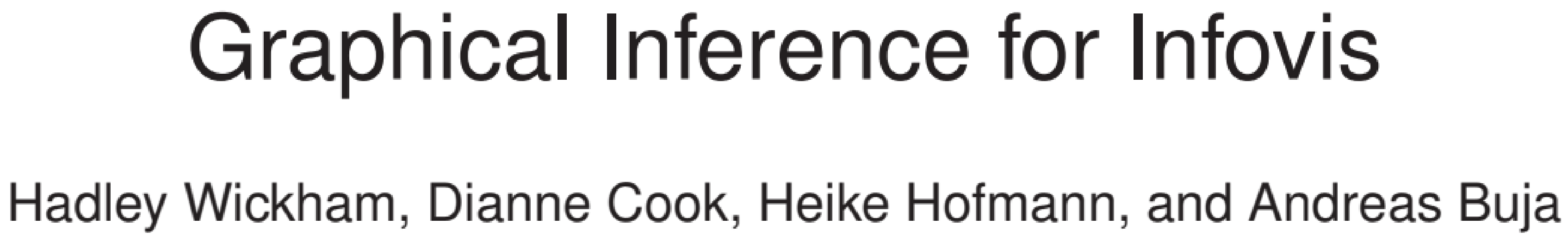
Week 7 A03

Reading Discussion



Graphical Inference for Infovis

Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja

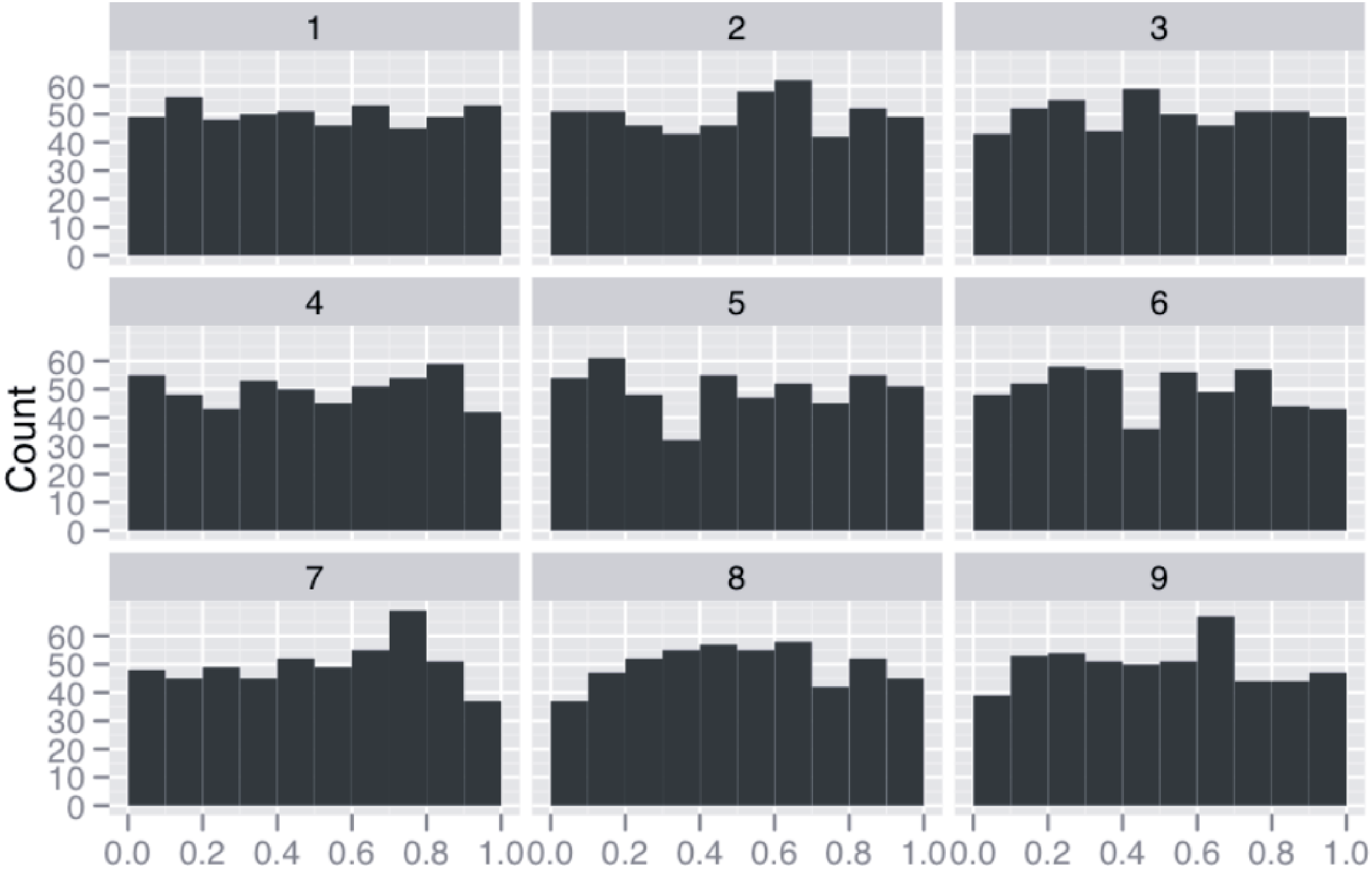
What is inference and why do we need it?

1) Statistical inference is the process of drawing conclusions about a population based on a sample of data.
2) Estimation involves using sample data to estimate the value of a population parameter, such as the mean or proportion.
3) Hypothesis testing involves making decisions about whether a particular hypothesis is supported by the data, based on a set of statistical criteria and a chosen level of significance.

Protocols of Graphical Inference

1) Rorschach: a calibrator, helping the analyst become accustomed to the vagaries of random.
2) Line-up (works like a police line-up): the suspect (test statistic plot) is hidden in a set of decoys. If the observer, who has not seen the suspect, can pick it out as being noticeably different, there is evidence that it is not innocent.

Misleading due to patterns in random noise



Example of Rorschach Protocol

To use the line-up protocol:

1) Identify the question the plot is trying to answer.

2) Characterize the null-hypothesis.

3) Figure out how to generate null datasets.

Selected visualizations in terms of their purpose and associated null distributions

1) Tag clouds: a visual representation of text data. It typically consists of a collection of tags, or keywords, that are displayed in different sizes or colors based on their frequency or importance.

2) Scatterplot: displays the relationship between two continuous variables, and answers the question: are x and y related in some way? The scatterplot can reveal many different types of relationships, e.g., linear trends, non-linear relationships and clustering.

Example of tag clouds

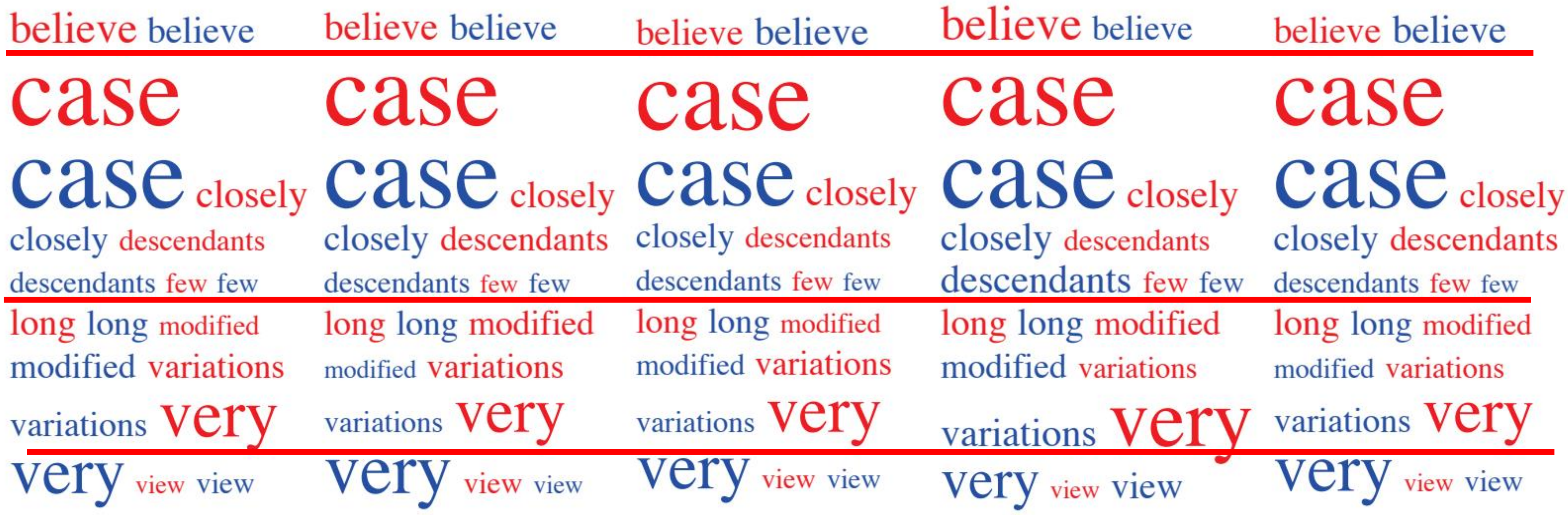


Fig. 5. Five tag clouds of selected words from the 1st (red) and 6th (blue) editions of Darwin's "Origin of Species". Four of the tag clouds were generated under the null hypothesis of no difference between editions, and one is the true data. Can you spot it?

Example of scatterplot

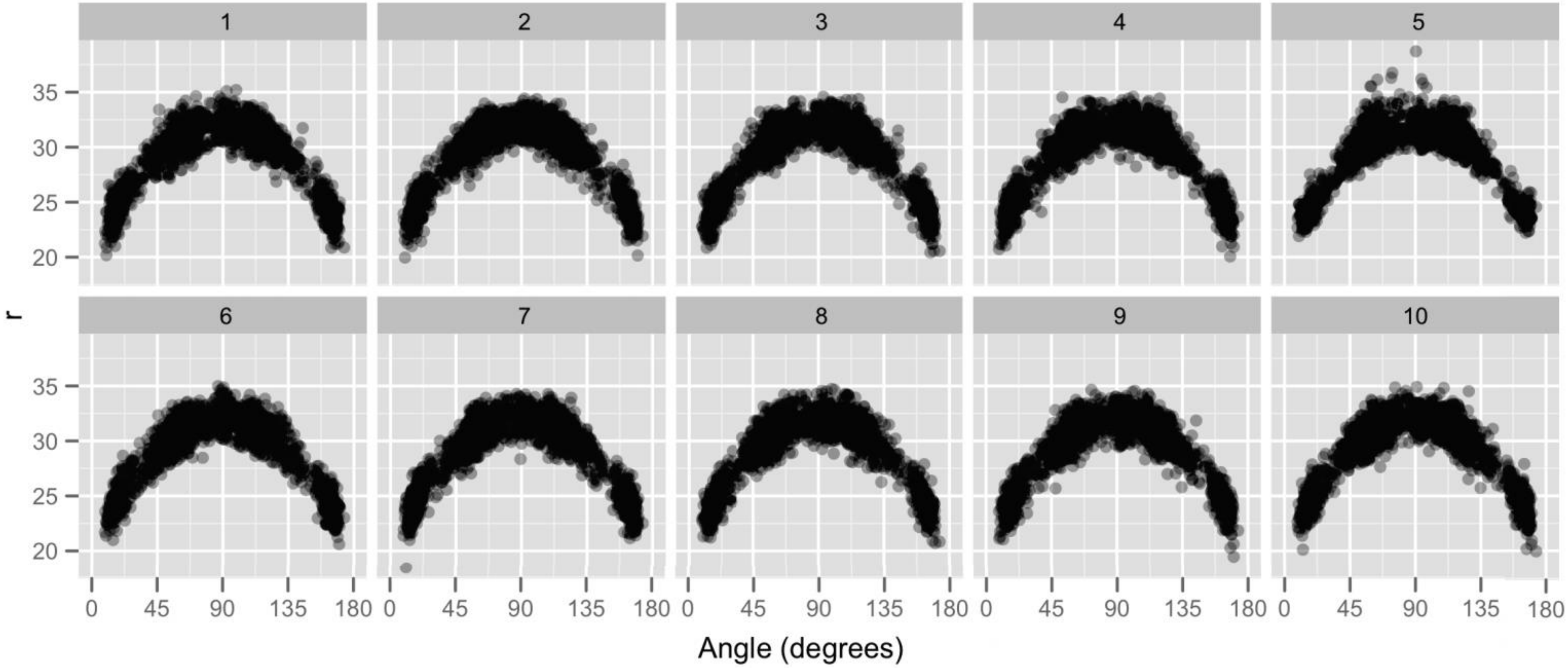


Fig. 6. Scatterplot of distance vs. angle for three pointers by the LA Lakers. True data is concealed in line-up of nine plots generated under the null hypothesis that there is a quadratic relationship between angle and distance.

The Power of Graphical Tests

- The probability of correctly convicting a guilty dataset. The capacity to detect specific structure in plots can depend on many things, including an appropriate choice of plot.
- The ability of graphical methods to detect patterns, trends, and differences in data that may not be apparent through traditional statistical tests.

Conclusion

- Rorschach and line-up protocols bring rigorous statistical inference to freeform data exploration.
- Both techniques center around identifying a null hypothesis, which then generates null datasets and null plots.
- The Rorschach provides a tool for calibrating our expectations of null data, while the line-up brings the techniques of formal statistical hypothesis testing to visualization.