

Visualizing Probability Distributions

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics
University of Minnesota



August 28, 2020

Copyright © 2020 by Nathaniel E. Helwig

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Estimating the CDF

The empirical cumulative distribution function (ECDF) is a simple and powerful approach for estimating the CDF.

Given an independent and identically distributed (iid) sample of data x_1, \dots, x_n from some distribution F , the ECDF is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

where $I(\cdot)$ is an indicator function, i.e., $I(x_i \leq x) = 1$ if $x_i \leq x$ and $I(x_i \leq x) = 0$ otherwise.

ECDF Properties

The ECDF simply calculates the proportion of observations in the sample that are less than or equal to the input x .

Since $\hat{F}_n(x)$ is a proportion estimate, we have that

$$E\left(\hat{F}_n(x)\right) = F(x) \quad \text{and} \quad \text{Var}\left(\hat{F}_n(x)\right) = \frac{1}{n}F(x)(1 - F(x))$$

which implies that $\hat{F}_n(x)$ is an unbiased estimate of $F(x) = P(X \leq x)$.

Furthermore, as the sample size gets large, i.e., as $n \rightarrow \infty$, we have that $\hat{F}_n(x) \xrightarrow{d} F(x)$, which is known as the Glivenko-Cantelli theorem.

ECDF Visualizations

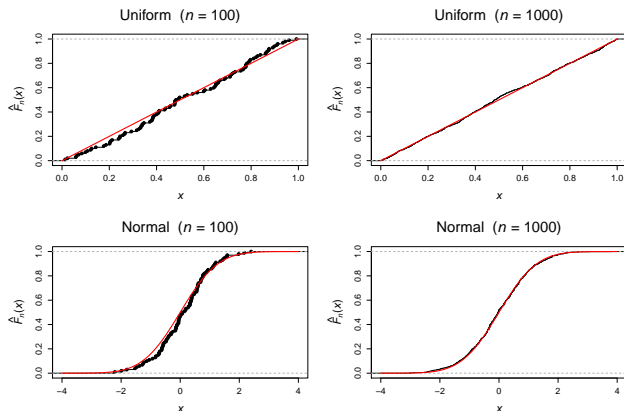


Figure 1: ECDF for $n \in \{100, 1000\}$ samples drawn from a $U[0, 1]$ distribution (top) and a $N(0, 1)$ distribution (bottom). The black dots denote the ECDF and the red line denotes the true CDF for each distribution.

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Quantile Overview

Q-Q plots are used to plot sample quantiles against one another or against population quantiles.

Such plots can be useful for assessing whether

- one sample of data follows a particular distribution
- two samples of data have a similar distribution

As a reminder, the population quantile function $Q(p)$ is the inverse of the CDF function, such that it takes in a probability $p \in [0, 1]$ and returns a value $x \in S$ such that $F(x) \geq p$.

Order Statistics and Sample Quantiles

Given an iid sample of data x_1, \dots, x_n from some distribution F , the order statistics are

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$$

which is simply the sample of data sorted from smallest to largest.

For convenience of notation, let's assume that the observations are sorted from smallest to largest, so that $x_i = x_{(i)}$ for $i = 1, \dots, n$. The sample quantiles are defined as

$$\hat{Q}_n(p) = x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor) (x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$$

where the value of h depends on what interpolation scheme is used to estimate the quantiles.

Two Uses of Q-Q Plots

Two ways in which Q-Q plots are typically used:

- If you have a single sample of data, it is typical to plot the theoretical quantiles $Q(p)$ on the x-axis and the sample quantiles $\hat{Q}_n(p)$ on the y-axis.
- If you have two samples of data with sizes m and n , it is typical to plot the sample quantiles of the first sample $\hat{Q}_m^1(p)$ on the x-axis and the sample quantiles $\hat{Q}_n^2(p)$ on the y-axis.

In both cases, having the points fall on the 45-degree line indicates that the two sets of plotted quantiles reasonably agree with one another.

Interpreting Q-Q Plots

Any deviations from the 45-degree line can provide graphical insights into how the quantiles differ from one another.

In the following example, note the following:

- for left-skewed data, the Q-Q points fall below the 45-degree line
- for right-skewed data, the Q-Q points fall above the 45-degree line
- for leptokurtic data, the points fall below the 45-degree line for negative values and above the 45-degree line for positive values
- for platykurtic data, the points fall above the 45-degree line for negative values and below the 45-degree line for positive values

Example Q-Q Plots

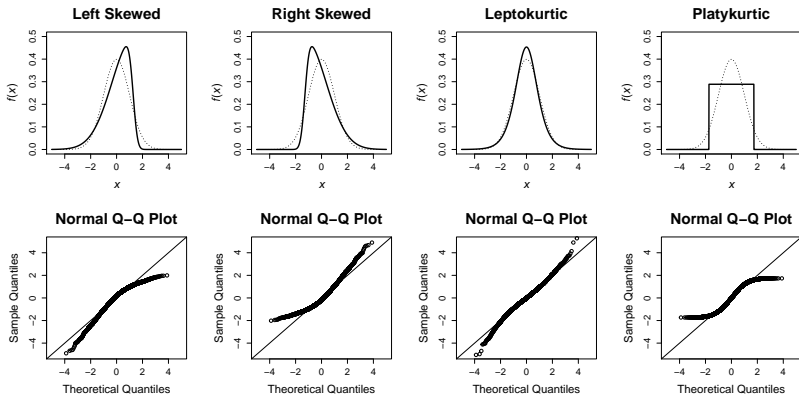


Figure 2: Top: probability density functions for distributions with different values of skewness and kurtosis (solid line), along with the standard normal density function (dotted line). Bottom: corresponding normal Q-Q plots with theoretical quantiles from a standard normal. Calculated using 10,000 independent samples.

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Boxplot Properties

A standard box plot consists of a few different components:

- a rectangle to denote the interquartile range, i.e., $IQR = Q_3 - Q_1$
- a line for the median, i.e., the second quartile Q_2
- whiskers on each end of the box plot to denote the data range

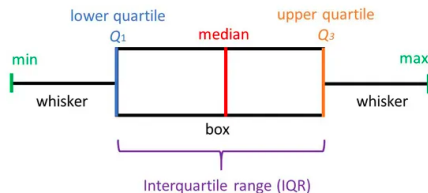


Figure 3: Properties of a box plot. From <https://www.simplypsychology.org/boxplots.html>

R's `boxplot()` function draws the whiskers to extend to $\pm 1.5IQR$.

Boxplot Visualization

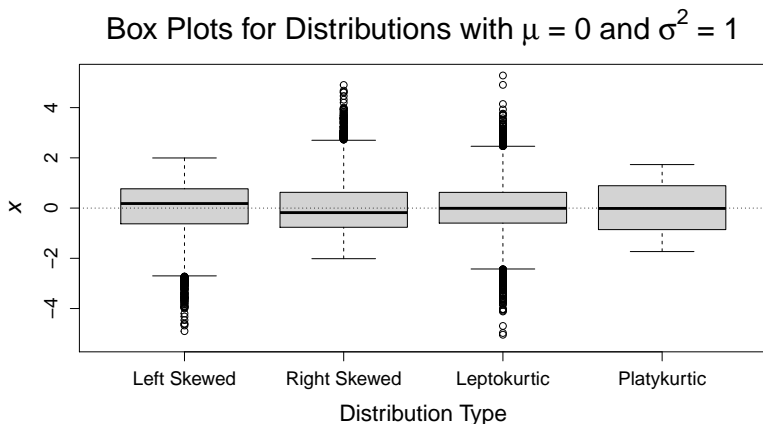


Figure 4: Box plots created with R's `boxplot()` function. The box plots were calculated using 10,000 independent samples from each distribution.

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Motivation for Histogram

If the PDF $f(x)$ is smooth, then we have that

$$\begin{aligned}P(x - h/2 < X < x + h/2) &= F(x + h/2) - F(x - h/2) \\&= \int_{x-h/2}^{x+h/2} f(z)dz \approx hf(x)\end{aligned}$$

where $h > 0$ is some small constant referred to as the “bin width”.

If the CDF $F(x)$ were known, then we could estimate the PDF using

$$\hat{f}(x) = \frac{F(x + h/2) - F(x - h/2)}{h}$$

but this isn't practical because we never know the true CDF $F(x)$.

Histograms in Practice

Plugging the ECDF estimate $\hat{F}_n(x)$ into the previous equation gives

$$\begin{aligned}\hat{f}_n(x) &= \frac{\hat{F}_n(x + h/2) - \hat{F}_n(x - h/2)}{h} \\ &= \frac{\sum_{i=1}^n I(x_i \leq x + h/2) - \sum_{i=1}^n I(x_i \leq x - h/2)}{nh} \\ &= \frac{\sum_{i=1}^n I(x_i \in (x - h/2, x + h/2])}{nh}\end{aligned}$$

Generally, we could estimate the PDF $f(x)$ in a window around x using

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n I(x_i \in w_j)}{nh} = \frac{n_j}{nh}$$

for all $x \in w_j = (b_j - h/2, b_j + h/2]$ where the $b_1 < b_2 \dots < b_{m+1}$ are chosen constants known as “break points”.

Choosing the Histogram Break Points

To form a histogram you just need to (i) break the real number line into m mutually exclusive bins at break points spanning your data, and (ii) count the number of observations n_j that fall within each bin.

Different choices of the number of bins m will affect the estimate

Different methods for choosing m and h for a histogram:

- Sturges (default in R's `hist()` function): $m = \lceil \log_2(n) + 1 \rceil$ and $h = (x_{(n)} - x_{(1)})/m$
- Freedman and Diaconis: $h = 2IQR/n^{1/3}$ and $m = \lceil (x_{(n)} - x_{(1)})/h \rceil$
- Scott: $h = 3.5s/n^{1/3}$ and $m = \lceil (x_{(n)} - x_{(1)})/h \rceil$ where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Histogram Examples

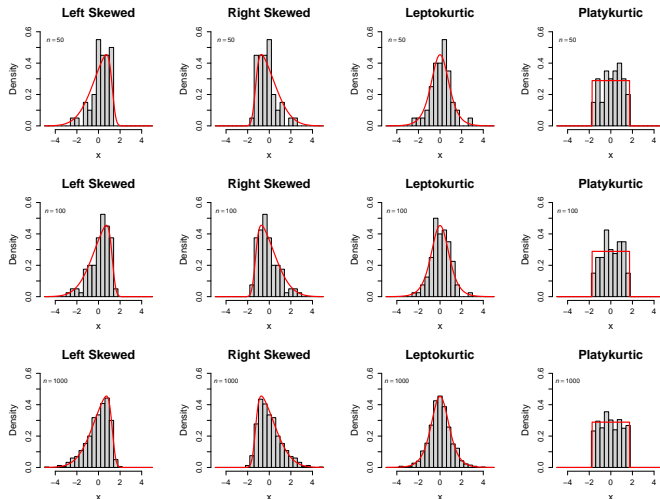


Figure 5: Created with R's `hist()` function. Red line denotes the true density.

Table of Contents

1. Empirical Cumulative Distribution Function
2. Quantile-Quantile (Q-Q) Plots
3. Boxplots
4. Histograms
5. Kernel Density Estimates

Improved Estimates of Densities

Histograms are simple to understand and create, but they provide rather crude (i.e., jagged) estimates of PDFs.

Given an iid sample of data x_1, \dots, x_n from some distribution F , a KDE has the form

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is a kernel function and $h > 0$ is the chosen bandwidth.

The kernel function $K(\cdot)$ can be any function that satisfies

- $K(x) \geq 0$ for all x (non-negative)
- $K(x) = K(-x)$ for all x (symmetric)
- $\int_{-\infty}^{\infty} K(x) = 1$ (unit measure)

Examples of Kernel Functions

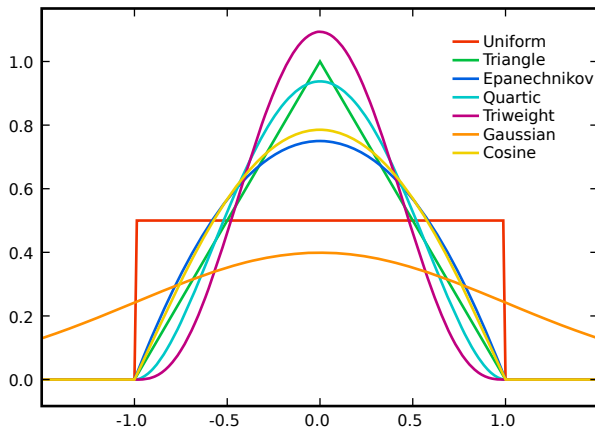


Figure 6: Different kernel functions. From <https://upload.wikimedia.org/wikipedia/commons/4/47/Kernels.svg>

Bandwidth Parameter

The bandwidth parameter h is analogous to the bin width parameter h in a histogram, such that different values of h will produce different estimates.

The bandwidth parameter controls the compactness of the kernel function, such that larger values of h use wider kernels

- As $h \uparrow$ the KDE gets smoother
- As $h \downarrow$ the KDE gets more jagged

It is typical to use Silverman's rule of thumb to define h , which has the form $h = 0.9n^{-1/5} \min(s, IQR/1.34)$ where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

- Using cross-validation is more ideal

Simple Example of KDE

Suppose that we $n = 6$ data points $(-2.1, -1.3, -0.4, 1.9, 5.1, 6.2)$, and we want to form a histogram and a KDE using a standard normal kernel function with $h = 1.5$.

- KDE centers a $N(0, 1.5^2)$ density at each data point x_i
- then calculates the average of the $N(x_i, 1.5^2)$ densities

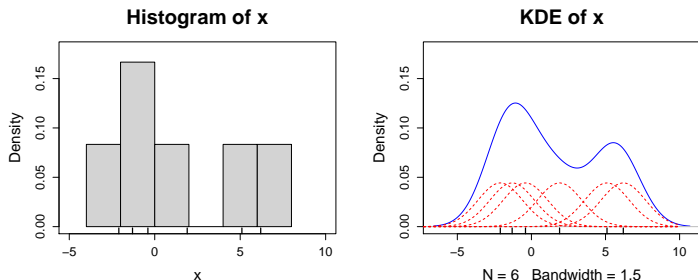


Figure 7: The red dashed lines are showing $\frac{1}{nh} K\left(\frac{x-x_i}{h}\right)$, which are summed together to obtain the blue line, which is the KDE.

More Examples of KDEs

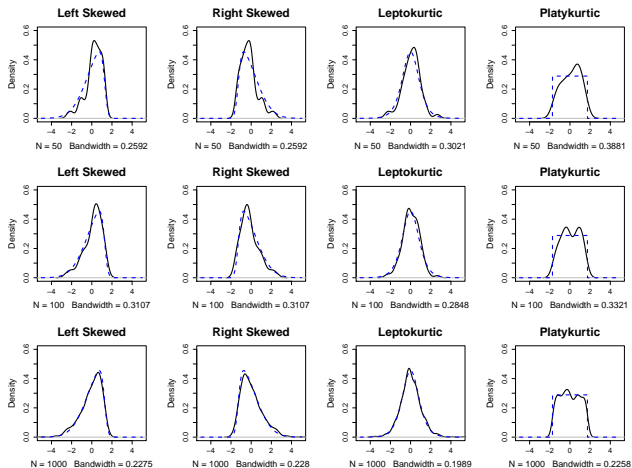


Figure 8: Kernel density estimates (KDEs) created with R's `density()` function. The blue dashed line denotes the true density.