QQPlots and Normality Testing

MA2003B Application of Multivariate Methods in Data Science

Team Members:

Raul Gomez
Adriana Reyna

August 12, 2025

1 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 section, 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. But before starting, let's load the necessary libraries and set some options.

```
library("tidyverse")
library("here")
library("patchwork")
library("krulRutils")
library("ISLR2")
library("magrittr")

options(scipen = 999)
```

1.1 Skewness

Definition 1.1. 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. A perfectly symmetric distribution has a skewness of zero. To illustrate this idea, we will introduce the gamma distributions.

Definition 1.2. Let X be a continuous random variable following a *Gamma distribution* with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. The probability density function (PDF) of X 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.

Remark 1.3. 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.

We will now show how to plot the PDF of the Gamma distribution for $\alpha = 2$ and $\beta = 1$, highlighting the mean and median.

```
# Parameters
alpha <- 2
beta <- 1

# Create sequence of x values
x_data <- seq(0, 10, length.out = 200)

# Calculate density values
gamma_distribution_tbl <- tibble(
  x_data = x_data,
  density = dgamma(x_data, shape = alpha, rate = beta)
)

# Calculate mean and median
mean_val <- alpha / beta
median_val <- qgamma(0.5, shape = alpha, rate = beta)

# Plot
gamma_distribution_tbl %>%
ggplot(aes(x = x_data, y = density)) +
  geom_line(
    color = c_palette["C blue"],
    linewidth = 1
  ) +
  geom_vline(
    xintercept = mean_val,
    color = c_palette["C red"],
    linetype = "dashed",
    linewidth = 1
```

```
) +
geom_vline(
  xintercept = median_val,
  color = c_palette["C green"],
  linetype = "dotted",
  linewidth = 1
) +
annotate(
  "text",
  x = mean_val,
  y = 0.35,
  label = "Mean",
  color = c_palette["C red"],
  hjust = -0.1
) +
annotate(
  "text",
  x = median_val,
  y = 0.35,
  label = "Median",
  color = c_palette["C green"],
  hjust = 1.1
) +
labs(
  title = paste(
    "Gamma Distribution PDF (shape =", alpha, ", rate =", beta, ")"
  ),
  x = "x",
  y = "Density"
) +
theme_krul()
```

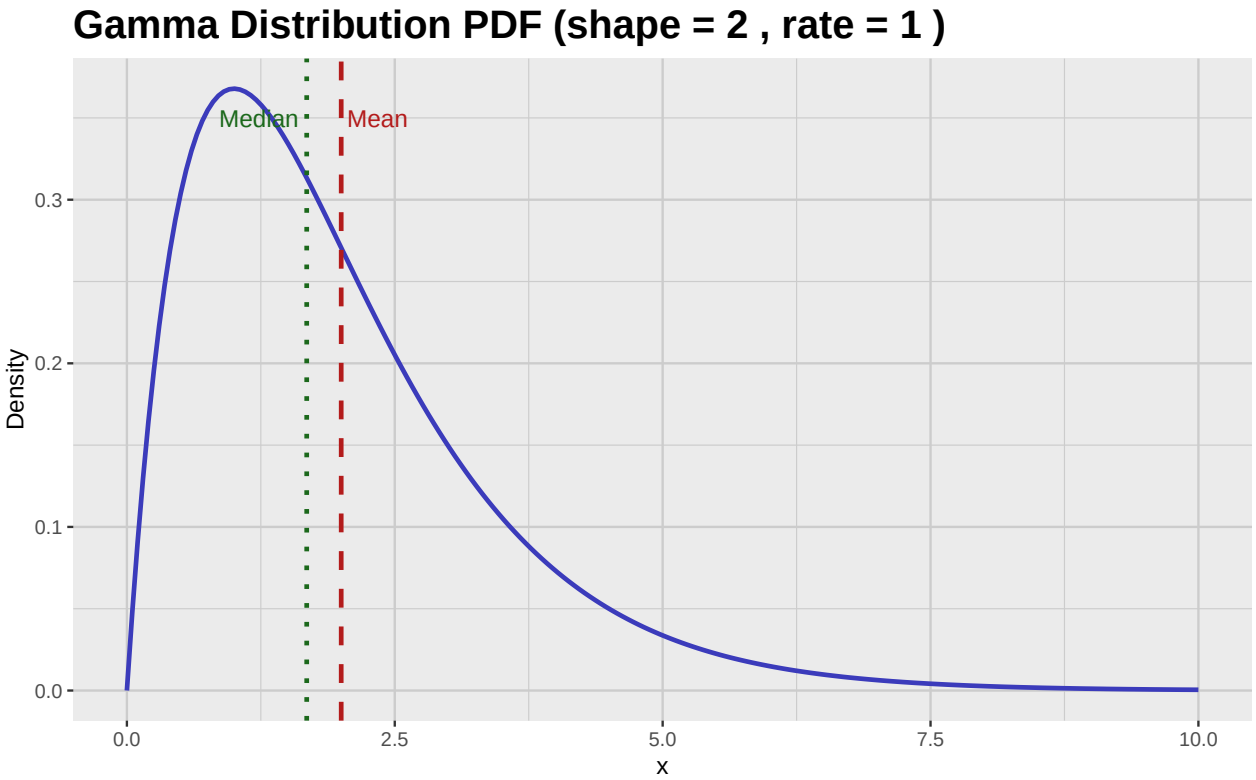

Figure 1: PDF of the gamma distribution with mean and median highlighted.

1.2 Kurtosis

Definition 1.4. The *kurtosis* of a random variable X is defined as the fourth standardized moment, given by

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4},$$

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 (also known as a *leptokurtic* distribution) has heavier tails and a sharper peak than a normal distribution, while a distribution with low kurtosis (also known as a *platykurtic* distribution) has lighter tails and a flatter peak. The kurtosis of a normal distribution is 3, and it is often adjusted to be 0 for comparison purposes. To illustrate this idea, we will introduce the Laplace distribution.

Definition 1.5. Let X be a continuous random variable following a *Laplace distribution* with location parameter $\mu \in \mathbb{R}$ and scale parameter $b > 0$. The probability density function (PDF) of X 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.

Remark 1.6. The kurtosis of a Laplace distribution is given by

$$\gamma_2 = 6.$$

This means that the Laplace distribution has heavier tails than a normal distribution, which has a kurtosis of 3.

We will now show how to plot the PDF of the Laplace distribution with mean zero and variance one.

```
# Parameters
mu <- 0      # mean (location)
b <- 1/sqrt(2)       # scale

# Sequence of x values covering enough range to see tails
x_data <- seq(-5, 5, length.out = 300)

# Laplace PDF function
dlaplace <- function(x, mu, b) {
  return(1/(2*b) * exp(-abs(x - mu)/b))
}

# Create data frame with PDF values
laplace_distribution_tbl<- data.frame(
  x_data = x_data,
  density = dlaplace(x_data, mu, b)
)

# Plot PDF
laplace_distribution_tbl %>%
  ggplot(aes(x = x_data, y = density)) +
  geom_line(color = c_palette["C blue"], linewidth = 1) +
  labs(
    title = "Laplace Distribution PDF (mean = 0, sd = 1)",
    x = "x",
    y = "Density"
  ) +
  theme_krul()
```

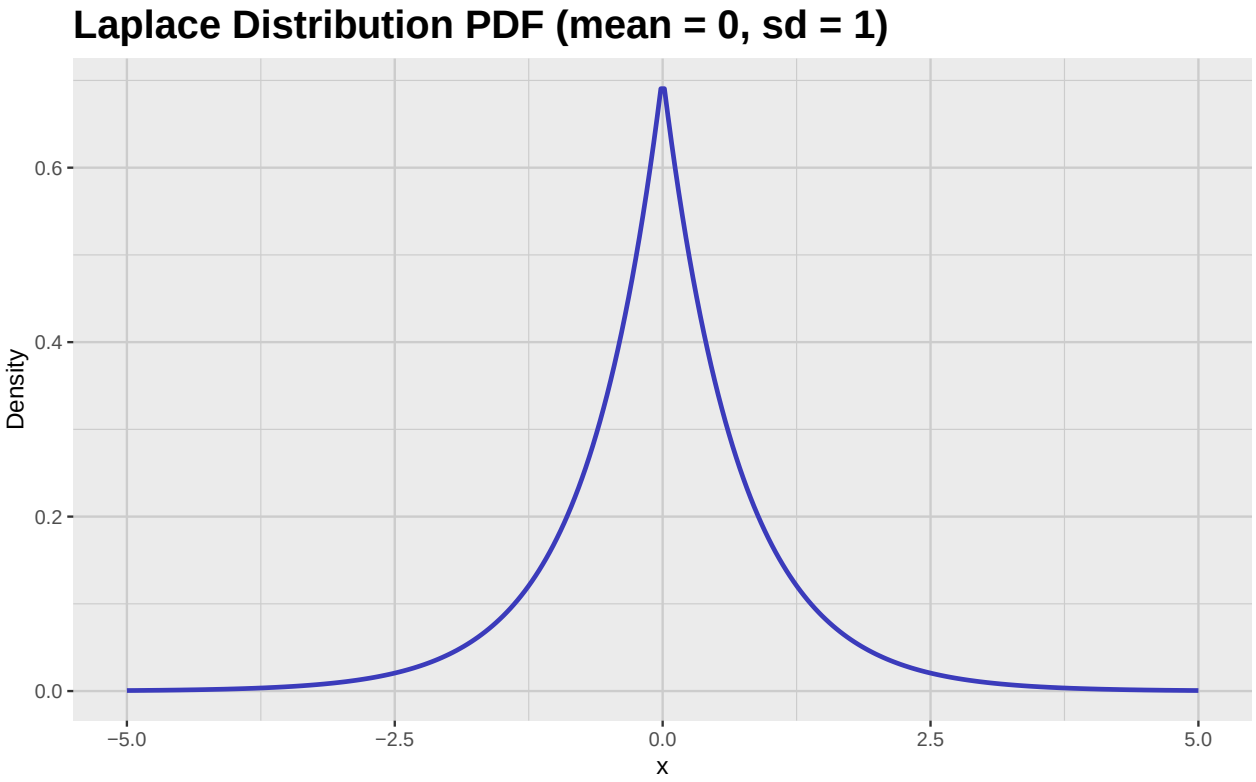

Figure 2: PDF of the Laplace distribution with mean zero and variance one.

1.3 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 whether a dataset follows a normal distribution.

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

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, 2, \ldots, n.$$

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.

To illustrate these ideas, we will now generate samples from the Gamma and Laplace distributions, and then we will construct thier associated QQ plots and histograms. First we construct the samples:

```
set.seed(123)  # for reproducibility

# Sample size
n <- 300
# Generate random samples from the Gamma distribution
gamma_samples <- rgamma(n, shape = alpha, rate = beta)

gamma_samples_tbl <- tibble(
  sample = gamma_samples
)

# Generate uniform random numbers in (0,1)
u <- runif(100)

# Apply inverse CDF of Laplace
laplace_sample <- ifelse(u < 0.5,
                         mu + b * log(2*u),
                         mu - b * log(2*(1 - u)))

laplace_samples_tbl <- tibble(
  sample = laplace_sample
)
```

We will now construct the corresponding QQ plots and histograms. We will start with the Gamma distribution samples:

```
gamma_qq_plot <- gamma_samples_tbl %>%
  ggplot(aes(sample = sample)) +
  geom_qq(
    color = c_palette["C red"],
    size = 0.3
```

```
  ) +
  geom_qq_line(
    color = c_palette["C blue"],
    linewidth = 0.5
  ) +
  labs(
    title = "QQ Plot",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_krul()

gamma_histogram <- gamma_samples_tbl %>%
  ggplot(aes(x = sample)) +
  geom_histogram(
    bins = 30,
    fill = c_palette["C blue"],
    color = "black",
    alpha = 0.7
  ) +
  labs(
    title = "Histogram",
    x = "Sample Values",
    y = "Density"
  ) +
  theme_krul()

gamma_plot <- gamma_qq_plot + gamma_histogram +
  plot_layout(ncol = 2) +

  plot_annotation(
    title = "Gamma Distribution: QQ Plot and Histogram",
    theme = theme(
      plot.title = element_text(
        size = 24,
        face = "bold"
      )
    )
  )

gamma_plot
```

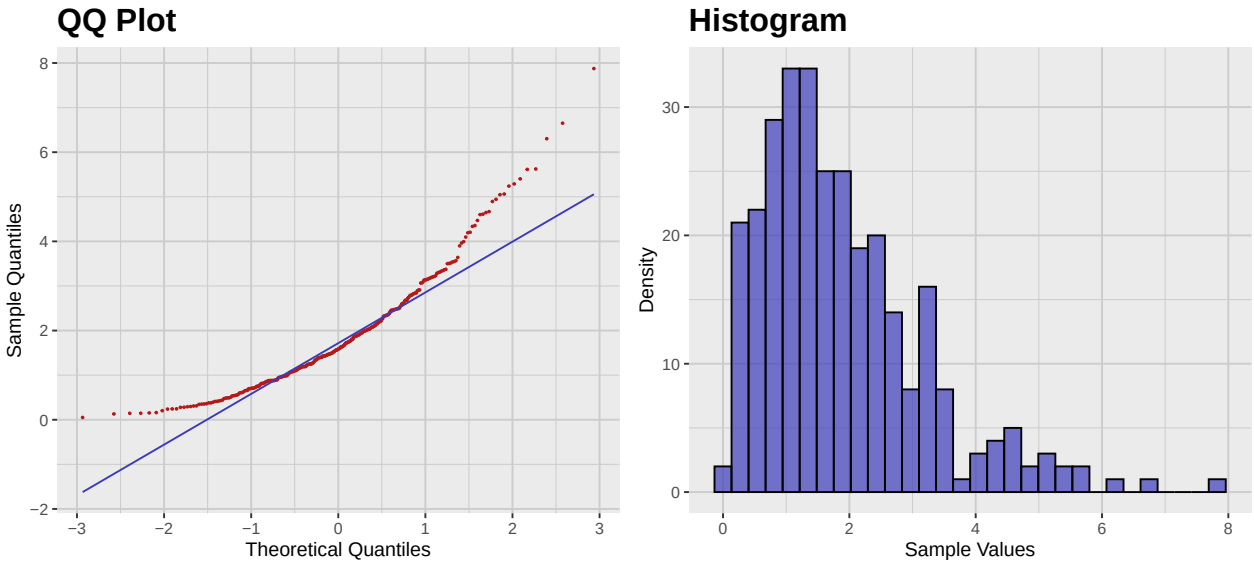

Figure 3: QQ plot and histogram of the Gamma distribution samples.

We will now construct the QQ plot and histogram for the Laplace distribution samples:

```
laplace_qq_plot <- laplace_samples_tbl %>%
  ggplot(aes(sample = sample)) +
  geom_qq(
    color = c_palette["C red"],
    size = 0.3
  ) +
  geom_qq_line(
    color = c_palette["C blue"],
    linewidth = 0.5
  ) +
  labs(
    title = "QQ Plot",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_krul()

laplace_histogram <- laplace_samples_tbl %>%
  ggplot(aes(x = sample)) +
  geom_histogram(
    bins = 30,
    fill = c_palette["C blue"],
```

```
    color = "black",
    alpha = 0.7
  ) +
  labs(
    title = "Histogram",
    x = "Sample Values",
    y = "Density"
  ) +
  theme_krul()
laplace_plot <- laplace_qq_plot + laplace_histogram +
  plot_layout(ncol = 2) +

  plot_annotation(
    title = "Laplace Distribution: QQ Plot and Histogram",
    theme = theme(
      plot.title = element_text(
        size = 24,
        face = "bold"
      )
    )
  )
laplace_plot
```

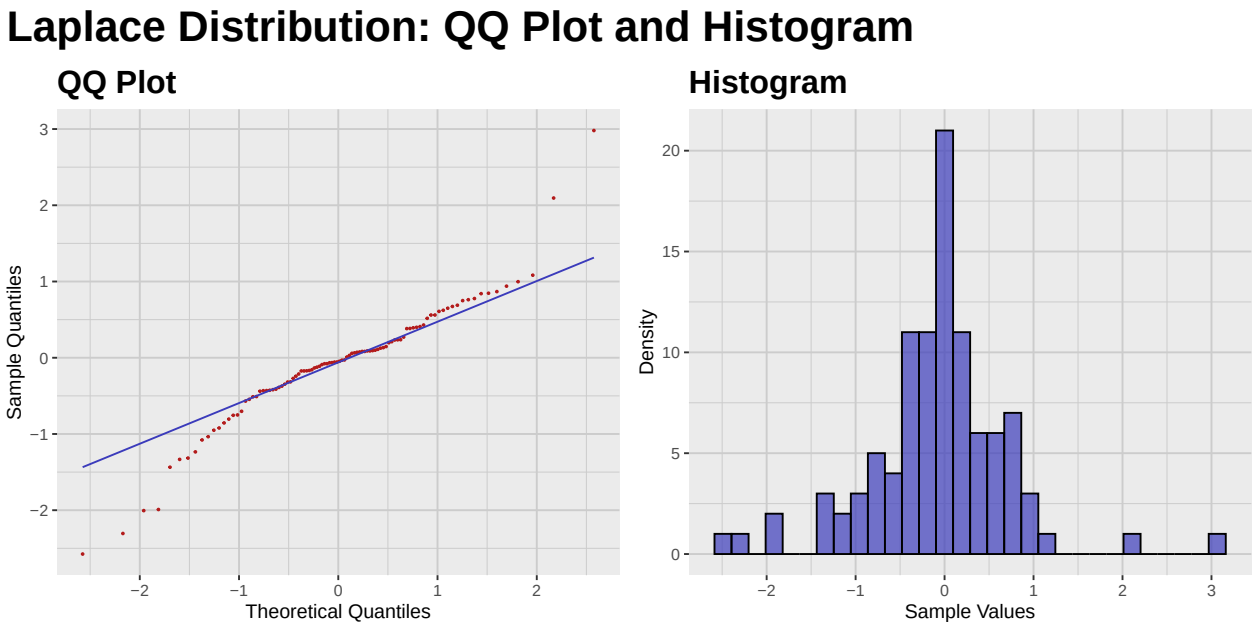

Figure 4: QQ plot and histogram of the Laplace distribution samples.

1.4 Exercises

Construct the QQ plot and histogram for the following distributions:

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 $k = 3$.