Distributions Walkthrough

MIT Biological Engineering Communication Lab

Understanding your data

- How many points do you have?
- What does your data look like?
- What is your message?

Example: "4 drug conditions, n = 3 mice per condition, each condition has small ranges"

▶ **Condition 1 results in longer survival.**

Choosing a representation

Bar graph
Discrete, n < 10
***must also show points**

Strip Plot
Discrete, n < 30

Boxplot
Discrete/continuous, n > 10
****should also show points***

Violin plot
Continuous, n > 30
****can also show points***

Density Plots
Discrete/continuous, n > 30

Elements checklist

- ☐ Plot + data labels (label directly if possible)
- ☐ Appropriate axes scales and tick marks
- ☐ Gridlines if desired and needed
- ☐ Redundant (color + shape) markers
- ☐ Lines and points are thick and clear
- ☐ All text is clearly legible
- ☐ Units and annotations are directly on plot
- ☐ Plot is reproducible from clean workspace
- ☐ High-resolution output (DPI > 300)
- ☐ Statistical markers added if needed
- ☐ Peer/mentor proofread

Building a figure

"Biomechanics of predator–prey arms race in lion, zebra, cheetah and impala"

*Data was pulled directly from **Wilson et al. Nature (2018)** and describes peak power/ kg of muscle for pairs of predatory and prey.*

What does our data look like?
4 groups, n > 30 per group with a broad distribution per group

Intended message
Predator peak muscle power is larger than prey muscle power.

Finding the plot that *best* fits your data can be challenging.

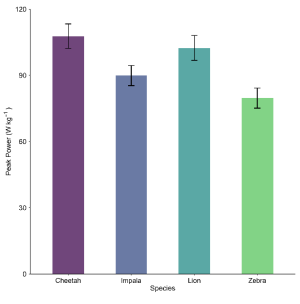
Focus on the message and remain true to the data.

Follow the checklist of the left to address all plot elements.

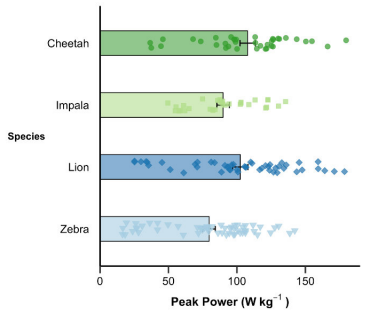
Bar Graph*

Let's start with a bar graph to illustrate its shortcomings.

*Don't use a bar graph for this data.



- ✗ Text sizes are too small
- ✗ Y-axis range is (0, 120), but the data is (0, 180)
- ✗ Underlying distribution is hidden
- ✗ Colors are meaningless
- ✓ Labels are clear
- ✓ Units are included

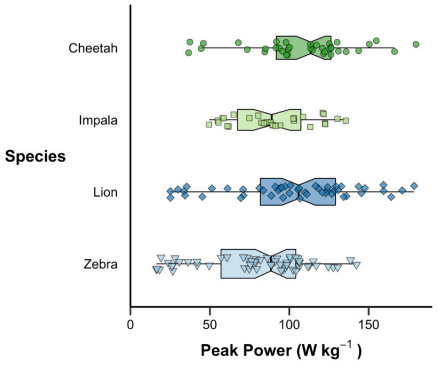


Boxplot

Boxplots can be a great way to display statistics related to the distribution plotted. However, they can still be misleading if points aren't shown.

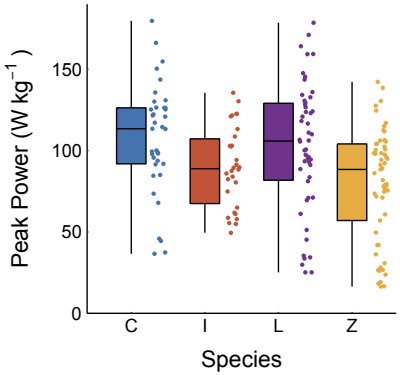
The notches shown here represent the 95% confidence interval of the medians, providing a metric meaningful to our message and ability for viewer to interpret differences between species

Building from the bar graph above, let's convert it into a boxplot



- ✓ Points are shown over the boxplot with transparency
- ✓ Entire range is shown
- ✓ Colors relate predator and prey
- ✓ Shapes encode different species
- ✓ Labels are large and clear

Authors chose a boxplot paired with a strip plot

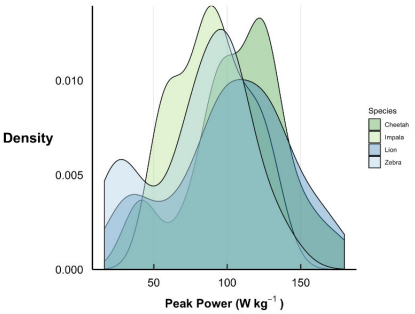


Density plot

Density plots are great for data with large *n*.

These density plots help visualize the entire distribution, but the parameters used to make the plot can be manipulating to over-smooth the underlying data.

For data with subtle differences, it's also harder for the reader to quickly see this.

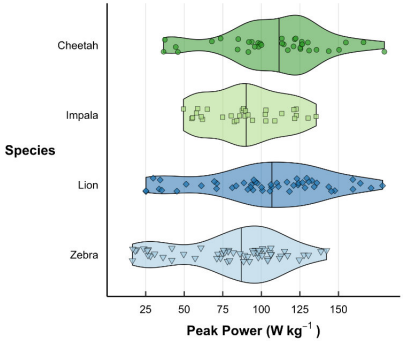


- ✗ Hard to discern statistical differences
- ✗ Hard to label with minor shifts in density
- ✗ Not enough points to use this plot effectively

Violin plot

Violin plots plot the densities at the right in a format similar to boxplots, which can combine the best of both worlds.

Statistical metrics can also be visualized on violin plots. Here we show the mean indicated by the black line within the violin distribution.



- ✓ Gridlines guide the reader to power thresholds
- ✓ Distribution across entire range is shown

All code and figures are stored at **https://github.com/MIT-BECL/awesome-becl-resources**

Learn more at **http://mitcommlab.mit.edu/be** – Updated August 2018