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 which display quantiles instead of quartiles in a traditional boxplot. Kernel density based plots for displaying a distribution could be violin plots and ridge line plots. Kernel density estimates provide more detailed information about the distribution but depend on smoothing parameters and hence can vary for the same data set. The descriptive statistics based methods do not allow us to see the shape, skewness, nature of the tail or multi-modality 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. Also, the summaries based on descriptive statistics remain free of any tuning parameters or Kernels, which is sometimes more desirable for exploration.

Choice of plots are dictated by the statistical transformations and geometric objects used for the visualization. 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 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. For example, although boxplots are really cool compact displays, they are not useful for displaying multimodal distributions as you can see from these plots.

Slide 4

Highest density regions in these cases are one of the the most appropriate subsets to use to summarize a probability distribution. There are several statistical methods to summarize a distribution by region of the sample space covering certain probability. For example, in a traditional boxplot, the central box bounded by the interquartile range represents 50% coverage and whiskers represents 99% coverage for large samples. The method of summarizing a distribution using highest density regions are useful for analyzing multimodal distributions.

Explain HDR

Slide 5

The R package hdrcde authored by Rob Hyndman provides tools for computing highest density regions in one and two dimensions. 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.

The HDR boxplot displaying the 99% and 50% highest density regions also shows the local mode in each of the regions. For this example, it implies that the eruption times are likely to be around 4.5 minutes or 2 minutes but rarely for around 3 minutes. This insight would not have been apparent through a boxplot.

A HDR scatterplot produces a scatterplot where the points 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 5

As I stated earlier that it is often useful to display such 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 6

Hence, we started with R package gghdr. The key elements of any graphic made in ggplot2 are geoms and stats. The "stats" which involves the statistical transformation of the data is inherited from the R package "hdrcde" and 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.

Slide 7

In package gghdr, we have different geoms for graphing HDRs. For example, the code here on the right side uses the geom_hdr_boxplot function to draw HDR boxplots. Similarly, geom_hdr_rug and hdr_bin could be used to plot HDR rug plots and scatter plots respectively. As you could see all these codes are very similar to any ggplot2 code.

Slide 8

With ggplot2, we have immense flexibility adding different elements or layers to thr plot. Here, I have plotted HDR box and jitter

Slide 9

Explain what you see by combining these plots.

Slide 10

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

Slide 11

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