3.1 Plots of empirical densities

Suppose we have measured a given variable in each of several groups. To visualise the distributional shape of each collective set of sample values, we might consider creating several *histograms* - one for each group - but it then might be difficult to compare them with one another.

We might alternatively consider using a *box plot*. Although a box plot may do a good job of summarising certain features (the median, inter-quartile range and overall range) of the data in each group, it may not necessarily provide insights about the *shape* of the collective set of values obtained within each group. What if, for example, certain groups actually show a pattern of having more than one mode?

There are many ways that one might consider visualising or approximating the underlying *probability density function* (pdf)¶ of a random variable, either on its own or separately within groups. PRIMER 8 now offers two empirical non-parametric tools that can help to visualise the density (shape) of points along the number line, also permitting comparisons of those shapes across several groups. If we have a discrete random variable, a simple *dot plot* is an appealing approach, while for a continuous random variable, using kernel density estimation to produce a smooth *violin plot* might be desirable, although either of these tools can, in practice, be used for either type of variable.

- **Dot plots** - Dot plots are a very simple way to represent data. We place a dot for every data point at its appropriate location on the number line (y axis). Observations that have the same value are simply 'stacked' alongside one another (along the x axis) at that same (y) position. Thus, visually, an empirical distribution of the collective set of points effectively 'builds itself', point by point, along the number line. Dot plots in PRIMER also include a horizontal line to show the median value for each group of observations.
- **Violin plots** - Violin plots show the median and inter-quartile range (like a box plot), but they also provide a smooth empirical non-parametric *kernel density estimate* (kde) of the probability density function, which is mirrored horizontally.

Kernel density estimation

Violin plots require kernel density estimation of the pdf, so we shall describe kde briefly here as implemented in PRIMER. Core references for this technique are Rosenblatt (1956) and Parzen (1962) ; see also Silverman (1986) . Suppose we have n independent and identically distributed random variables, $Y_1, Y_2, \ldots, Y_n$, with a common (but unknown) probability density function (pdf) of $f(y)$, and in our sample we have a set of corresponding observed values $y_1, y_2, \ldots, y_n$. We can estimate the shape of the pdf by the following *kernel density estimator*:

$$ \hat{f}_ h(y) = \frac{1}{nh} \sum_{j=1}^n K \left(\frac{y - y_j}{h} \right) $$

where K is the kernel (a non-negative function) and $h>0$ is a smoothing parameter called a *bandwidth*.

There are a range of kernel functions commonly used. PRIMER 8 uses the standard normal kernel, so $K(y) = \phi(y)$, and ϕ is the standard normal density function, hence:

$$ \hat{f}_ h(y) = \frac{1}{nh}\cdot\frac{1}{\sqrt{2\pi}} \sum_{j=1}^n \text{exp} \left(\frac{ -(y - y_j)^2 }{ 2h^2 } \right) $$

Choice of bandwidth

The bandwidth controls the degree of smoothing. The greater the bandwidth, the greater the degree of smoothing and, hence, the less important any individual data point will appear to be in producing the resulting kde function (a smooth line on the plot). A simple and widely used choice of bandwidth is obtained using Silverman's rule-of-thumb (Silverman (1986)). In PRIMER, the default is to apply Silverman's rule to calculate a suitable bandwidth separately for each group.

Suppose we have $i = 1, \ldots, g$ separate groups of observations, and the i^th^ group has a sample size of n_i and a within-group sample standard deviation of s_i. Silverman's rule to calculate a bandwidth h_i for group i (i.e., for each 'violin' being shown in the plot), is:

$$ h_i = 0.9 \cdot \text{min} \left(s_ i, \frac{\text{IQR}_ i}{1.34} \right) \cdot n_i^{-1/5} $$

where $\text{IQR}_ i$ is the inter-quartile range of group i.

Alternatively, one also has the option in PRIMER to type in manually a custom bandwidth for each group. This manual tool is also handy to use if you want all groups to have the same bandwidth. Note that, if n_i < 2 for any group, then an exception is thrown and a warning is issued stating that there are too few points to calculate Silverman's rule of thumb. In such cases (where there is a single data point), a custom bandwidth must be specified.

Re-scaling

PRIMER offers the following options for *re-scaling the widths* of the 'violins' (i.e., the relative 'heights' of the pdfs) produced in the plot:

- **None**: No rescaling is applied.
- **Area**: All violins are rescaled according to the *global* maximum density (i.e., every point in the violin is divided by the global maximum density obtained in any group).
- **Width**: Each violin is rescaled according to its *own group's* maximum density (i.e., every point in each violin is divided by the maximum density of its own group).
- **Count**: Each violin is rescaled in proportion to how many data points it has; specifically, every point in each violin is divided by its own maximum density and then multiplied by n_i/N, where $N = \sum_{i=1}^g n_i$^†^.

By default, PRIMER re-scales the densities in the violin plots by *area* (the global maxium density) so that every group has a density (area) that scales to a constant (as all pdfs integrate to 1.0, regardless of the sample size of each group). In contrast, densities that are scaled by *count* will have widths that will depend on their sample size.

Trimming

One consequence of placing a small normal distribution onto every data point and then summing and smoothing the resulting curves (as is done by any kde) is that the 'tails' of the plot (minimum and maximum values of the violin) will naturally exceed the empirical range of the data itself. This is not inappropriate, in general, because indeed the purpose of the kde is to give us an (albeit entirely empirical) estimate of a smooth probability function from which our observed data may have been drawn. Nevertheless, these 'tails' may appear illogical in practice. For example, if we have a random variable that is strictly non-negative (such as the biomass of a particular species), then it might be disconcerting to see the lowest values of the violin plot descend below zero. Clearly, we would never observe a biomass less than zero.

One of the options in PRIMER, therefore, is to permit *trimming* of the violins. One can choose to trim any (or all) of the violins at some set lower and/or upper value(s). It should be noted, however, that doing this kind of 'trimming' rather fundamentally changes the interpretation of the violin plot. The resulting shapes can no longer be considered to represent probability densities (pdfs) *per se*, but rather should be considered purely as visual representations of the general distributional shape of the underlying set of sample points in each group - like a kind of smoothed dot plot.

For data of type 'Abundance' or 'Biomass', PRIMER will, by default, trim the violins at a lower bound of zero, but will not trim by any upper bound. For percentage (e.g., cover) data, one might consider trimming the violins at a lower bound of 0 % and an upper bound of 100 %. For any other data type, the default in PRIMER is not to do any trimming.[‡]

[¶]*Or, in the case of a discrete random variable, a probability mass function (pmf).*

[†]*Note that n_i/N is just:*

(the number of data points making up the violin)/(the total number of data points across all violins).

[‡]*Recall that one identifies the '*Data type*' for a given dataset upon import as being one of '*Abundance*', '*Biomass*', '*Environmental*', or '*Unknown/other*'. For existing data (e.g., any PRIMER 8 example data files), you can always click Edit > Properties to see and/or alter the data type.*'

Revision #100
Created 27 March 2025 02:51:15 by Marti
Updated 16 November 2025 22:39:55 by Marti