Assessing Normality

1. Visual data exploration

A big part of our theory of linear models has been under the normal model; that is the most successful part of the theory applies when $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where we assumed that $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}\,, \sigma^2 \boldsymbol{I})$; equivalently, that $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{N}(0\,, \sigma^2)$'s.

A natural question, for a given data set, is to ask, "is the noise coming as i.i.d. $\mathrm{N}(0\,, \sigma^2)$'s"?

Since the noise is not observable in our model, it seems natural that we "estimate" it using the residuals $\boldsymbol{e} := \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. It stands to reason that if $\boldsymbol{\varepsilon}$ is a vector of n i.i.d. $\mathrm{N}(0\,, \sigma^2)$'s, then the histogram of $e_1\,, \ldots, e_n$ should look like a normal density. One should not underestimate the utility of this simple idea; for instance, we should see, roughly, that:

- approximately 68.3% of the e_i's should fall in $[-S^2\,, S^2]$,
- approximately 95.4% of the e_i's should fall in $[-2S^2\,, 2S^2]$, etc.

This is very useful as a first attempt to assess the normality of the noise in our problem. But it is not conclusive since we cannot assign confidence levels [the method is a little heuristic]. It turns out that this method can be improved upon in several directions, and with only a little more effort.

2. General remarks

We can use the Pearson's χ^2-test in order to test whether a certain data comes from the $N(\mu_0\,, \sigma_0^2)$ distribution, where μ_0 and σ_0 are known. Now we wish to address the same problem, but in the more interesting case

that μ_0 and σ_0 are *unknown*. [For instance, you may wish to know whether or not you are allowed to use the usual homoskedasticity assumption in the usual measurement-error model of linear models.]

Here we discuss briefly some solutions to this important problem. Although we write our approach specifically in the context of linear models, these ideas can be developed more generally to test for normality of data in other settings.

3. Histograms

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. The pressing question is, "is it true that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$"?

To answer this, consider the "residuals,"

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

If $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then one would like to think that the histogram of the e_i's should look like a normal pdf with mean 0 and variance σ^2 (why?). How close is close? It helps to think more generally.

Consider a sample $U_1, \ldots, U_n$ (e.g., $U_i = e_i$). We wish to know where the U_i's are coming from a normal distribution. The first thing to do is to plot the histogram. In R you type,

```
hist(u,nclass=n)
```

where u denotes the vector of the samples $U_1, \ldots, U_n$ and n denotes the number of bins in the histogram.

For instance, consider the following exam data:

```
16.8 9.2 0.0 17.6 15.2 0.0 0.0 10.4 10.4 14.0 11.2 13.6 12.4
14.8 13.2 17.6 9.2 7.6 9.2 14.4 14.8 15.6 14.4 4.4 14.0 14.4
0.0 0.0 10.8 16.8 0.0 15.2 12.8 14.4 14.0 17.2 0.0 14.4 17.2
0.0 0.0 0.0 14.0 5.6 0.0 0.0 13.2 17.6 16.0 16.0 0.0 12.0 0.0
13.6 16.0 8.4 11.6 0.0 10.4 0.0 14.4 0.0 18.4 17.2 14.8 16.0
16.0 0.0 10.0 13.6 12.0 15.2
```

The command hist(f1.dat,nclass=15) produces Figure 8.1(a).[1]

Try this for different values of nclass to see what types of histograms you can obtain. You should always ask, "which one represents the truth the best"? Is there a unique answer?

Now the data $U_1, \ldots, U_n$ is probably not coming from a normal distribution if the histogram does not have the "right" shape. Ideally, it would be symmetric, and the tails of the distribution taper off rapidly.

[1] You can obtain this data freely from the website below:
http://www.math.utah.edu/~davar/math6010/2011/Notes/f1.dat.

In Figure 8.1(a), there were many students who did not take the exam in question. They received a '0' but this grade should probably not contribute to our knowledge of the distribution of all such grades. Figure 8.1(b) shows the histogram of the same data set when the zeros are removed. [This histogram appears to be closer to a normal density.]

4. QQ-Plots

QQ-plots are a better way to assess how closely a sample follows a certain distribution.

To understand the basic idea note that if $U_1, \ldots, U_n$ is a sample from a normal distribution with mean μ and variance σ^2, then about 68.3% of the sample points should fall in $[\mu - \sigma, \mu + \sigma]$, 95.4% should fall in $[\mu - 2\sigma, \mu + 2\sigma]$, etc.

Now let us be more careful still. Let $U_{1:n} \le \cdots \le U_{n:n}$ denote the order statistics of $U_1, \ldots, U_n$. Then no matter how you make things precise, the fraction of data "below" $U_{j:n}$ is $(j \pm 1)/n$. So we make a continuity correction and *define* the fraction of the data below $U_{j:n}$ to be $(j - \frac{1}{2})/n$.

Consider the normal "quantiles," $q_1, q_2, \ldots, q_n$:

$$\Phi(q_j) := \int_{-\infty}^{q_j} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = \frac{j - \frac{1}{2}}{n}; \qquad \text{i.e., } q_j := \Phi^{-1}\left(\frac{j - \frac{1}{2}}{n}\right).$$

Now suppose once again that $U_1, \ldots, U_n \sim \mathrm{N}(\mu, \sigma^2)$ is a random [i.i.d.] sample. Let $Z_j := (U_j - \mu)/\sigma$, so that $Z_1, \ldots, Z_n \sim \mathrm{N}(0, 1)$. The Z's are standardized data, and we expect the fraction of the standardized data that fall below q_j to be about $(j - \frac{1}{2})/n$. In other words, we can put together our observations to deduce that $Z_{j:n} \approx q_j$. Because $U_{j:n} = \sigma Z_{j:n} + \mu$, it follows that $U_{j:n} \approx \sigma q_j + \mu$. In other words, we expect the sample order statistics $U_{1:n}, \ldots, U_{n:n}$ to be very close to some linear function of the normal quantiles $q_1, \ldots, q_n$. In other words, if $U_1, \ldots, U_n$ is a random sample from some normal distribution, then we expect the scatterplot of the pairs $(q_1, U_{1:n}), \ldots, (q_n, U_{n:n})$ to follow closely a line. [The slope and intercept are σ and μ, respectively.]

QQ-plots are simply the plots of the $\mathrm{N}(0, 1)$-quantiles $q_1, \ldots, q_n$ versus the order statistics $U_{1:n}, \ldots, U_{n:n}$. To draw the qqplot of a vector $\boldsymbol{u}$ in R, you simply type

qqnorm(u).

Figure 8.2(a) contains the qq-plot of the exam data we have been studying here.

5. The Correlation Coefficient of the QQ-Plot

In its complete form, the R-command qqnorm has the following syntax:

$$\text{qqnorm(u, datax = FALSE, plot = TRUE).}$$

The parameter u denotes the data; datax is "FALSE" if the data values are drawn on the y-axis (default). It is "TRUE" if you wish to plot $(U_{j:n}, q_j)$ instead of the more traditional $(q_j, U_{j:n})$. The option plot=TRUE (default) tells R to plot the qq-plot, whereas plot=FALSE produces a vector. So for instance, try

$$\text{V = qqnorm(u, plot = FALSE).}$$

This creates two vectors: V\$$x$ and V\$$y$. The first contains the values of all q_j's, and the second all of the $U_{j:n}$'s. So now you can compute the correlation coefficient of the qq-plot by typing:

$$\begin{aligned}&\text{V = qqnorm(u, plot = FALSE)}\\&\text{cor(V\$}x\text{, V\$}y\text{).}\end{aligned}$$

If you do this for the qq-plot of the grade data, then you will find a correlation of ≈ 0.910. After censoring out the no-show exams, we obtain a correlation of ≈ 0.971. This produces a noticeable difference, and shows that the grades are indeed normal.

In fact, one can analyse this procedure statistically ["is the sample correlation coefficient corresponding to the line sufficiently close to ± 1"?].

6. Some benchmarks

Figures 8.3 and 8.4 contain four distinct examples. I have used qq-plot in the prorgam environment "R." The image on the left-hand side of Figure 8.3 shows a simulation of 10000 standard normal random variables (in R, you type `x=rnorm(10000,0,1)`), and its qq-plot is drawn on typing `qqnorm(x)`. In a very strong sense, **this figure is the benchmark**.

The image on the right-hand side of Figure 8.3 shows a simulation of 10000 standard Cauchy random variables. That is, the density function is $f(x) = (1/\pi)(1 + x^2)^{-1}$. This is done by typing `y=rcauchy(10000,0,1)`, and the resulting qq-plot is produced upon typing `qqnorm(y)`. We know that the Cauchy has much fatter tails than normals. For instance,

$$\mathrm{P}\{\text{Cauchy} > a\} = \frac{1}{\pi}\int_a^\infty \frac{dx}{1+x^2} \sim \frac{1}{\pi a} \quad (\text{as } a \to \infty),$$

whereas $\mathrm{P}\{\mathrm{N}(0,1) > a\}$ decays faster than exponentially.[2] Therefore, for a large,

$$\mathrm{P}\{\mathrm{N}(0,1) > a\} \ll \mathrm{P}\{\text{Cauchy} > a\}.$$

This heavy-tailedness can be read off in Figure 8.3(b): The Cauchy qq-plot grows faster than linearly on the right-hand side. *And this means that the standard Cauchy distribution has fatter right-tails.* Similar remarks apply to the left tails.

Figure 8.4(a) shows the result of the qq-plot of a simulation of 10000 iid uniform-$(0,1)$ random variables. [To generate these uniform random variables you type, `runif(10000,0,1)`.]

Now uniform-$(0,1)$ random variables have much smaller tails than normals because uniforms are in fact *bounded*. This fact manifests itself in Figure 8.4(a). For instance, we can see that the right-tail of the qq-plot for uniform-$(0,1)$ grows less rapidly than linearly. And this shows that the right-tail of a uniform is much smaller than that of a normal. Similar remarks apply to the left tails.

A comparison of the figures mentioned so far should give you a feeling for how sensitive qq-plots are to the effects of tails. [All are from distributions that are symmetric about their median.] Finally, let us consider Figure 8.4(a), which shows an example of 10000 Gamma random variables with $\alpha = \beta = 1$. You generate them in R by typing `x=rgamma(10000,1,1)`. Gamma distributions are inherently *asymmetric*. You can see this immediately in the qq-plot for Gammas; see Figure 8.4(b). Because Gamma random variables are nonnegative, the left tail is much smaller than that of a normal. Hence, the left tail of the qq-plot grows more slowly than linearly. The right tail however is fatter. [This is always the case. However, for the sake of simplicity consider the special case where Gamma=Exponential.] This translates to the faster-than-linear growth of the right-tail of the corresponding qq-plot (Figure 8.4(b)).

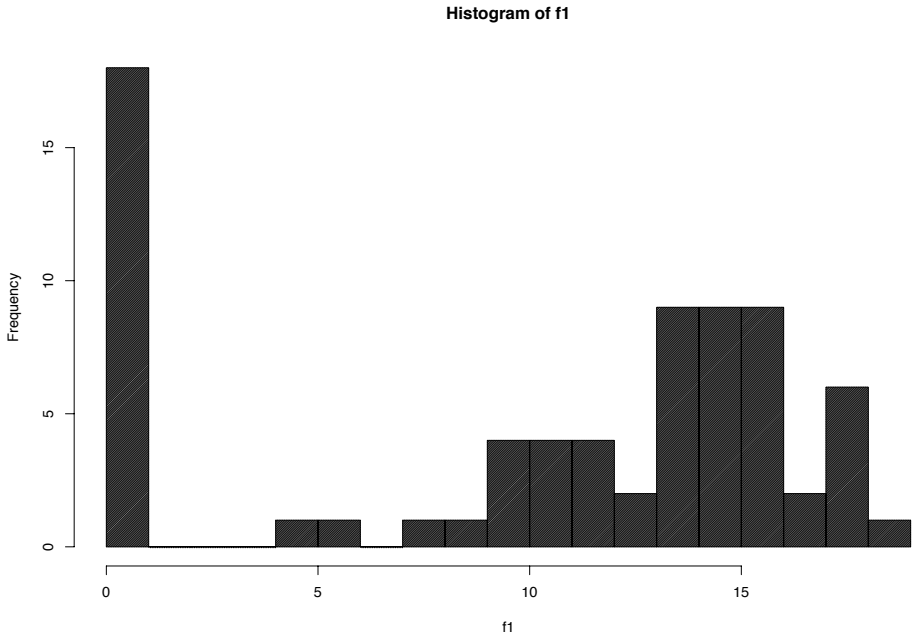
I have shown you Figures 8.3 and 8.4 in order to high-light the basic features of qq-plots in ideal settings. By "ideal" I mean "simulated data," of course.

Real data does not generally lead to such sleek plots. Nevertheless one learns a lot from simulated data, mainly because simulated data helps identify key issues without forcing us to have to deal with imperfections and other flaws.

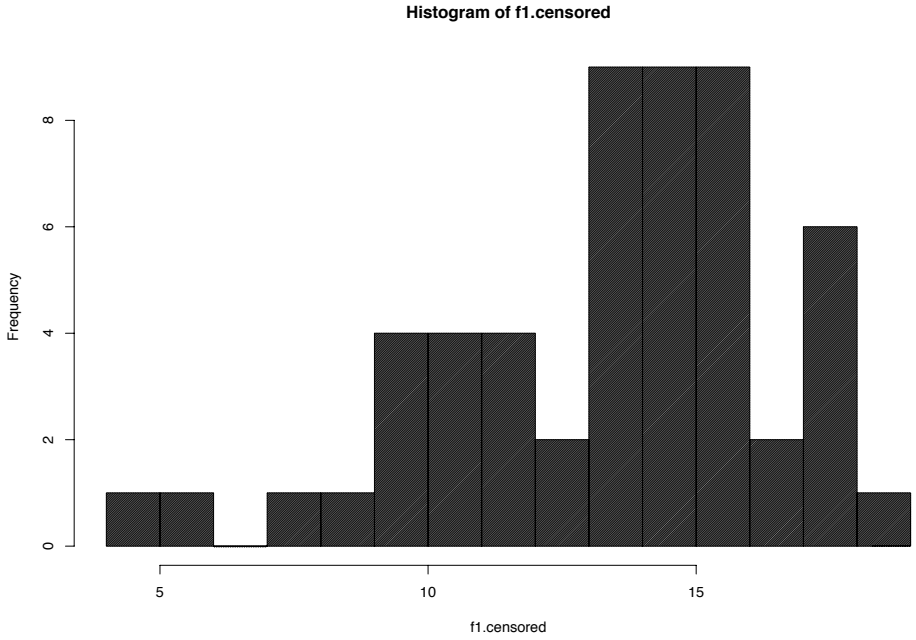
---

[2] In fact it can be shown that $\bar{\Phi}(a) := \int_a^\infty \varphi(x)\,dx \approx a^{-1}\varphi(a)$ as $a \to \infty$, where φ denotes the $\mathrm{N}(0,1)$ pdf. Here is why: Let $G(a) := a^{-1}\varphi(a)$. We know from the fundamental theorem of calculus that $\bar{\Phi}'(a) = -\varphi(a) = -aG(a)$. Also, $G'(a) = -a^{-1}G(a) - aG(a) \approx -aG(a)$ as $a \to \infty$. In summary: $\bar{\Phi}(a), G(a) \approx$ and $\bar{\Phi}'(a) \approx G'(a)$. Therefore, $\bar{\Phi}(a) \approx G(a)$, thanks to the L'Hôpital's rule of calculus.

But it is important to keep in mind that it is real data that we are ultimately after. And so the histogram and qq-plot of a certain real data set are depicted in Figure 8.5. Have a careful look and ask yourself a number of questions: Is the data normally distributed? Can you see how the shape of the histogram manifests itself in the shape and gaps of the qq-plot? Do the tails look like those of a normal distribution? To what extent is the "gap" in the histogram "real"? By this I mean to ask what do you think might happen if we change the bin-size of the histogram in Figure 8.5?

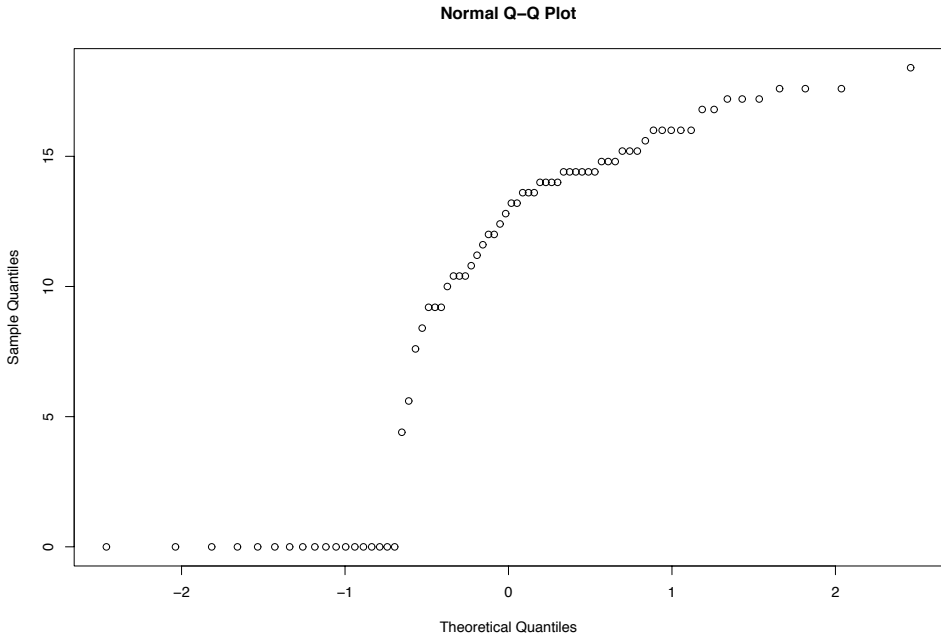


(a) Grades

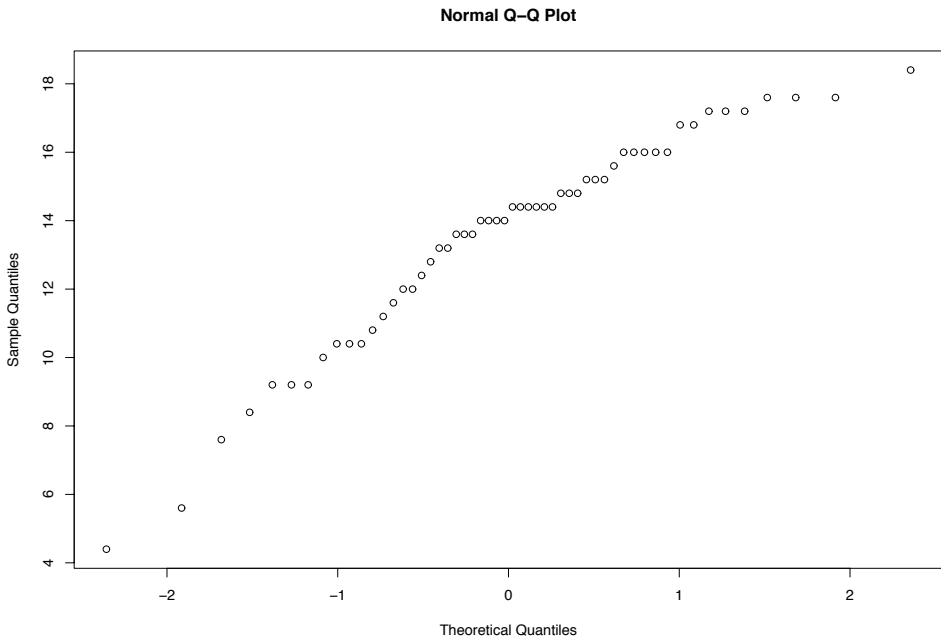


(b) Censored Grades

Figure 8.1. Histogram of grades and censored grades



(a) QQ-plot of grades



(b) QQ-plot of censored grades

Figure 8.2. QQ-plots of grades and censored grades

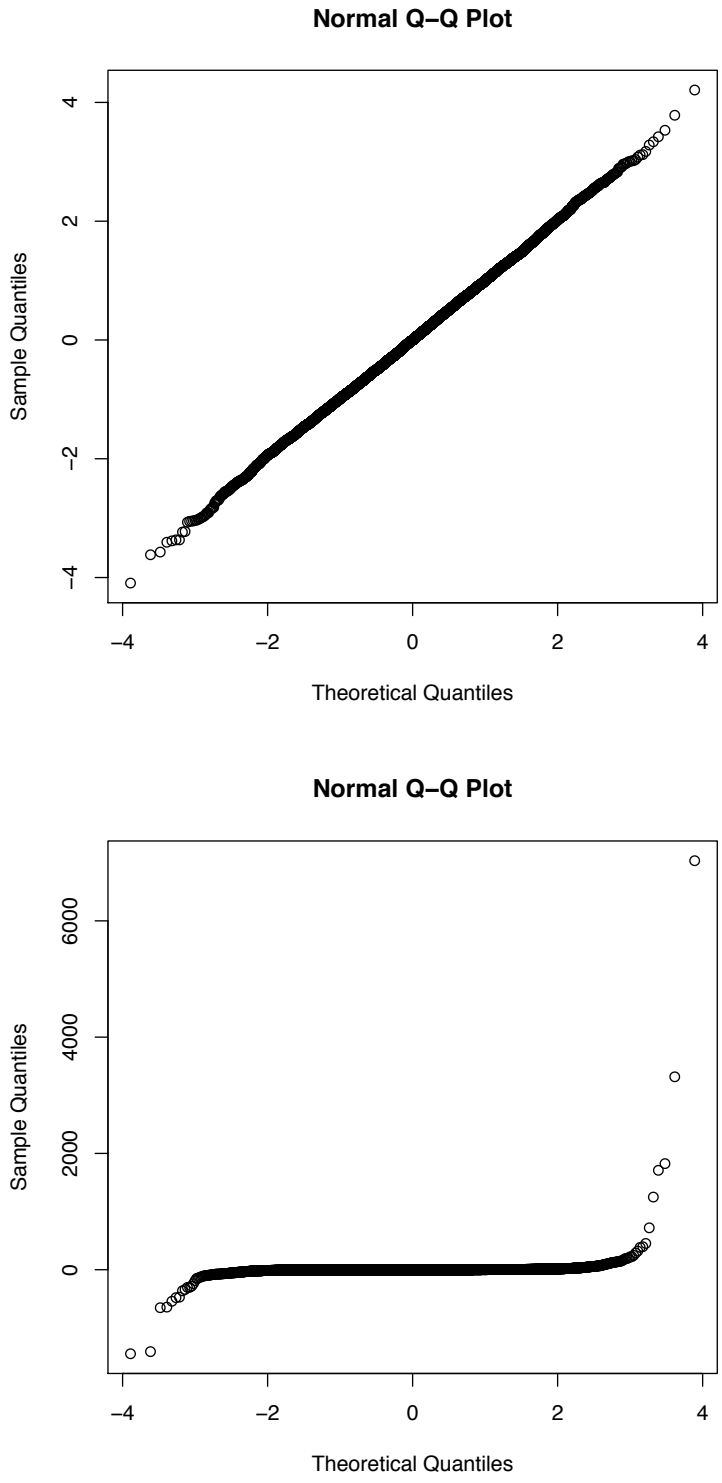


Figure 8.3. (a) is N(0 , 1) data; (b) is Cauchy data

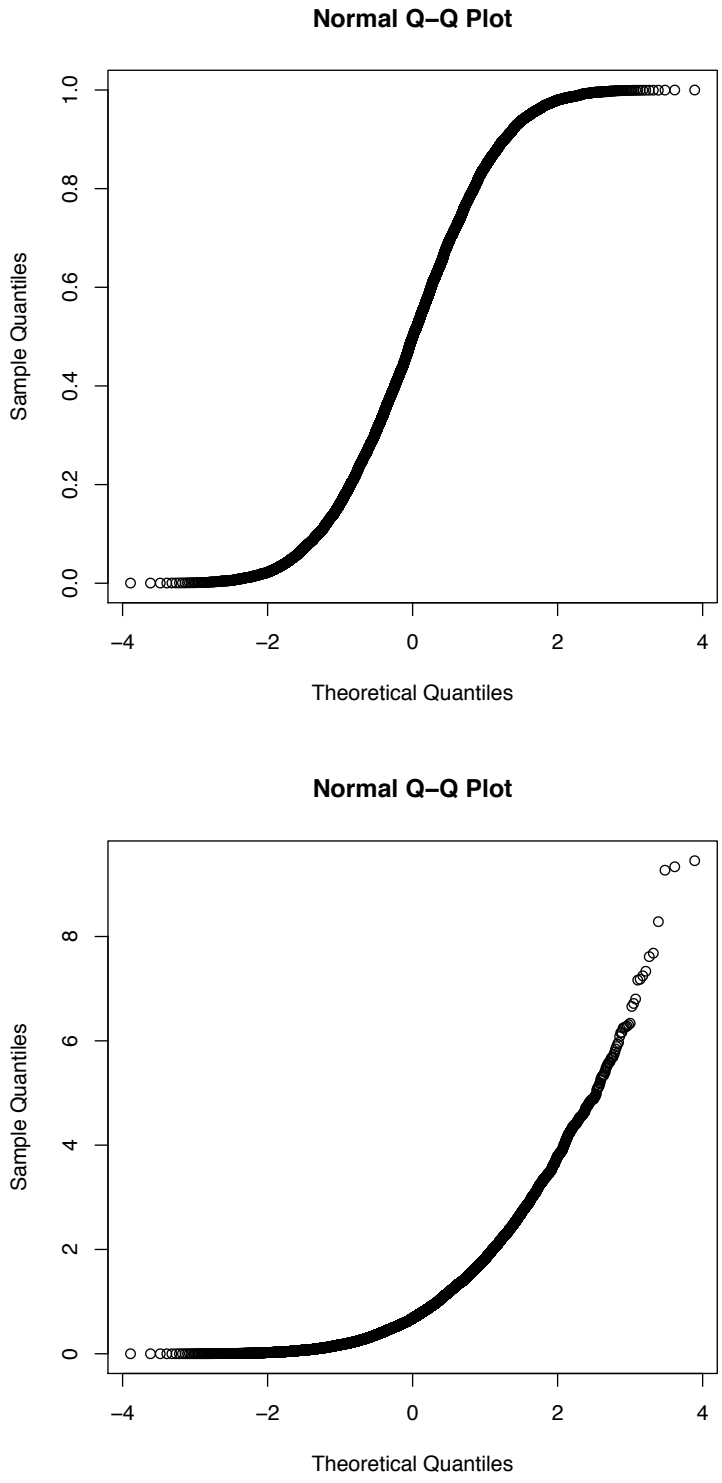


Figure 8.4. (a) is the qq-plot of unif-$(0\,,1)$; (b) is the qq-plot of a Gamma$(1\,,1)$.

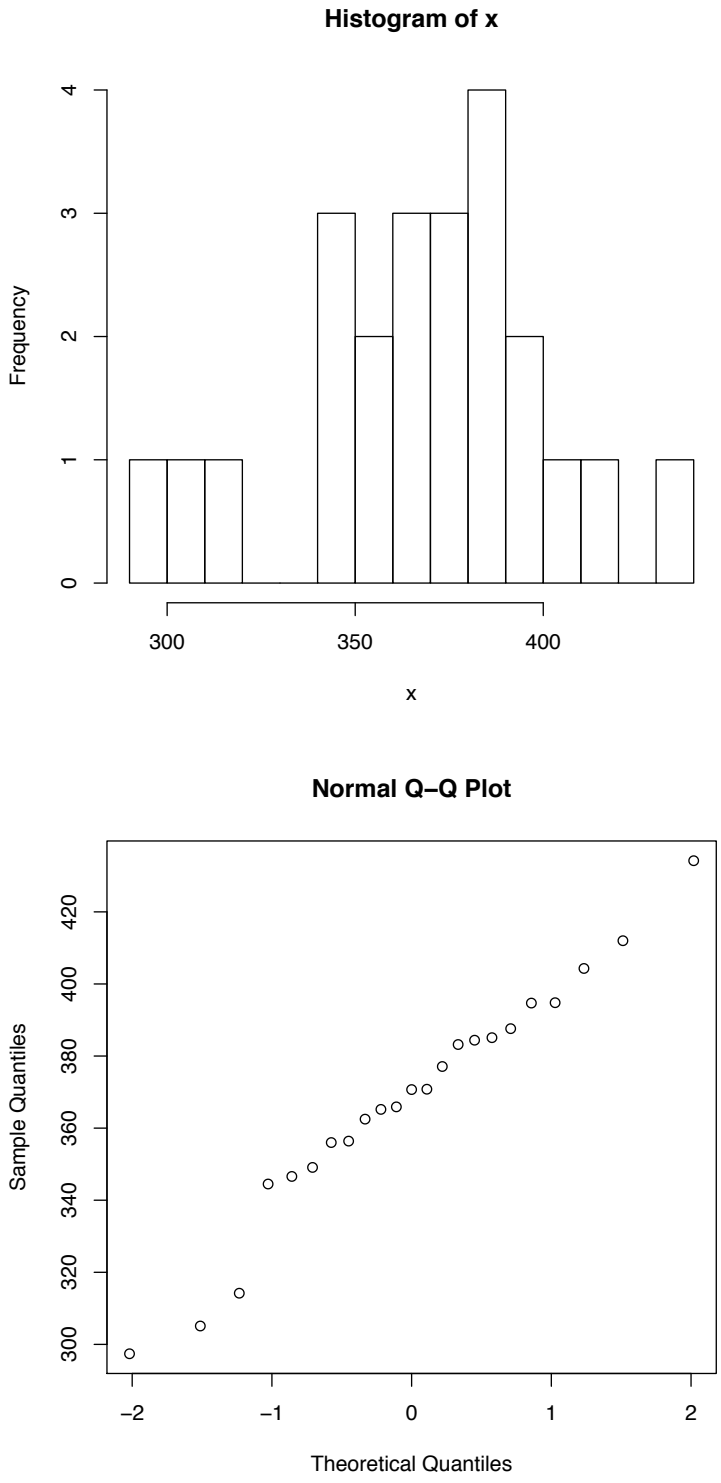


Figure 8.5. Histogram and qq-plot of the data