QQPlots and Normality Testing

MA2003B Application of Multivariate Methods
in Data Science

Raul Gomez

Skewness, Kurtosis and QQ Plots

As it is well known, any normal distribution is completely determined by its mean and standard deviation.

rgomez@tec.mx

Skewness, Kurtosis and QQ Plots

As it is well known, any normal distribution is completely determined by its mean and standard deviation.

However, many real-world datasets do not follow a normal distribution.

rgomez@tec.mx

Skewness, Kurtosis and QQ Plots

As it is well known, any normal distribution is completely determined by its mean and standard deviation.

However, many real-world datasets do not follow a normal distribution.

In this presentation, we will explore some other distributions and show how the concepts of skewness and kurtosis can help us understand its shape.

rgomez@tec.mx

Skewness, Kurtosis and QQ Plots

As it is well known, any normal distribution is completely determined by its mean and standard deviation.

However, many real-world datasets do not follow a normal distribution.

In this presentation, we will explore some other distributions and show how the concepts of skewness and kurtosis can help us understand its shape.

We will also introduce QQ plots, which are useful graphical tools for comparing the distribution of a dataset against a theoretical distribution, which is usually taken to be normal.

rgomez@tec.mx

Skewness

Definition

The **skewness** of a random variable X is defined as the third standardized moment, given by

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Skewness

Definition

The **skewness** of a random variable X is defined as the third standardized moment, given by

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Skewness measures the asymmetry of a probability distribution.

Skewness

Definition

The **skewness** of a random variable X is defined as the third standardized moment, given by

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Skewness measures the asymmetry of a probability distribution.

A distribution is said to be **positively skewed** (or right-skewed) if it has a longer tail on the right side, and **negatively skewed** (or left-skewed) if it has a longer tail on the left side.

Gamma Distribution

Definition

We say that a continuous random variable X follows a **Gamma distribution** with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, if its associated probability density function (PDF) is given by

$$f_X(x;\alpha,\beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \tag{1}$$

where $\Gamma(\alpha)$ denotes the Gamma function, defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$

The parameters α and β control the shape and rate of the distribution, respectively.

Among other applications, the Gamma distribution is frequently used to model the number of total insurance claims over a given time period.

Among other applications, the Gamma distribution is frequently used to model the number of total insurance claims over a given time period.

Remark

The skewness of a Gamma distribution is given by

$$\gamma_1 = \frac{2}{\sqrt{\alpha}}.$$

This means that the skewness is always positive, and it approaches zero as α increases. In other words, the distribution becomes more symmetric as α becomes larger.

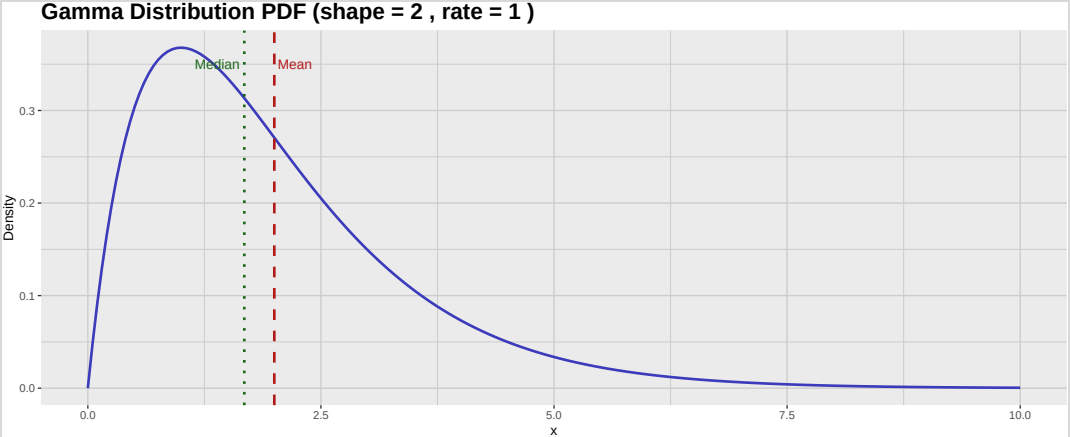

Figure 1: PDF of the gamma distribution with mean and median highlighted.

Kurtosis

Definition

The **kurtosis** of a random variable X is defined as the fourth standardized moment, given by

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Kurtosis

Definition

The **kurtosis** of a random variable X is defined as the fourth standardized moment, given by

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Kurtosis measures the "tailedness" of a probability distribution.

Kurtosis

Definition

The **kurtosis** of a random variable X is defined as the fourth standardized moment, given by

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Kurtosis measures the "tailedness" of a probability distribution.

A distribution with high kurtosis has heavier tails and a sharper peak than a normal distribution, while a distribution with low kurtosis has lighter tails and a flatter peak.

Kurtosis

Definition

The **kurtosis** of a random variable X is defined as the fourth standardized moment, given by

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4},$$

where $\mu = \mathbb{E}[X]$ is the mean and $\sigma = \sqrt{\mathbb{E}[(X-\mu)^2]}$ is the standard deviation of X.

Kurtosis measures the "tailedness" of a probability distribution.

A distribution with high kurtosis has heavier tails and a sharper peak than a normal distribution, while a distribution with low kurtosis has lighter tails and a flatter peak.

The kurtosis of a normal distribution is 3, which is why frequently we are interested in computing the *excess kurtosis* which is given by:

$$\gamma_2 = \beta_2 - 3.$$

Laplace Distribution

Definition

We say that a continuous random variable X follows a **Laplace distribution** with location parameter $\mu \in \mathbb{R}$ and scale parameter $b > 0$, if its associated probability density function (PDF) is given by

$$f_X(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right), \quad x \in \mathbb{R}. \tag{2}$$

The parameters μ and b control the location and scale of the distribution, respectively.

Among other applications, the Laplace distribution is often used in image processing and computer vision, particularly in modeling noise in images.

Among other applications, the Laplace distribution is often used in image processing and computer vision, particularly in modeling noise in images.

Remark

The excess kurtosis of a Laplace distribution is given by

$$\gamma_2 = 3.$$

This means that the Laplace distribution has heavier tails than a normal distribution.

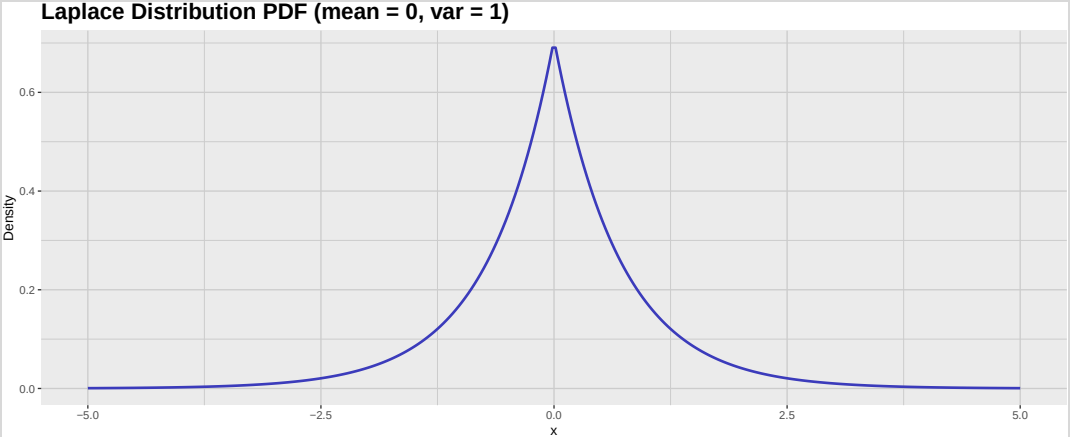

Figure 2: PDF of the Laplace distribution with mean zero and variance one.

QQ Plots

QQ plots, or quantile-quantile plots, are graphical tools used to compare the distribution of a dataset against a theoretical distribution, such as the normal distribution.

QQ Plots

QQ plots, or quantile-quantile plots, are graphical tools used to compare the distribution of a dataset against a theoretical distribution, such as the normal distribution.

They are particularly useful for assessing the normality of the data.

Constructing a QQ Plot

To construct a QQ plot, you should follow these steps:

1. **Sort the sample data:** Arrange the data in increasing order:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

2. **Calculate plotting positions (empirical quantiles):** For each ordered value $x_{(i)}$, compute the corresponding probability

$$p_i = \frac{2i-1}{2n}, \quad i = 1, 2, \ldots, n.$$

Constructing a QQ Plot

3. **Find theoretical quantiles of the normal distribution:** Compute the theoretical quantiles from the inverse cumulative distribution function (CDF) of the normal distribution:

$$q_i = \Phi^{-1}(p_i),$$

where Φ^{-1} is the quantile function of the standard normal distribution.

4. **Plot the points:** Plot the pairs $(q_i, x_{(i)})$ with q_i on the x-axis and $x_{(i)}$ on the y-axis.

5. **Add a reference line:** Typically, we add a line through the first and third quartiles to help assess normality.

The closer the points lie to this line, the more the sample resembles the specified normal distribution.

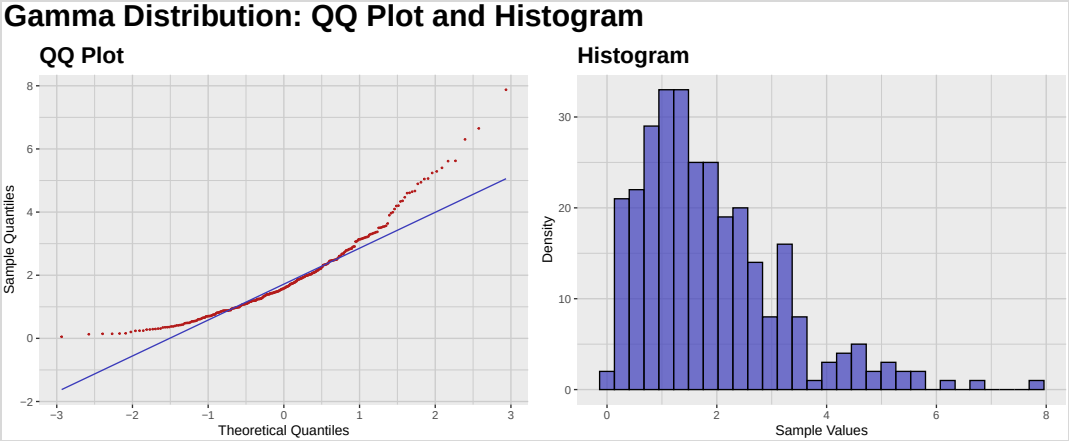

Figure 3: QQ plot and histogram of the Gamma distribution samples.

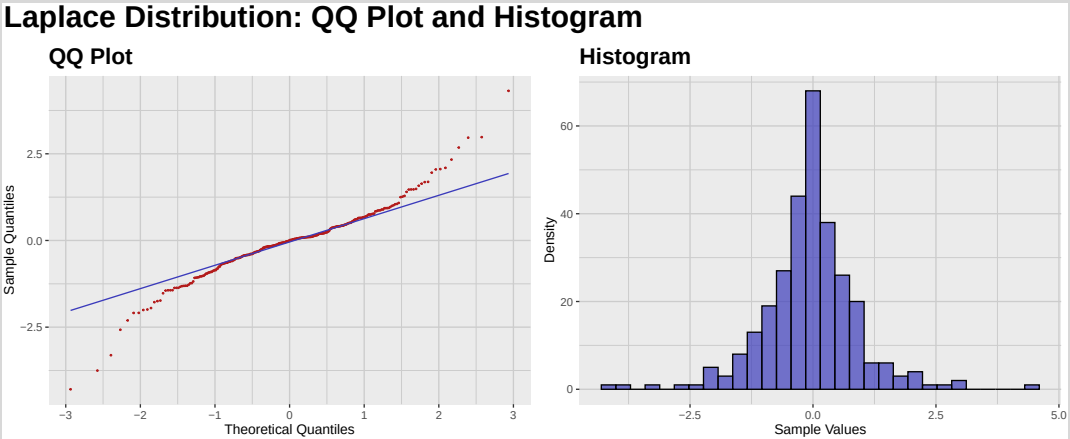

Figure 4: QQ plot and histogram of the Laplace distribution samples.

Sample Skewness and Kurtosis

So far we have defined skewness and kurtosis for random variables.

Sample Skewness and Kurtosis

So far we have defined skewness and kurtosis for random variables.

However, in practice, we often work with samples rather than entire populations.

Sample Skewness and Kurtosis

So far we have defined skewness and kurtosis for random variables.

However, in practice, we often work with samples rather than entire populations.

Therefore, we need to define sample skewness and sample kurtosis.

Definition

Let $x_1, x_2, \ldots, x_n$ be a sample of size n. The *sample mean* $\bar{x}$ and *sample variance* s^2 are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *sample skewness* g_1 given by

$$g_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3.$$

The *sample excess kurtosis* g_2 is given by

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

As an example, we will compute the sample skewness and sample excess kurtosis of the Gamma and Laplace samples we showed earlier.

As an example, we will compute the sample skewness and sample excess kurtosis of the Gamma and Laplace samples we showed earlier.

For the Gamma sample we have:

```
Rows: 1
Columns: 2
$ sample_skewness        <dbl> 1.293772
$ sample_excess_kurtosis <dbl> -0.8356471
```

As an example, we will compute the sample skewness and sample excess kurtosis of the Gamma and Laplace samples we showed earlier.

For the Gamma sample we have:

```
Rows: 1
Columns: 2
$ sample_skewness        <dbl> 1.293772
$ sample_excess_kurtosis <dbl> -0.8356471
```

For the Laplace sample we have:

```
Rows: 1
Columns: 2
$ sample_skewness        <dbl> -0.07876743
$ sample_excess_kurtosis <dbl> 3.726374
```

Activity: Assessing Normality

For each of the following distributions, construct a sample of size 300 and compute their associated sample Skewness and Excess Kurtosis. After that, construct and compare their associated QQ plots and histograms.

1. Exponential distribution with rate parameter $\lambda = 1$.
2. Uniform distribution on the interval $[0, 1]$.
3. Beta distribution with shape parameters $\alpha = 2$ and $\beta = 5$.
4. Cauchy distribution with location parameter $x_0 = 0$ and scale parameter $\gamma = 1$.
5. Chi-squared distribution with degrees of freedom $k = 3$.