Working through Distribution Plots

Josh Peters

2019-03-20

Recapping the cheatsheet

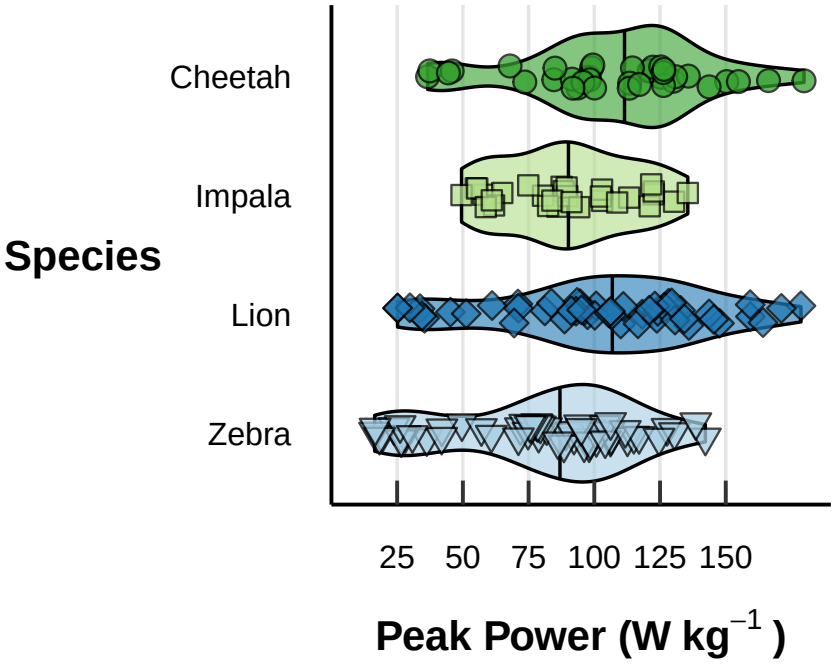
The distributions plotting cheatsheet aims to provide a quick idea about the necessary elements and considerations for plotting distributions and comparisons of one variable data across distinct categories. Common examples in the biological field include RNA and protein levels across samples and replicates.

We first looked at this dataset, a summary of which is shown below. We used this data to generate several plots, which iteratively improved upon different plotting elements.

Table 1: Summary of the predatory/prey data

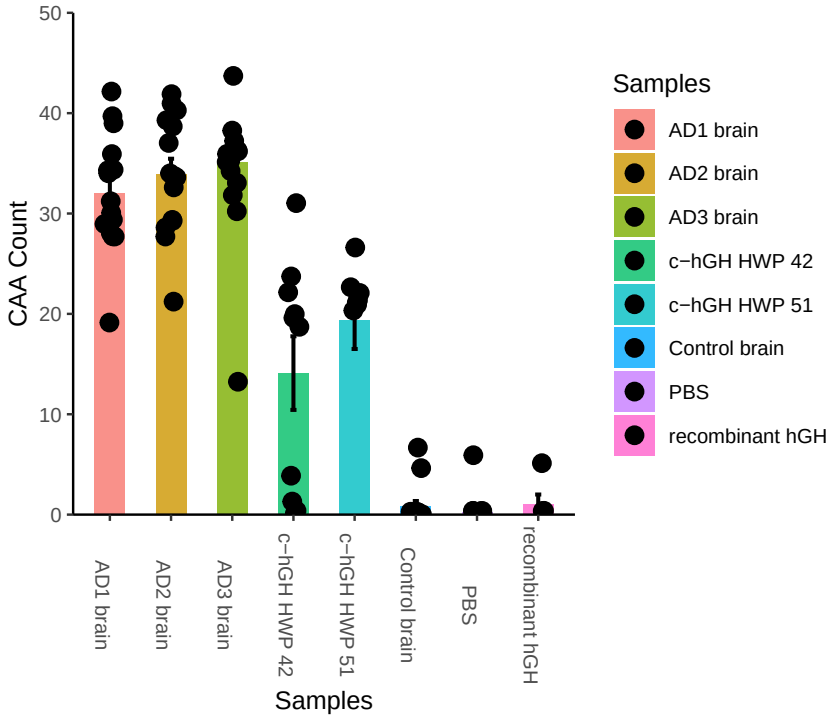
Species	N	mean	sd	se
Cheetah	37	107.77994	33.78020	5.553430
Impala	30	89.90193	24.89221	4.544675
Lion	50	102.44448	39.83166	5.633048
Zebra	57	79.67570	34.28224	4.540794

The cheatsheet produced this final plot, which chose to use a `violin plot` to display the data across the 4 species. What do you think are the biggest improvements with this plot, compared to previous iterations?



Below is a new plot.[1] First, critique the current plot using the checklist from the cheatsheet and attributes we improved upon throughout the plotting process.

[1] Data is from 1. Purro, S. A. et al. Transmission of amyloid-B protein pathology from cadaveric pituitary growth hormone. Nature 564, 415–419 (2018).



Take it one step further and use the data provided within this

repository to create a better plot. The markdown file used to create this worksheet is included. You can use the code provided as a starting point to practice plotting in R and fixing the plot above. Remember the checklist of items to consider, which focus on these areas of plotting:

1. How the data is encoded within the plot (e.g. figure choice, color)
2. How the data is described within the plot (e.g. axes, scales, grids)
3. Additional information needed to understand the data (e.g. statistics)