Slide1

Hello everyone! My name is Sayani Gupta. I am pursuing my PhD at the Department of Econometrics and Business Statistics at Monash University, Australia. Hope you all are well and this unprecedented situation will be over us soon. 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 in 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 its different variations, letter-value box plots or quantile plots. Kernel density based plots like violin plots and ridge line plots provide 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 or regions of the sample space and hence are very desirable for exploratory purposes.

The R package ggdist provides a flexible set of ggplot2 methods to visualize distributions and uncertainty.

Slide 3

Summarizing a distribution tend to expose different features of the distribution 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 bounded by Q1 and Q3 typically covers 50% and the whiskers cover 99% of the sample space for large samples. This visualization technique, however, limits our ability to see multimodality in distributions as could be seen through graphics here.

Slide 4

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

Slide 5

Now what are highest density regions?

As we see here, there can be several possible ways to cover a specified probability in the sample space. In this plot, all of these regions cover 75% Probability Regions, but only the highest density region shows the bimodality. Thus the principle of highest density 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. Hence the method of summarizing a distribution using highest density regions are useful for analyzing multimodal distributions.

Slide 6

Rob Hyndman in his R package hdrcde has already implemented plotting 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. Clearly, boxplot does not give any indication about the bimodality of the distribution of eruptions, but it could be observed from the density plot.

Slide 7

Now let us use HDR boxplot to display the same variable. Along with displaying the 99% and 50% highest density regions, it also shows the local mode. 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 that we saw in the previous plots.

Similarly, HDRs can also be represented through a density plot and marking the highest density regions. This could be extended for two variable, with a HDR conditional density plotting continuous display of the density of eruptions times conditional on different bins of the waiting time or with a HDR scatter plot producing points that are colored according to the bivariate HDRs in which they fall.

Slide 8

While all of these are already great implementation, it is often useful to display these summary plots in unison to have more involved perspectives about the data and also having the flexibility to customise them. ggplot2 has become the de-facto standard visualization package in R because of the excellent flexibility it provides in terms of adding new elements to a display and customization. It creates graphics based on The Grammar of Graphics. HDR is not yet implemented in the ggplot2 framework, and 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 summary statistics and the geometries are the lines, boxes or points used to represent them.

Slide 10

The required components of creating a ggplot2 object is typically specifying the data, some or all of the aesthetics and adding at least one layer to render the data and aesthetics to the screen. These layers typically take the form of a ggplot2 geom functions. And then we can keep adding elements and geoms to customize the plot, using +. 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. One exciting feature `geom_hdr_boxplot()` supports (unlike `hdr.boxplot`) is the ability to create separate HDR boxplots for different groups in a dataset.

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

We can keep adding one or many geoms and other elements to create plots to add flexibility and customization. Here, I have superimposed a jitter plot on HDR box to supplement the insight drawn from the HDR boxplot.

Slide 13

Similarly, by adding different layers we are able to visualize both marginal and bivariate highest density regions and see that there is multimodality in both.

Slide 14

The authors of this package are Mitch, Stephen, Ryo, Darya, myself, Thomas and Emi. This was a great team and it was a pleasure working with them. 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.