

2.1 What is an empirical distribution?

Overview

What is an empirical distribution? The empirical distribution of a variable is able to be characterised by considering each unique numerical value observed for that variable in a given sample of size n . If certain values are repeated, then we simply tally the number of each unique value obtained. These tallies are essentially *raw frequencies* of the values. We can order the values obtained for the variable from smallest to largest and then look at these frequencies cumulatively, as a percentage of the entire sample. A plot of these *cumulative percentages* as a function of the ordered values in the sample is known as the *empirical cumulative distribution function*.

Description

More formally, suppose we have n independent and identically distributed random variables, $Y_1, Y_2, \dots, Y_n$, with a common (but unknown) probability density function (pdf) of $f(y)$ and cumulative distribution function (cdf) of $F(y) = \text{Pr}\{Y \leq y\}$.

For the discrete case, we have $F(y) = \sum_{t \leq y} f(t)$.

For the continuous case, we have $F(y) = \int_{-\infty}^y f(t) \, dt$.

We obtain corresponding observed values $y_1, y_2, \dots, y_n$ in a sample of size n . Now, let $I(y_i \leq t)$ be an indicator that takes the value of 1 if $y_i \leq t$ is true, and zero otherwise. The empirical cumulative distribution function $\hat{F}_n(t)$ is defined as the proportion of data points in the sample that are less than or equal to t , i.e.,

$$\hat{F}_n(t) = \frac{1}{n} \cdot \sum_{i=1}^n I(y_i \leq t)$$

This is therefore a step function, continuous from the right, that jumps up by a quantity of $1/n$ at each of the n data points. Its shape gives us a basic visual understanding of the distributional shape of the data values.

A small example

Suppose we had the following data with a sample size of $n = 10$ for a variable, Y :

Sample	y
1	2
2	7
3	10

Sample	y
4	12
5	2
6	5
7	5
8	8
9	10
10	3

The raw frequencies are as follows:

Value of y	Frequency
2	2
3	1
5	2
7	1
8	1
10	2
12	1

Looking at these values cumulatively, we have:

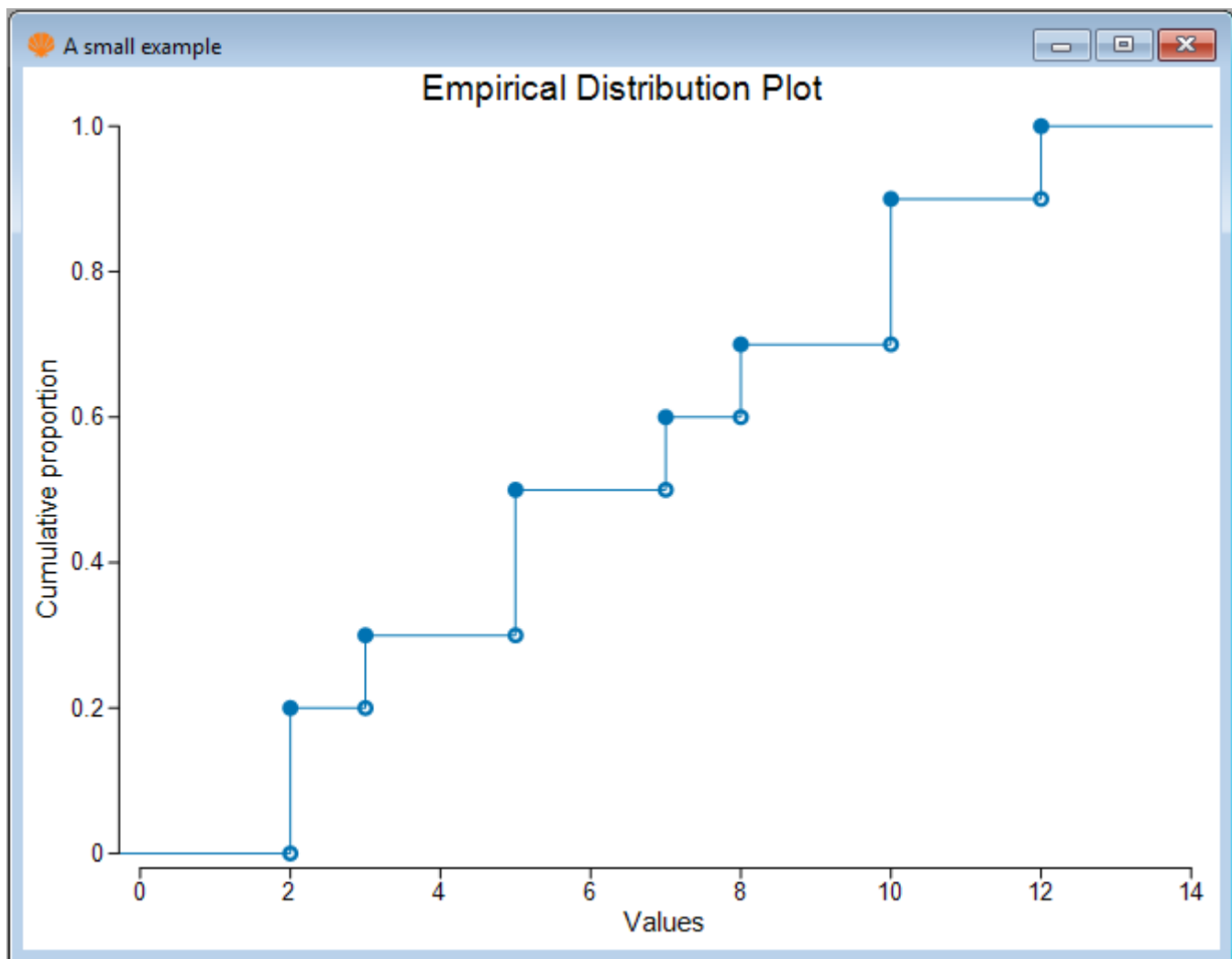
Value of y	Cumulative frequency
2	2
3	3
5	5
7	6
8	7
10	9
12	10

Expressing these frequencies as *cumulative proportions* of the total sample, we have:

Value of y	Cumulative proportion
2	0.2
3	0.3
5	0.5

Value of y	Cumulative proportion
7	0.6
8	0.7
10	0.9
12	1.0

These cumulative proportions comprise the **empirical cdf**. A plot of this empirical distribution (a step function) is shown below, with open circles being used to show the discontinuities (i.e., at the point of each step).



The above plot was obtained by running **Plots > Empirical Distribution Plot...** in PRIMER 8, and ticking the option to: '☒ Express as proportions: [0, 1]'. Note that you can alternatively look at an empirical cdf with the values expressed as percentages instead of proportions (i.e., a plot of $100 \times \hat{F}_n(t)$ versus t), which is the default.

As an aside, a related graphical tool for examining distributional shapes of variables is a **histogram**. A histogram is a plot of the raw frequencies (as bars on the y-axis) vs the empirical values (on the x-axis). For a histogram, we would typically pool together (sum) the raw frequencies into larger 'bin' sizes (instead of having one bin for every unique value in the dataset), which can be very useful if n is large. That's what the **Plots > Histogram Plot...** function in PRIMER 8 does. Once you get a histogram, you can also change the bin size by clicking **Graph > Special**.

Revision #46

Created 26 March 2025 20:27:45 by Marti

Updated 1 December 2025 21:47:36 by Marti