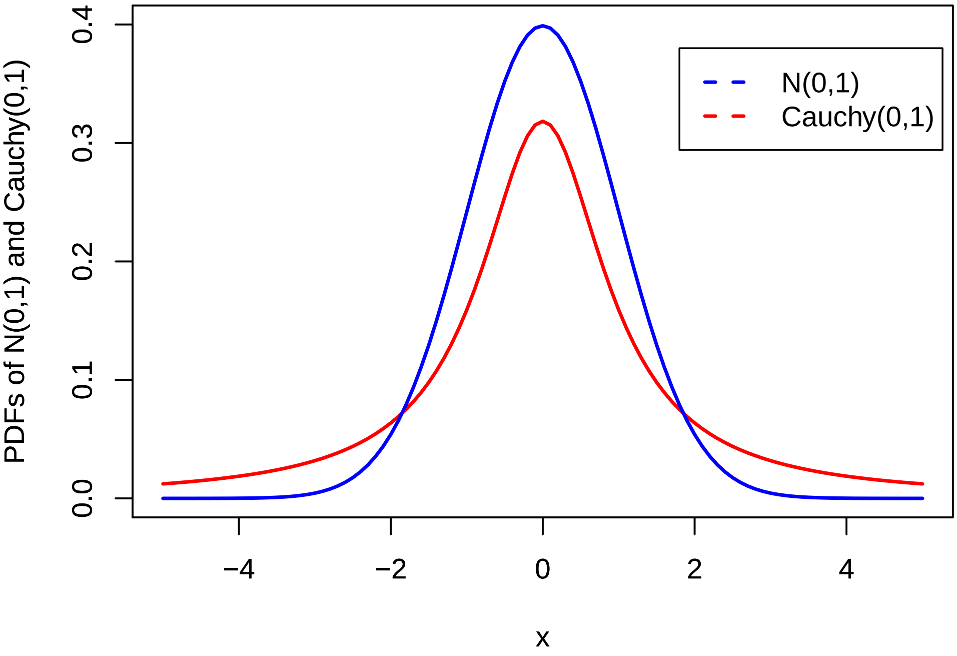


Fig. 5.8: PDFs of the $\mathcal{N}(0,1)$ (blue) and $\mathtt{Cauchy}(0,1)$ (red)

You can see that $q^{\mathcal{N}}_{0.5} = q^{C}_{0.5} = 0$, $q^{\mathcal{N}}_{0.7} < q^{C}_{0.7}$, $q^{\mathcal{N}}_{0.8} = q^{C}_{0.8}$ and $q^{\mathcal{N}}_{0.9} = q^{C}_{0.9}$, and, in fact, $q^{\mathcal{N}}_{\alpha}$ grows much slower that q^{C}_{α} (can you explain what this statement means?). This means that the shape of the Q-Q Plot on the left will be *convex*. To explain why the quantiles of the Standard Normal grow much slower than the ones for Cauchy, we draw the CDF's and corresponding α levels, see Fig. 5.10. On the Fig. 5.10, the quantiles are the x-coordinates of the corresponding line $y = \alpha$ with the CDFs. Visually, when α increases, the quantile of the Cauchy Distribution (the intersection point of the line $y = \alpha$ with the red curve) grows faster that the quantile of the Standard Normal Distribution.

This phenomenon is specific for fatter-tailed distributions: if we draw the Q-Q Plot for two distributions $\mathcal{A}$ and $\mathcal{B}$, both distributions have right tails (are non-zero in $[a, +\infty)$ for some a), and the tails of $\mathcal{A}$ are thinner than the tails of $\mathcal{B}$, then the right-hand side of the Q-Q Plot (assuming that the quantiles of $\mathcal{A}$ are on the x-axis) will be convex-shaped. The inverse is true for the left-tailed distributions: if the left tail of $\mathcal{B}$ is fatter than the left tail of $\mathcal{A}$, then on the left-hand side of the Q-Q Plot (again assuming that the quantiles of $\mathcal{A}$ are on the x-axis) we will have a concave-shaped graph.

Example, Comparison of distribution tails with Q-Q Plot: The Fig. 5.11 shows the Q-Q Plot for $\mathcal{N}(0,1)$ (quantiles are on the x-axis) and $\mathtt{Cauchy}(0,1)$ (on the y-axis). $\mathtt{Caucy}(0,1)$ has fatter tails on the left and right hand sides, so the Q-Q Plot shape is convex on the right-hand side, and concave on the left one. Also, you can clearly see the symmetry of quantiles (because of the symmetry of distributions).

Another example is in Fig. 5.12, where the Q-Q plot of $\mathcal{N}(0,1)$ vs $\mathtt{Exp}(1)$ is given. Here again, $\mathtt{Exp}(1)$ has fatter tails on the right hand side, hence the convex shape on the right. $\mathtt{Exp}(1)$ does not have left tail (the PDF is 0, if $x < 0$), so the quantiles of $\mathtt{Exp}(1)$ will approach $0+$, as $\alpha \to 0$, and the

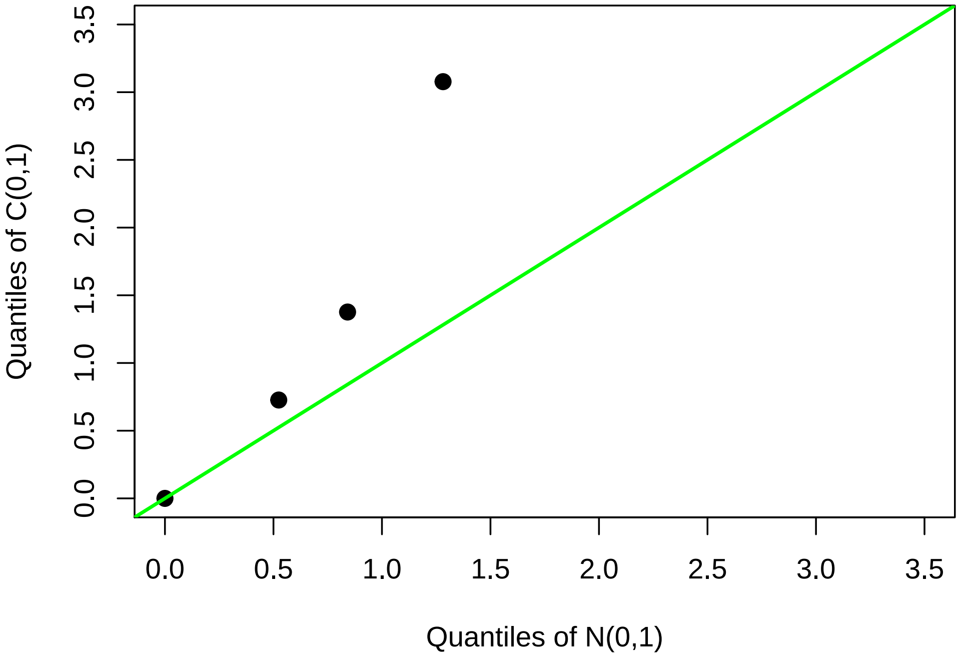


Fig. 5.9: $0.5, 0.7, 0.8$ and 0.9 quantiles of $\mathcal{N}(0,1)$ vs $\mathrm{Cauchy}(0,1)$

quantiles of $\mathcal{N}(0,1)$ will tend to $-\infty$, when $\alpha \downarrow 0$ (can you explain this?).

Yet another example is in Fig. 5.13, where the Q-Q plot of $\mathrm{Exp}(1)$ vs $\mathrm{LogNormal}(0,1)$ is given (see `https://en.wikipedia.org/wiki/Log-normal_distribution` for the definition and properties of the LogNormal distribution). Here, LogNormal distribution has fatter tails compared to the Exponential.

R code, Q-Q Plot of $\mathcal{N}(0,1)$ vs $\mathrm{Cauchy}(0,1)$:

```
# Q-Q Plot for N(0,1) and C(0,1), alpha runs from 0.01 to 0.99
alpha <- seq(from = 0.01, to = 0.99, by = 0.001)
xx <- qnorm(alpha)
yy <- qcauchy(alpha)
plot(xx,yy, type = "l", lwd = 3, xlab = "Quantiles of N(0,1)",
     ylab = "Quantiles of Cauchy(0,1)")
par(new = TRUE)
abline(0,1, col = "green", lwd = 3)
```

R code, Q-Q Plot of $\mathcal{N}(0,1)$ vs $\mathrm{Cauchy}(0,1)$:

```
# Q-Q Plot for N(0,1) and Exp(1), alpha runs from 0.001 to 0.999
```

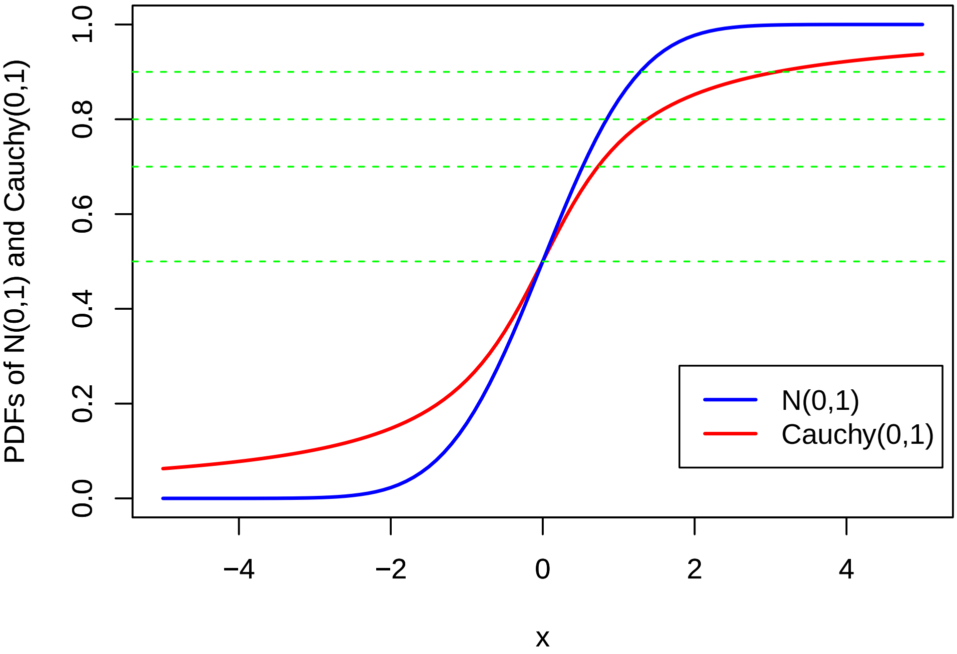


Fig. 5.10: The CDFs of $\mathcal{N}(0,1)$ vs $\mathrm{Cauchy}(0,1)$, and the lines $\alpha = 0.5, 0.7, 0.8$ and 0.9

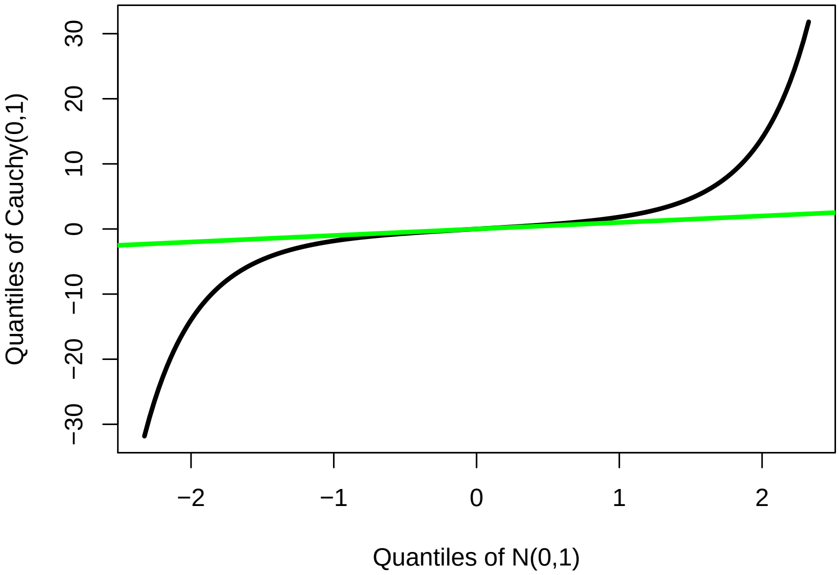


Fig. 5.11: Q-Q Plot of $\mathcal{N}(0,1)$ vs $\mathrm{Cauchy}(0,1)$

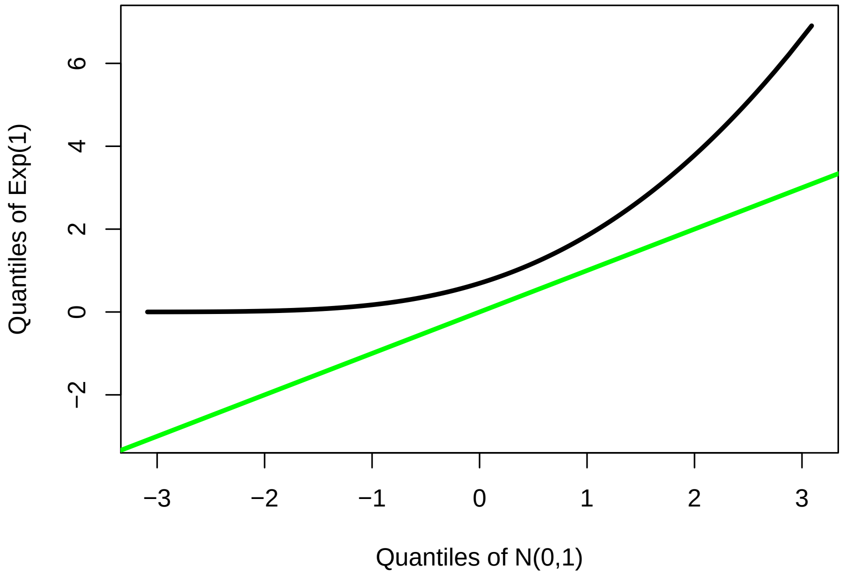


Fig. 5.12: Q-Q Plot of $\mathcal{N}(0,1)$ vs $\mathrm{Exp}(1)$

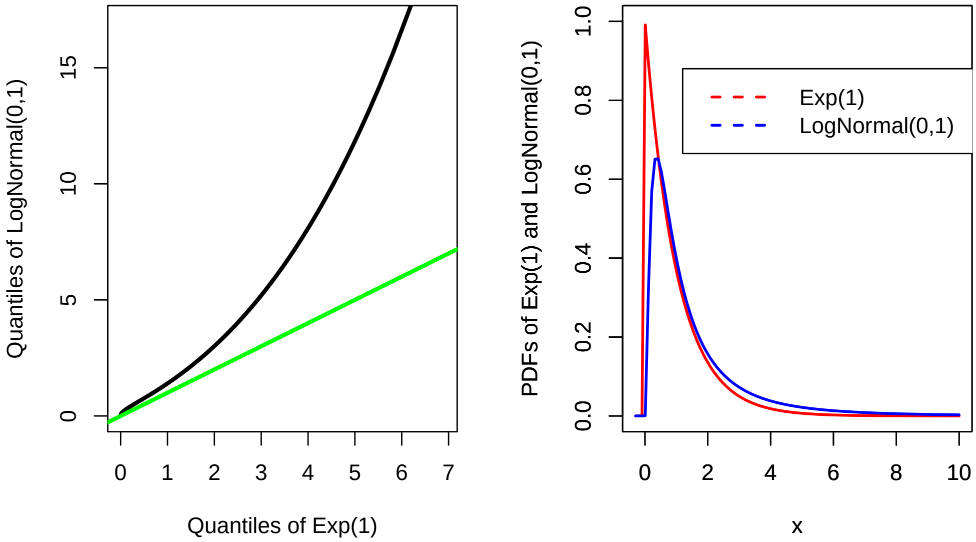


Fig. 5.13: Q-Q and PDF Plots of $\mathrm{Exp}(1)$ vs $\mathrm{LogNormal}(0,1)$

```
alpha <- seq(from = 0.001, to = 0.999, by = 0.001)
xx <- qnorm(alpha)
yy <- qexp(alpha)
plot(xx,yy, type = "l", lwd = 3, ylim = c(-3,7), xlab = "Quantiles of N(0,1)",
     ylab = "Quantiles of Exp(1)")
abline(0,1, col = "green", lwd = 3)
```

R code, Q-Q Plot of $\mathcal{N}(0,1)$ vs Cauchy$(0,1)$:

```
# Q-Q Plot for Exp(1) and LogNormal(0,1)
alpha <- seq(from = 0.001, to = 0.999, by = 0.001)
xx <- qexp(alpha)
yy <- qlnorm(alpha)
par(mfrow = c(1,2))
plot(xx,yy, type = "l", lwd = 3, ylim = c(0,17), xlab = "Quantiles of Exp(1)",
     ylab = "Quantiles of LogNormal(0,1)")
abline(0,1, col = "green", lwd = 3)
curve(dexp, xlim = c(-0.3,10), ylim = c(0,1), lwd = 2, col = "red",
      ylab = "PDFs of Exp(1) and LogNormal(0,1)")
par(new = TRUE)
curve(dlnorm, xlim = c(-0.3,10),  ylim = c(0,1), lwd = 2, col = "blue",
      ylab = "PDFs of Exp(1) and LogNormal(0,1)")
legend(1.2, 0.88, c("Exp(1)", "LogNormal(0,1)"), lty = c(2,2), lwd = c(2,2),
       col = c("red", "blue"))
```

Remark, Q-Q Plot:

- It is easy to see that the Q-Q Plot has the shape of a graph of an increasing function;
- Unfortunately, on the Q-Q Plot, one cannot identify some specific quantile, say, having the Q-Q Plot, one cannot find the 20% quantile for the distributions, or, say, the Medians. Unfortunately, we are not showing α on the graph

Example, t-distribution and Standard Normal Distribution: Student's t Distribution is one of the important distributions in Statistics. We will define and meet this distribution a lot of times in the rest of our Stat course. t distribution comes with a parameter called the degrees of freedom, $t(n)$ is the t distribution with n degrees of freedom[8] Here we want to compare $t(n)$ with $\mathcal{N}(0,1)$. The idea is that for large n, $t(n)$ is very close to $\mathcal{N}(0,1)$, for example, as a rule of thumb (that you will find in many Stat textbooks), if $n \geqslant 30$, then one is using $\mathcal{N}(0,1)$ as an approximation of $t(n)$.

Now, let us give the Q-Q Plots for $t(n)$ vs $\mathcal{N}(0,1)$ for different n-s. You can find the plots in Fig. 5.14-5.19.

5.3.2 Q-Q Plot for a Dataset vs Distribution

Assume here that we have a dataset x, and a fixed distribution given by its CDF $F(x)$. Our task is to check if the dataset is coming from the distribution defined by F or not. To check this graphically, we are using the Q-Q Plot defined below:

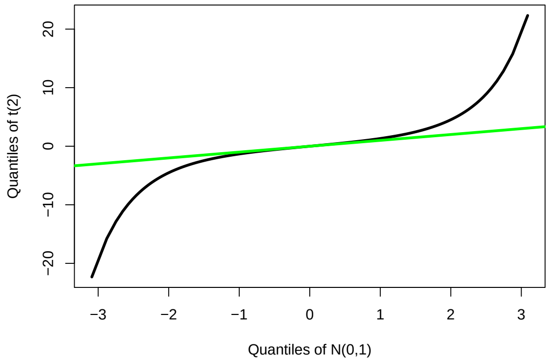


Fig. 5.14: Q-Q Plot: $t(2)$ vs $\mathcal{N}(0,1)$

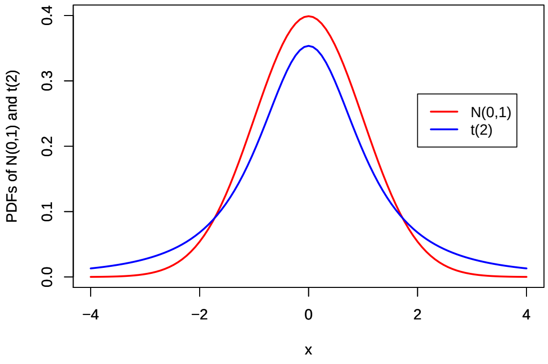


Fig. 5.15: PDF Plot: Plot: $t(2)$ vs $\mathcal{N}(0,1)$

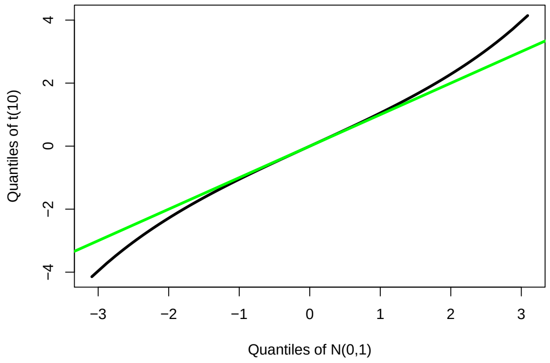


Fig. 5.16: Q-Q Plot: $t(10)$ vs $\mathcal{N}(0,1)$

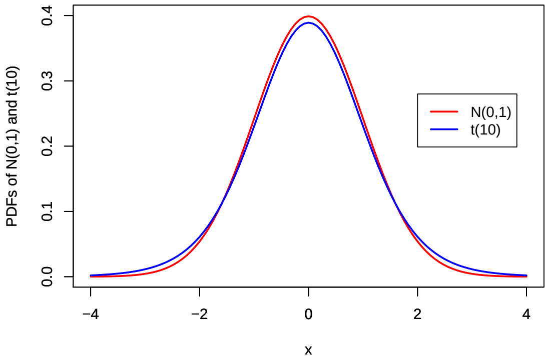


Fig. 5.17: PDF Plot: Plot: $t(10)$ vs $\mathcal{N}(0,1)$

Definition 5.5. *The* ***Q-Q Plot*** *for the dataset x and distribution $F(x)$ is the plot of all points (q_α^F, q_α^x), where α runs over some values in $(0,1)$, and q_α^F and q_α^x are the α-th quantiles of the distribution F and the dataset x, respectively.*

Usually, one uses $\alpha = \frac{k}{n}$-th quantiles, or $\alpha = \frac{k}{n+1}$-th $(k = 1, ..., n)$ quantiles[9] or $\alpha = \frac{k-0.5}{n}$-th quantiles, where n is the size of our dataset[10].

The interpretation of the Q-Q Plot obtained this way is just the same as for the previous case, Theoretical-Theoretical Q-Q Plot.

R code, Financial Returns are not Normally Distributed: It is somehow classical fact now that the returns of stocks are not following a Normal Distribution.

Let us consider here an example to explain the claim: we consider some stock, say, FB stock, and we want to see if their weekly rates of returns are Normally Distributed or not. To that end we want to use the Q-Q Plot.

We start by downloading the data: we navigate to http://finance.yahoo.com/. In the search bar we enter "FB" and chose the "FB, Facebook, Inc.". The first page will show some basic info about

[9] This is the most used one, for other choices see Wikipedia, https://en.wikipedia.org/wiki/Q-Q_plot

[10] No need to take more that n quantiles, since if we have n data points, then taking more than n quantiles will not give new ones.

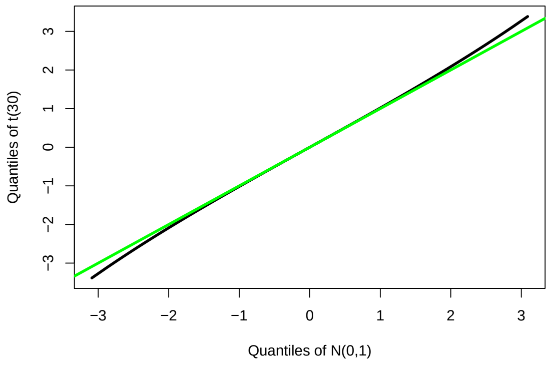


Fig. 5.18: Q-Q Plot: $\mathrm{t}(30)$ vs $\mathcal{N}(0,1)$

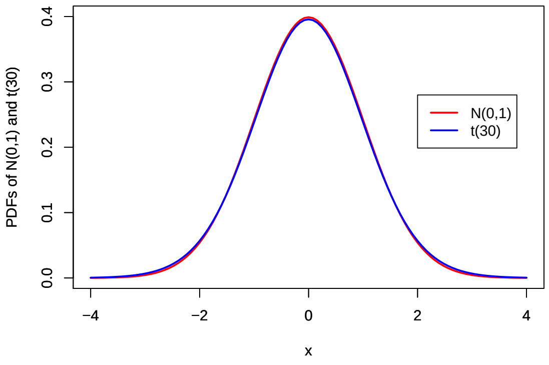


Fig. 5.19: PDF Plot: Plot: $\mathrm{t}(30)$ vs $\mathcal{N}(0,1)$

the company and stock, and, particularly, the current price for 1 share. Then we go to "Historical Data", choose the Time Period "Max" (this will download all available data), choose the Frequency "Weekly" (because we want to calculate weekly returns), hit "Apply", and then choose "Download Data". This will download the historical price data for FB stock to *FB.csv* file (.csv stands for the Comma Separated File). This file can be viewed in Excel. It has a header, the top row, with the names of variables (features) - "Date", "Open", "High", "Low", "Close", "Adj Close", "Volume". Here "Date" is the date ☺, "Open" is the price at the very beginning of that week, "Close" is the price at the very end of the week, "High" is the highest price during that week, "Low" is the lowest price for that week, "Adj Close" is the price at the end of the week, adjusted, if dividend payments or splits happened during that week. "Volume" is the number of shares traded (bought or sold) that week. We will use the "Adj Close" Prices, and will do our calculations in **R**.

So first, we read import the dataset into **R**. To that end, we will use the command

```
xx <- read.csv(file.choose(), header = TRUE)
```

Here *read.csv* is obviously to read the .csv file, *file.choose()* is to open the "Open" dialog to choose the .csv file (otherwise, you need to specify the path to that file), and *header* = *TRUE* is to indicate that in our .csv file the first row is the header, it shows the variables names.

After running this command, xx will be a *data frame* having the same structure as the .csv file. You can see the content of xx just by clicking on it in the R-Studio's Environment tab.

Now, we select the "Adj Close" column values:

... To Be Continued ....

```
#Financial returns Q-Q plot and non-Normality
dataset <- read.csv(file.choose(), header = TRUE)
adjcloseprices <- dataset$Adj.Close
rate_of_ret <- diff(adjcloseprices)/adjcloseprices[1:(length(adjcloseprices)-1)]
hist(rate_of_ret)
qqnorm(rate_of_ret)
qqline(rate_of_ret)
```

5.3.3 Q-Q Plot for two Datasets

The third possible Q-Q plot is the plot for two datasets. The problem is that we want to check how similar are our datasets, and we want to check if they are coming from the same distribution. To that end, we plot in 2D the quantiles of the first dataset vs the quantiles of the second one.

Definition 5.6. *The **Q-Q Plot** for datasets x and y is the plot of all points (q^y_α, q^x_α), where α runs over some values in $(0,1)$, and q^y_α and q^x_α are the α-th quantiles of the datasets y and the x, respectively.*

As above, usually one uses $\alpha = \frac{k}{n}$-th quantiles, or $\alpha = \frac{k}{n+1}$-th $(k = 1, ..., n)$ quantiles, where n is the minimum of the lengths of datasets x and y. In fact, for Q-Q Plot, we can have that x contains more datapoints then y or vice-versa.

The interpretation of the Q-Q Plot is similar to the previous cases. If the datasets are coming from the same distribution, then the Q-Q Plot will show points well-aligned with the line $q^y = q^x$, the bisector. If the Q-Q Plot is well-aligned with some line which is parallel to $q^y = q^x$, then the datasets, most probably[11], have the same distribution but with some shifted location parameter (e.g., one is from $\mathcal{N}(0,1)$, and the other is from $\mathcal{N}(-2,1)$). And if the datasets are well aligned around some other line, then, most probably, they are from the same distributions, but with different scale and location parameters. The convex or concave shapes on the right or left are signaling about the heavier or lighter tails.

Fig. 5.20-5.23 below show some experiments in **R** for the Q-Q Plots: the datasets are from the same distributions. Fig. 5.24 is a Q-Q Plot for datasets from different distributions. The code is given below[12]

R CODE, EXPERIMENTS WITH Q-Q PLOT FOR 2 DATASETS:

```
#Q-Q Plot for 2 Datasets, experiment no. 1
x <- rnorm(100, mean = 0, sd = 1)
y <- rnorm(200, mean = 0, sd = 1)
qqplot(x,y, pch = 16)
abline(0,1, col = "red", lwd = 2)

#Q-Q Plot for 2 Datasets, experiment no. 2
x <- rnorm(100, mean = 0, sd = 1)
y <- rnorm(200, mean = -0.7, sd = 1)
qqplot(x,y, pch = 16)
abline(0,1, col = "red", lwd = 2)
abline(-0.7, 1, col = "green", lwd = 2)

#Q-Q Plot for 2 Datasets, experiment no. 3
x <- rnorm(100, mean = 0, sd = 1)
y <- rnorm(200, mean = -0.7, sd = 5)
qqplot(x,y, pch = 16)
abline(0,1, col = "red", lwd = 2)
abline(-0.7,5, col = "green", lwd = 2)
```

[11] I am using "most probably", because I cannot say for sure, and nobody can say for sure!

[12] of course, if you will run these codes, you will not get exactly the same picture, because every time computer generates different random samples. We can fix the random sample by using the command *set.seed(n)*, where n is some number.

```
#Q-Q Plot for 2 Datasets, experiment no. 4
x <- rexp(100, rate = 4)
y <- rexp(100, rate = 10)
qqplot(x,y, pch = 16)
abline(0,1, col = "red", lwd = 2)
abline(0,0.4, col = "green", lwd = 2)

#Q-Q Plot for 2 Datasets, experiment no. 5
x <- rnorm(200)
y <- rexp(100, rate = 4)
qqplot(x,y, pch = 16)
```

Please note that the lengths of x and y differ in our code, except the 4th experiment, when, after doing a copy-paste, I forgot to change the number of samples ⌣̈

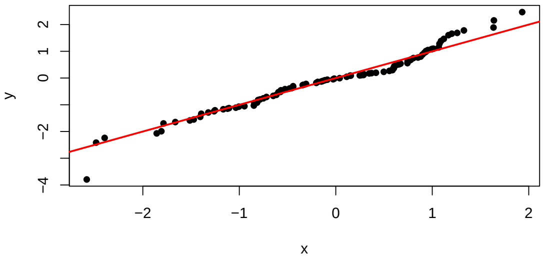


Fig. 5.20: Q-Q Plot: x and y are from $\mathcal{N}(0,1)$, red line is the line $q^y = q^x$, the bisector

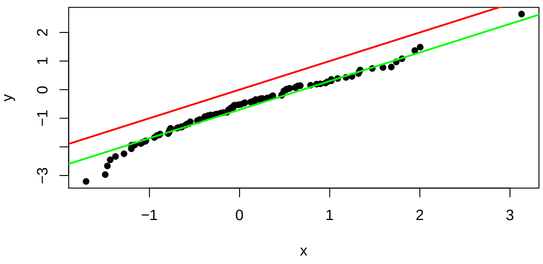


Fig. 5.21: Q-Q Plot: x is generated from $\mathcal{N}(0,1)$, y is from $\mathcal{N}(-0.7,1)$. Red line is the line $q^y = q^x$, and green line is $q^y = q^x - 0.7$

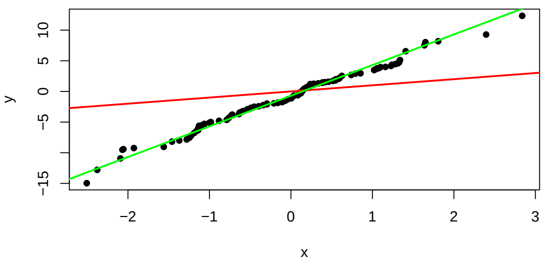


Fig. 5.22: Q-Q Plot: x is generated from $\mathcal{N}(0,1)$, y is from $\mathcal{N}(-0.7,5^2)$. Red line is the line $q^y = q^x$, and green line is $q^y = 5 \cdot q^x - 0.7$

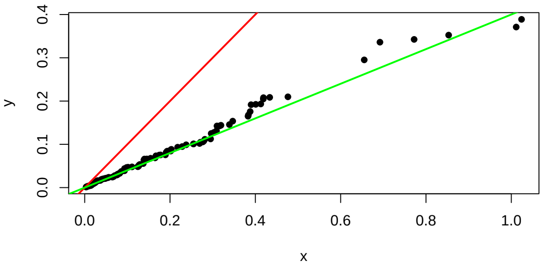


Fig. 5.23: Q-Q Plot: x is generated from $\text{Exp}(4)$, y is from $\text{Exp}(10)$. Red line is the line $q^y = q^x$, and green line is $q^y = \frac{4}{10} q^x$.

Remark, Q-Q Plots:

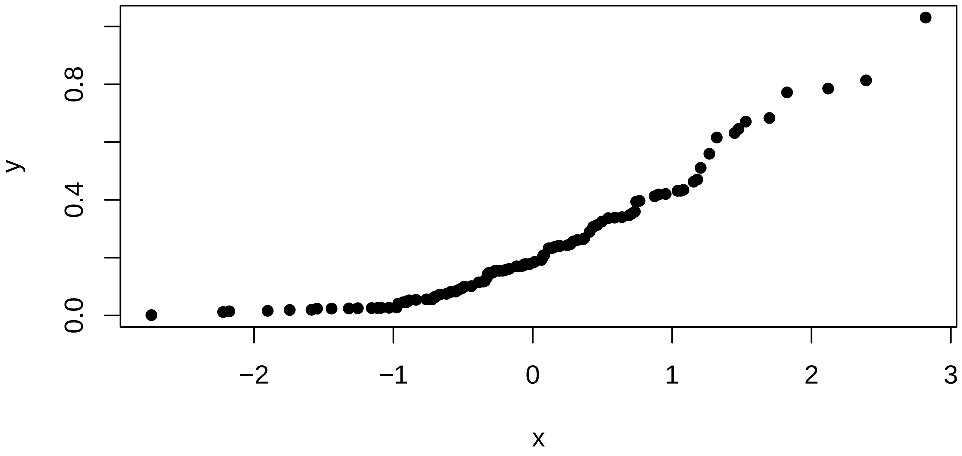


Fig. 5.24: Q-Q Plot: x is generated from $\mathcal{N}(0,1)$, y is from $\mathrm{Exp}(4)$

- If we have 2 datasets of the same size n, x and y, then, if we will use many α-s for the quantile orders, in some sense the Q-Q plot of that dataset will be the plot of points $(x_{(i)}, y_{(i)})$, $i = 1, ..., n$.
- Please note that in Q-Q Plot we are not graphing the data values, rather we plot quantiles. So you will not be able to recover data values from the Q-Q Plot.

Remark, Interpretation of Q-Q Plots : Nice interpretation for Q-Q Plots is given at http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot.

Exercise: Write **R** functions **qqunif**, **qqexp** that will do similar things like **qqnorm**.

Week 6

Exploratory Data Analysis for Bivariate Data: Covariance and Correlation

Now assume we have two datasets of the same size for the variables x and y:

$$x_1, ..., x_n \qquad \text{and} \qquad y_1, ..., y_n,$$

and we want to explore the dependencies between that variables x and y. Say, x is the height of a person and y is the width of the same person (or the salary ⌣̈. Btw, how to calculate the width of a person ? ⌣̈). Or, say, we want to find a relationship between the time spent in FaceBook and Statistics Grade, or the (stroong!) relationship between the number of missed Stat classes and Stat Grade.

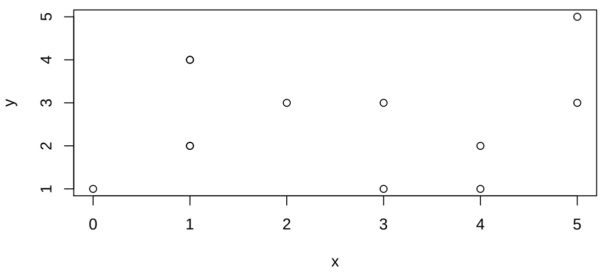
Here, like in 1D case, we will describe two methods - Geometric methods to visualize the relationship, and some numerical measures for that.

6.1 Visualizing the Data: ScatterPlot or the Point Cloud

One of the natural methods to visualize the data is to draw y vs x, i.e., to draw the points (x_i, y_i) for $i = 1, ..., n$:

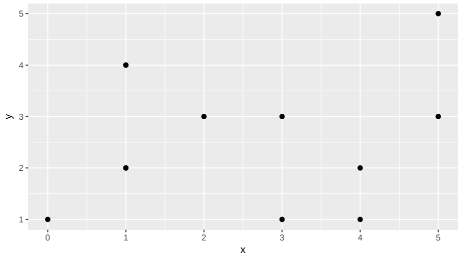
R code, Scatter Plot:

```
#Scatterplot or Points Cloud
x <- c(0,1,2,3,1,4,1,5,1,5,4,3)
y <- c(1,2,3,3,2,1,4,5,4,3,2,1)
plot(x,y)
```



R code, Scatter Plot, with ggplot2:

```
#Scatterplot or Points Cloud with ggplot2 library
library(ggplot2)
x <- c(0,1,2,3,1,4,1,5,1,5,4,3)
y <- c(1,2,3,3,2,1,4,5,4,3,2,1)
z <- data.frame(x,y)
ggplot(z, aes(x=x, y=y)) + geom_point(size=2)
```



In this case we assume that the observation x_i is related somehow (or maybe unrelated) to y_i, **with the same index** i. Say, x_1 and y_1 are two features of the same object (e.g., the height and age of the person no. 1; or the stock price for General Electric Stock at some time instant and the value of the DJIA Index at the same time). So we plot y_i vs x_i. And if you will shuffle the datasets, the scatter plot will not be the same!

6.2 Sample Covariance and the Correlation Coefficient

Now we want to give some numerical measure of relationship between our datasets x and y. Recall from the Probability course that the covariance and the correlation coefficient are measures for the linear relationship between two r.v.'s. Now, for our observations, we define similar notions[1]:

Definition 6.1. *The Sample Covariance of the datasets x and y is*

$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

or

$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n-1}$$

Here $\overline{x}$ and $\overline{y}$ are the sample means for the datasets x and y.

Definition 6.2. *We say that datasets x and y are* ***uncorrelated****, if* $\mathrm{cov}(x,y) = 0$.

Remark, Uncorrelatedness and Independence: In probability theory, we also have the notion of independence. And we are then describing the relationship between these two notions: independence and uncorrelatedness. And in Probability Theory, if X and Y are independent r.v.s, then they are also uncorrelated. The inverse statement is not true in the general case.

Here, for the datasets, the notion of independence is not defined.

You can see that, as in the case of the sample variance, we introduce 2 different formulas for the sample covariance. And, as in the case of the sample variance, different authors use either the first or the second one. Later we will explain why sometimes it is preferable to choose $n-1$ as a denominator instead of n.

Example, Sample Covariance: Assume we are given the following datasets:

$$x: 1,2,3,1,2,3,4,3,2,4,5, \qquad \text{and} \qquad y: -1,2,3,-1,-1,0,0,2,3,4,1.$$

Then you can surely calculate the covariance between x and y ⌣̈

One of the drawbacks in covariance is that it can be any number, anything from $-\infty$ to $+\infty$, and, when comparing the relationships between 2 pairs of datasets, we cannot use covariances. I

[1] Recall again, that there is no uncertainty in our case here, there is no anything probabilistic in our observation yet: we just have some numbers recorded. Of course, we can make from that numbers a r.v. taking the recorded values with equal probabilities. This will explain the introduction of the sample covariance and sample correlation coefficient.

mean, if we have 2 pairs of datasets, (x, y) is the first pair, and (z, t) is the other pair of datasets, and we know that $cov(x, y)$ is very large compared to $cov(z, t)$, that will not show that the relationship between x and y is stronger that the relationship between z and t. Even worse, if z and t will be the same as x and y, respectively, but with other units of measurements, then $cov(x, y)$ will not be equal to $cov(z, t)$.

Example, Covariances for two dataset pairs: Here we need to have an example of calculation.

The normalized version of the covariance is the correlation coefficient.

Definition 6.3. *The Sample Correlation Coefficient of the datasets x and y is*

$$cor(x, y) = \rho_{xy} = \frac{cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x, y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

where s_x and s_y are the standard deviations for x and y, respectively.

If $s_x = 0$ or $s_y = 0$, then we take $cor(x, y) = 0$ by definition.

Important is to remember to take the same denominator for the covariance and standard deviations - either n everywhere or $n-1$ everywhere. So it is not correct to calculate the covariance using n in the denominator, then take $n-1$ when calculating the standard deviations, and then calculate the correlation coefficient.

In both cases, when one calculates Standard Deviations and Covariance by using n simultaneously or $n-1$ simultaneously in the denominator, we will obtain

$$cor(x, y) = \rho_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \overline{x})^2 \cdot \sum_{k=1}^{n} (y_k - \overline{y})^2}}$$

Another formula to calc the correlation coefficient is

$$cor(x, y) = \rho_{xy} = \frac{\sum_{k=1}^{n} x_k y_k - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{k=1}^{n} x_k^2 - n \cdot (\overline{x})^2} \cdot \sqrt{\sum_{k=1}^{n} y_k^2 - n \cdot (\overline{y})^2}}.$$

Note: Again Cov and Cor can be interpreted as the cov and cor for r.v.'s X and Y taking the values x_i and y_i, correspondingly, with the probabilities $\frac{1}{n}$ (or $\frac{1}{n-1}$).

What are measuring covariance and correlation coefficient - they are giving us some "measure of linear dependence" between x and y, a "measure of joint linear variability", the strength and the direction of the linear relationship between the data. If, say, we get $\rho_{xy} = 0$, then x and y are **uncorrelated**, and we mean that there is no (linear) relationship between x and y. Soon we will see that if $\rho_{xy} = 1$, then there is an exact linear and increasing relationship between x and y, and if ρ_{xy} is very close to 1, then there is a strong linear increasing relationship between x and y.

```
#Covariance and Correlation
x <- rnorm(40)
y <- rnorm(40)
plot(x,y)
cov(x,y)
cor(x,y)
```

Now, if $cov(x, y)$ or ρ_{xy} are positive, the we say that x and y are positively correlated. This means, roughly[2], if $x_k > \overline{x}$, then also y_k tends to be larger than $\overline{y}$. So there is a tendency: if x increases, then y tends to increase also.

```
#Covariance and Correlation, positive correlation
x <- rnorm(40)
e <- rnorm(40)
y <- 2.5*x+e
plot(x,y)
cov(x,y)
cor(x,y)
```

```
#Covariance and Correlation, negative correlation
x <- rnorm(40)
e <- rnorm(40)
y <- -1.4*x+e
plot(x,y)
cov(x,y)
cor(x,y)
```

Example:

```
#Covariance and Correlation, real data
state.x77
state <- as.data.frame(state.x77)
str(state) #structure of the dataset state
head(state)
tail(state)
x <- state$Illiteracy
y <- state$Murder
plot(x,y)
cov(x,y)
cor(x,y)
```

and

```
#Covariance and Correlation, real data, cont
state <- as.data.frame(state.x77)
x <- state$Illiteracy
y <- state$‘Life Exp‘
plot(x,y)
cov(x,y)
cor(x,y)
```

[2] Veery roughly!

Example: See https://en.wikipedia.org/wiki/Pearson_correlation_coefficient or https://en.wikipedia.org/wiki/Correlation_and_dependence for some graphical examples.

The difference between cov and cor is that cor is normalized, in the sense that

Proposition 6.1. *For any datasets x, y,*

$$-1 \leqslant \rho_{xy} \leqslant 1.$$

Moreover,

- $\rho_{xy} = 1$ *iff there exists a constant $a > 0$ and $b \in \mathbb{R}$ such that*[3] $y_i = a \cdot x_i + b$ *for any* $i = 1, ..., n$.
- $\rho_{xy} = -1$ *iff there exists a constant $a < 0$ and $b \in \mathbb{R}$ such that*[4] $y_i = a \cdot x_i + b$ *for any* $i = 1, ..., n$.

Exercise: Prove this Proposition.

Example: For example, the correlation coefficient of a dataset x with itself gives 1, i.e.,

$$\mathrm{cor}(x, x) = \rho_{xx} = 1,$$

and the covariance of a dataset with itself is the variance:

$$\mathrm{cov}(x, x) = s_{xx} = \mathrm{var}(x) = s_x^2.\blacksquare$$

Another important aspect of the correlation coefficient is that it is dimensionless, it is independent on the units we are calculating the data x and y. Say, if we measure the weight x in Kg's and the height y of a person in meters, then we will obtain some number for the covariance between x and y. If we will change our units to grams for x and centimeters for y, the covariance will be another number (some multiple of the previous one). But in these both cases, the correlation coefficient will be the same.

REMARK, CORRELATION AND CAUSATION: It is important to note that high correlation between two datasets x and y doesn't mean that there is a causal relationship. In general, it is not true that x influences y or y influences x. It may be the case that some other feature, called a latent feature, is influencing both x and y.

6.3 Appendix: Sample Statistics and Random Variable characteristics

As we have seen above, many descriptive statistics measures are the analogues for the corresponding ones form the Probability Theory. For example, the idea of the Sample Mean:

Sample Covariance: Recall that for the jointly distributed r.v. X and Y, the covariance $\mathrm{Cov}(X, Y)$ is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big((x - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big).$$

Now, let us obtain from this the sample covariance formula. Assume we have 2 datasets $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$.

Let X be a discrete r.v. taking the values $x_1, x_2, ..., x_n$ (it is possible, of course, that some x_i's coincide, but this is OK for us), and Y be a discrete r.v. with values $y_1, y_2, ..., y_n$. Now, we need to

[3] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

[4] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

describe the probabilities of taking that values. If we will describe marginal (individual) PMF's of X and Y, that wil not be enough for calculating the $\mathrm{Cov}(X,Y)$. For this calculation, we need to have the Joint PMF of X and Y. We define

Table 6.1: The PMF of X and Y

$Y \setminus X$	x_1	x_2	$\cdots$	x_n
y_1	$\frac{1}{n}$	0	$\cdots$	0
y_2	0	$\frac{1}{n}$	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	0	0	$\cdots$	$\frac{1}{n}$

So we give the equal probabilities for the value (x_k, y_k), $k = 1, ..., n$, and also we assume that the event $X = x_1$, $Y = y_3$ is impossible - x_k and y_k are linked to each other, if we observe x_1, then we observe y_1 (say, x_1 is the year of study of a person, and y_1 is his/her salary). This can be written also in the form:

$$\mathbb{P}(X = x_i, Y = y_j) = \begin{cases} \frac{1}{n}, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \neq j. \end{cases}$$

Now, clearly,

$$\mathbb{E}(X) = \frac{\sum_{k=1}^{n} x_k}{n} = \overline{x}, \qquad \text{and} \qquad \mathbb{E}(Y) = \frac{\sum_{k=1}^{n} y_k}{n} = \overline{y}$$

and if we will calculate the covariance $\mathrm{Cov}(X,Y)$, then we will obtain

$$\mathrm{Cov}(X,Y) = \sum_{i,j=1}^{n} (x_i - \overline{x})(y_j - \overline{y}) \cdot \mathbb{P}(X = x_i, Y = y_j) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}).$$