Comparing Groups

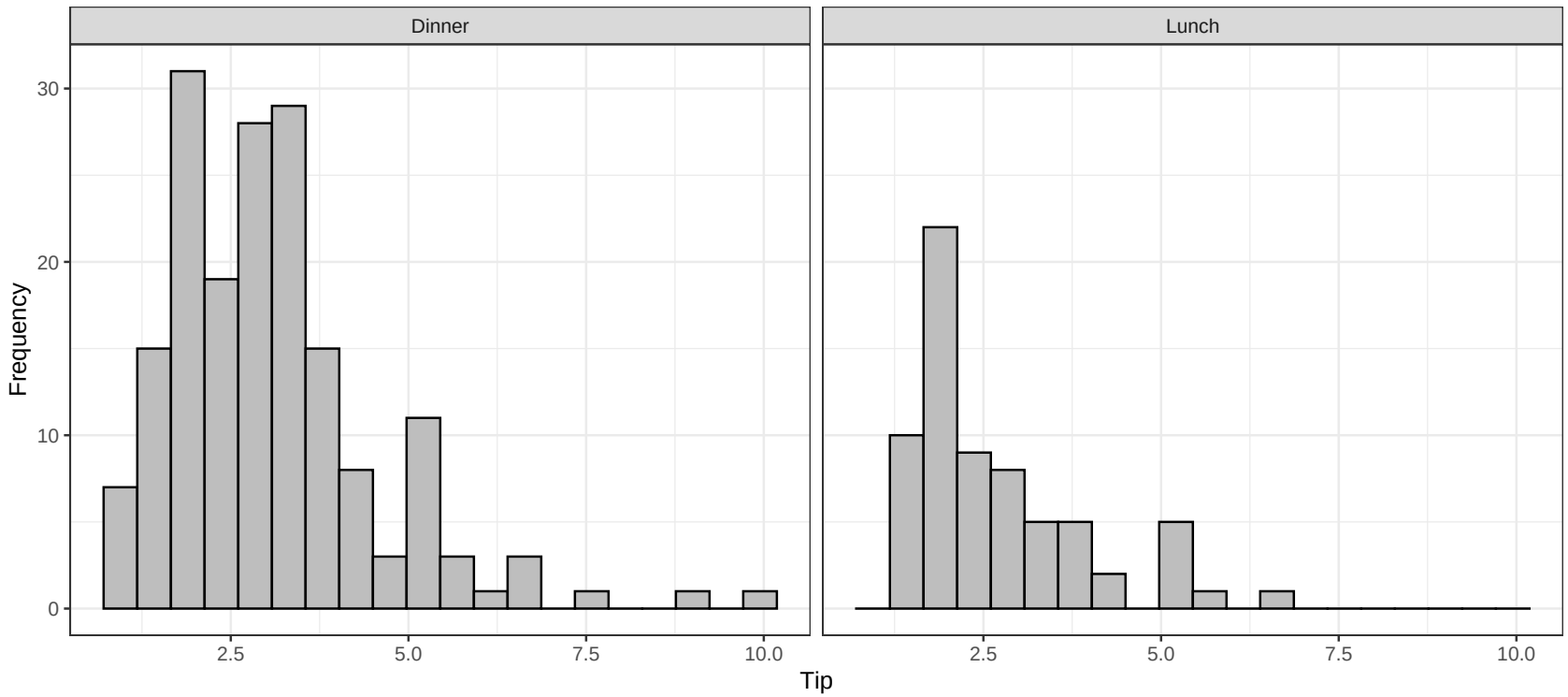
Ryan Miller

Introduction

- So far we've used contingency tables to find relationships between categorical variables
- We'll now look at relationships involving one categorical variable and one quantitative variable
  - This presentation will cover visual and numeric approaches to understanding these relationships

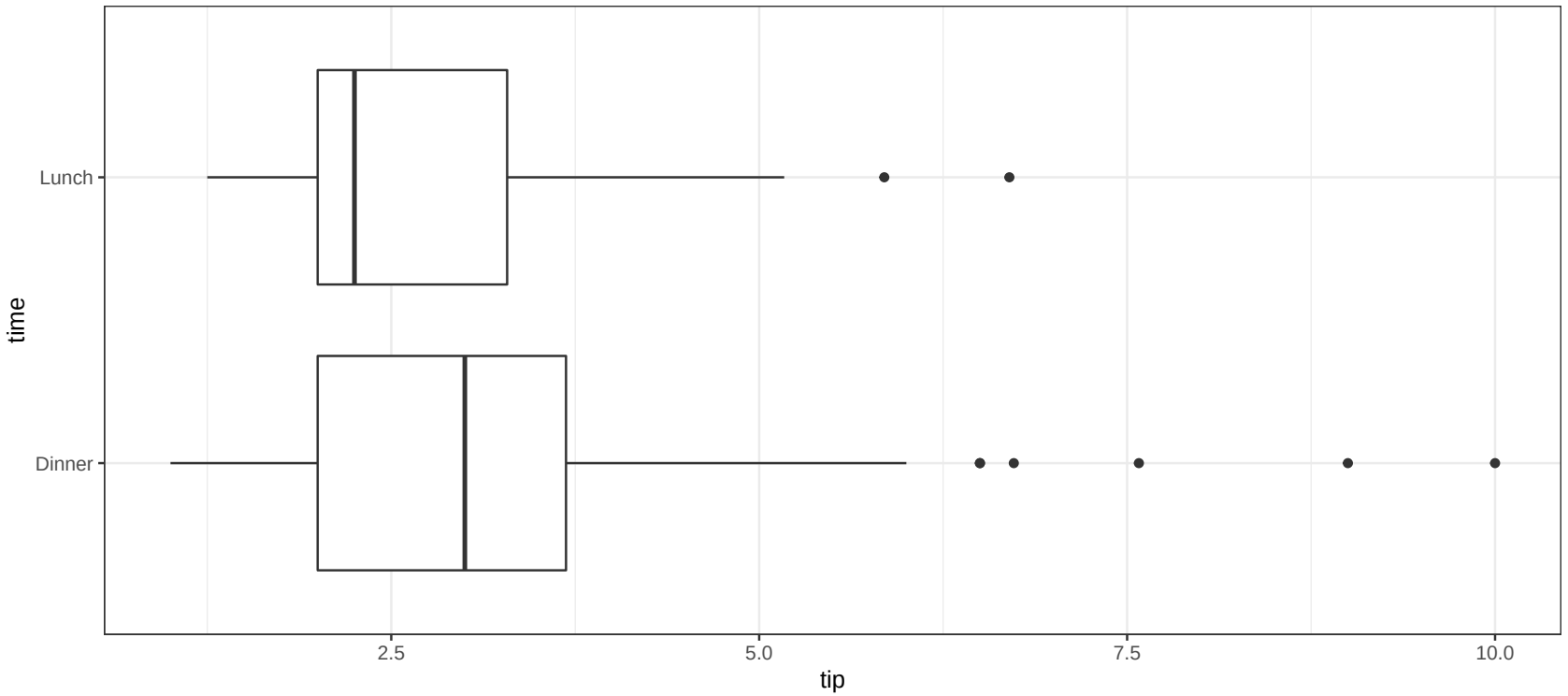
Side-by-side Graphs

- A simple way of comparing two or more groups (as defined by a categorical variable) is split up the cases by group and graph them side-by-side



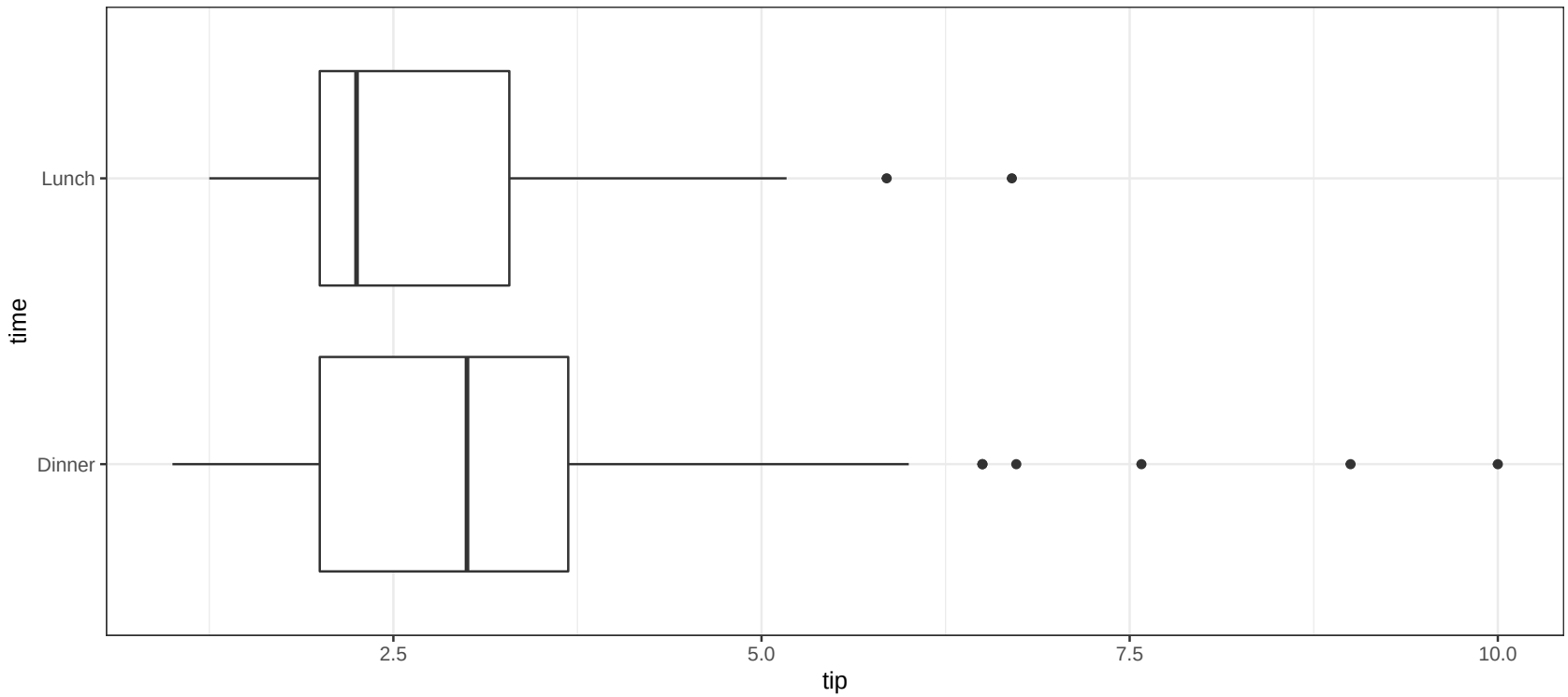
Side-by-side Graphs

- Boxplots tend to be more suitable for this approach since they allow for more direct comparisons (ie: median vs. median)



Association

- Recall we describe two variables as **associated** if the distribution of one variable depends upon the other
  - Thus, substantial differences in any single summary measure (medians, Q1, etc.) is enough to suggest an association, even if other parts of the distributions are similar



Numeric Summaries

- Boxplots are really just a visual representation of several different numeric summaries (minimum and/or potential outliers, Q1, median, Q3, maximum and/or potential outliers)
  - This suggests we can also find and describe associations by comparing side-by-side numeric summaries

time	min	Q1	median	mean	Q3	max
Dinner	1.00	2	3.00	3.102670	3.6875	10.0
Lunch	1.25	2	2.25	2.728088	3.2875	6.7

Reporting Associations

- Being able to identify an association is important, but we also need to be able to describe it to others with sufficient precision
  - As an example, we might report an association between tip and time in the Tips dataset by saying:

"The mean tip at Dinner is 38 cents (0.38 dollars) higher than the mean tip at Lunch"

- In this class, the **difference in means** will be our go-to when reporting on an association between two groups
  - That said, nothing prevents us from reporting a *difference in medians* or a *difference in 90th percentiles*

Practice

1. Open the Tips dataset in the "data explorer" app
2. Using boxplots, does there appear to be an association between smoking status and tip amount?
3. Report the difference in *means* for tips given by smokers and non-smokers
4. Report the difference in *medians* for tips given by smokers and non-smokers, when might it be wise to report this difference instead of a difference in medians?