Slide 1

Slide1

Hello everyone! Hope you are well and this unprescedented situation will be over us soon. My name is Sayani Gupta. I am pursuing my PhD at the Department of Econometrics and Business Statistics at Monash University, Australia and today I will be talking to you about highest density regions and how to go about plotting them in the ggplot2 framework. This is an R package that was developed during rOpenSci 2019 held at Sydney, Australia.

Slide 2

There are several ways to summarize a distribution using both Kernel density estimates and descriptive statistics. Descriptive statistics based displays include box plots or different variations of it like notched box plots, letter-value box plots or quantile plots. Kernel density based plots like violin plots and ridge line plots provide more detailed information about the shape, skewness, nature of tail or multimodality of the distribution. The descriptive statistics based methods do not allow us to see all these features all at once with so much clarity as the former, but they do avoid clutter and help us to focus on some specific properties of the data and hence are very desirable for exploratory purposes.

Slide 3

Summarizing a distribution tend to put forward different features of the distribution and hides other ones and hence it is often useful to display them in unison. Most statistical methods involve summarizing a probability distribution by a region of the sample space covering a specified probability. For example, in a box plot, the central box is bounded by Q1 and Q3, which represents the interquartile range. - The whiskers extend from Q1 - 1.5 (Q3 - Q1) to Q3 + to 1.5 (Q3 - Q1), which represents 99% coverage for large samples. Finally, the horizontal line represents the median. But this visualisation technique limited our ability to see multimodality in distributions.

Slide 4

Rob Hyndman in his paper "Computing and Graphing Highest Density Regions" in 1996 proposed methods to compute and display highest density regions.

Slide 5

Now what are highest density regions?

Highest density regions are one of the the most appropriate subsets to use to summarize a probability distribution. The criterion of this region is - 1. The region should occupy the smallest possible vol- ume in the sample space; 2. Every point inside the region should have probability density at least as large as every point outside the region.

The formal definition suggests that: HDRs could consists of disjoint regions and the mode is contained in every HDR.

something # Slide 6

Rob Hyndman in his R package hdrcde has already implemented plotting highest density regions in one and two dimensions. The method of summarizing a distribution using highest density regions are useful for analyzing multimodal distributions. We illustrate this by exploring the data set `faithful` which contains the waiting time and duration of eruptions for the old faithful geyser in the Yellowstone National Park, USA. Clearly, boxplot does not give any indication about the multimodality of the distribution.

Slide 7

We can use HDR boxplot to display the same variable. Along with displaying the 99% and 50% highest density regions, it also shows the local mode in each of the regions. This shows that eruption times are likely to be aroud 4.5 minutes or 2 minutes but rarely for around 3 minutes. This insight was not apparent through boxplot.

Similarly, HDRs can also be represented through a density plot and marking the higest density regions. For two variable, a HDR scatterplot can produce points that are coloured according to the bivariate HDRs in which they fall and a HDR conditional density plots highest density regions for a conditional density estimate.

Slide 8

While all of these are already great implementation, as I already stated earlier that it is often useful to display these summary plots in unison to have more involved perspectives about the data. HDR is not yet implemented in the ggplot2 framework, which provides excellent flexibility in terms of adding new elements to a display and customisation. Hence, we decided to extend the functionality of ggplot2 to be able to use HDR in the ggplot2 framework.

Slide 9

Hence, we started with R package gghdr.

The key elements of any graphic made in ggplot2 are geoms and stats. The "stats" component involves the statistical transformation of the data the "geoms" are defined through a class defined by ggproto function to specify the number of attributes and functions that describe how data should be drawn on a plot. For example, for a simple box plot which displays a compact distribution with median, quartiles, hinges and extreme outliers, the statistical transformations include the five number summary and the geometries are the lines, boxes or points used to represent them.

Slide 10

We create an object of the ggplot class, typically specifying the data and some or all of the aesthetics; Add on geoms and other elements to create and customize the plot, using +. We can keep adding one or many geoms and other elements to create plots that range from very simple to very customized. In package gghdr, we have built different geoms for graphing HDRs. For example, the code here on the right side uses the geom_hdr_boxplot function to draw HDR boxplots.

Slide 11

Similarly, geom_hdr_rug and hdr_bin could be used to plot HDR marginal distributions and scatter plots respectively. Thus, creating these geoms enable us to add layers to ggplot2 objects.

Slide 12

With ggplot2, we have immense flexibility adding different elements or layers to the plot. Here, I have superimposed a jitter plot on HDR box to supplement the insight drawn from the HDR boxplot.

Slide 13

Similarly, by adding the marginal distribution of HDR of the two variables in the scatterplot. Since both the modes lie in the 50% HDR, it implies that there is bimodality in both the variables.

Slide 14

The authors of this package are Mitch, Stephen, Ryo, myself, Thomas and Emi. We all met during rOpensci and developed this package.

Slide 15

For more information on the package, please visit the github page or the vignette. The slides are created using Rmarkdown, /knitr, xaringan and xaringanthemer. Thank you so much for listening.