Comparing Groups

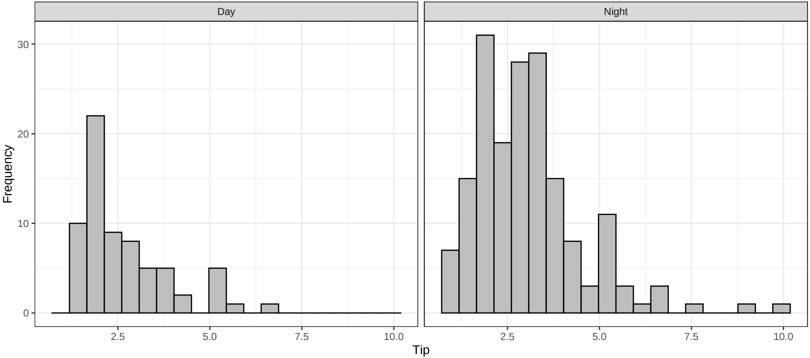
Ryan Miller

Introduction

- Previously, we introduced *contingency tables* as a method for summarizing relationships between *two categorical variables*
- Today we'll introduce methods for summarizing relationships between *one categorical and one quantitative variable*

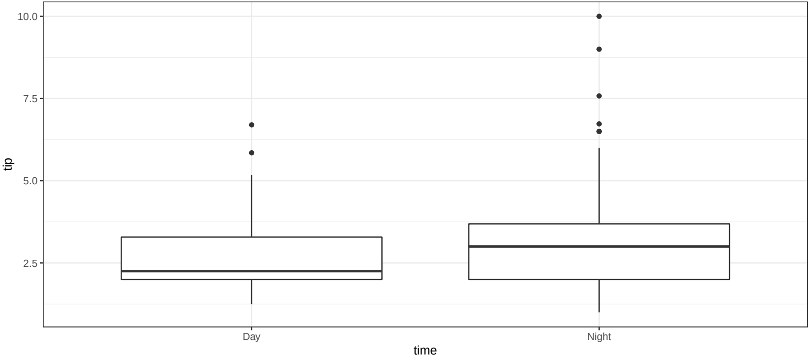
Side-by-side Graphs

- 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



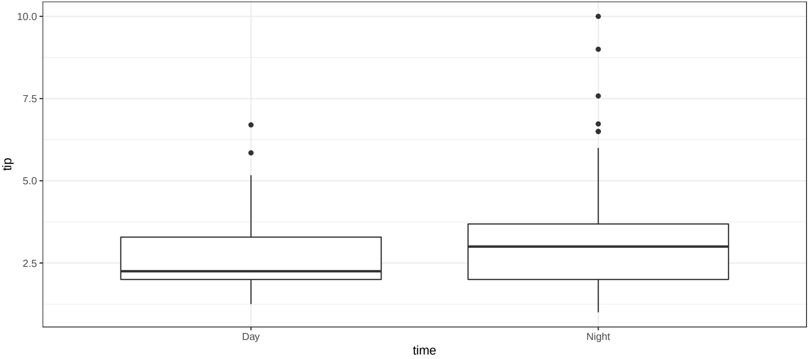
Side-by-side Graphs

- Boxplots tend work better for this since they easily facilitate direct comparisons (ie: median vs. median)



Association

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



Numeric Summaries

- Boxplots are just a visual representation of several different numeric summaries (minimum, Q1, median, Q3, and maximum)
  - So we can also find and describe associations using side-by-side numeric summaries

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

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 an association between two groups
  - That said, nothing prevents us from reporting a *difference in medians* or a *difference in 90th percentiles*

Practice

Using the “Tips” dataset, available by clicking here or on our website, go to https://www.lock5stat.com/StatKey/index.html, and click on the “One Quantitative and One Categorical” menu in the “Descriptive Statistics and Graphs” section

1) Upload the relevant columns from the “Tips” data to create boxplots that show the relationship between smoking status and tip amount
2) Report the *difference in means* for tips given by smokers and non-smokers
3) Report the *difference in medians* for tips given by smokers and non-smokers
4) Which difference do you think is better to report?

Closing Remarks (common misconceptions)

- At this point in the course, a substantial difference in any portion of the distribution across groups is sufficient to claim an association
- It is not necessary that all groups have differences in distribution. Instead, differences across any two categories is sufficient to claim an association