ISTA 116: Statistical Foundations for the Information Age

Grouped Numeric Data

21 and 26 September 2011

Outline

Reminders/Announcements

- Web Quiz 3 due Friday.
- Web Quiz 4 (up soon) due next Wednesday
- Lab 3 (up soon) due a week from Friday

Multivariate Data: Three Cases

- The kinds of relationships we can identify depend on the types of variables we have
- Three Cases:
  - All categorical variables ✓
  - **A mix of categorical and numeric**
  - All numeric

One (Binary) Categorical, One Numeric Variable

- What can we do with data like this?

Sex	Height (in.)
M	74
F	64
F	61
M	68
F	70
F	69
M	72
M	68

- As with categorical data, we can *condition* on one variable (here, the categorical one)

Back-to-Back Stem Plots

- One way of displaying the conditional distributions is via back-to-back stem-and-leaf plots
- **Note: the smaller "leaf" values always go closest to the stem**

Males	Stem	Females
	6	1 4
8 8	6	9
4 2	7	0

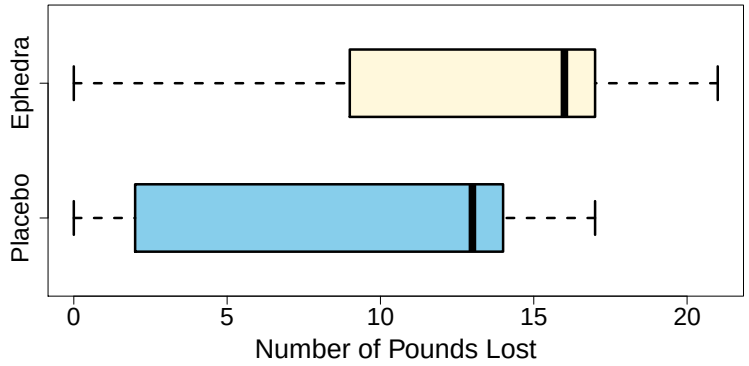
Weight Loss in Ephedra Trial

- Number of pounds lost during a clinical trial for placebo and ephedra diet pill users

Placebo	Stem	Ephedra
4 2 0 0 0	0	0
5	0	6 7 9
4 4 4 3	1	1 3
7 7 5	1	6 6 6 7 8
	2	0 1

Weight Loss in Ephedra Trial

- Since we're measuring the same numeric variable for both groups, we can plot multiple boxplots on the same scale
- Easy to compare quartiles/hinges



Time to Taxi for Flights out of EWR

- This extends easily to more than two groups

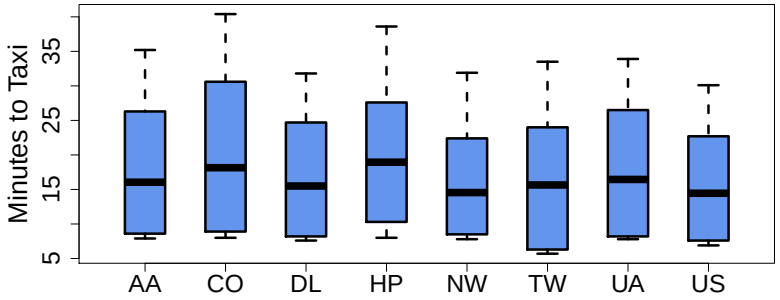
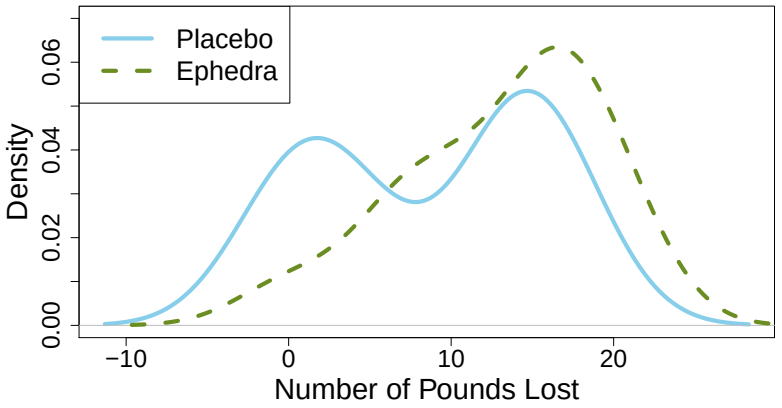


Figure: Minutes to Taxi for Flights from Newark Airport for Various Carriers

Overlaid Density Curves

- With two groups, we can see even more detail by overlaying two density curves:



Review: Quantiles

- With side-by-side box plots, we could visually compare the quartiles (well, hinges) for two numeric distributions
- We can extend this to comparing arbitrary quantiles
- Recall: the p^{th} **quantile** is the same as the $100p^{\text{th}}$ percentile.
  - The 50^{th} percentile is the 0.5 quantile
  - The 33^{rd} percentile is the 0.33 quantile
  - etc.

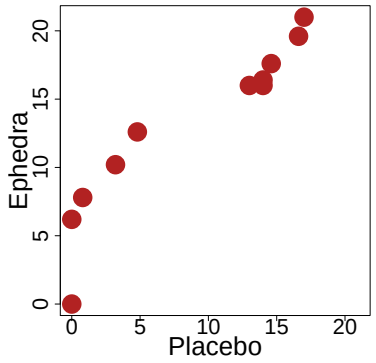
Quantile Pairs

- If we select a set of quantiles (e.g., $\{0.0, 0.1, 0.2, 0.3, ..., 1.0\}$), then for each quantile, we have a value for *each* distribution.

Quantile	Placebo	Ephedra
0.0	0.0	0.0
0.1	0.0	6.2
0.2	0.8	7.8
...	...	...
0.9	16.6	19.6
1.0	17.0	21.0

- Since we have two distributions, we can treat each quantile as a coordinate in 2D, and plot it. This is called a **Quantile-Quantile Plot** (or **QQ Plot**).

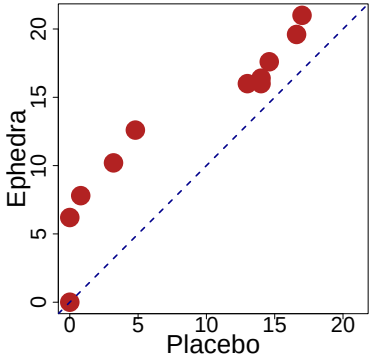
QQ Plot of Ephedra Data



- Notice that the x and y ranges are the same.
- What can we tell from this graph?
- Where would the points be if the two distributions were the same?
- What if one were just shifted up or down?
- What if one had greater variability?

ISTA 116: Statistical Foundations for the Information Age
Comparing Numeric Data Across Groups
Quantile-Quantile (QQ) Plots

QQ Plot of Ephedra Data



- It's useful to plot a reference line based on identical distributions.
- Now we can see that people at each quantile besides the 0.0 quantile lost more weight with the drug than with the placebo.