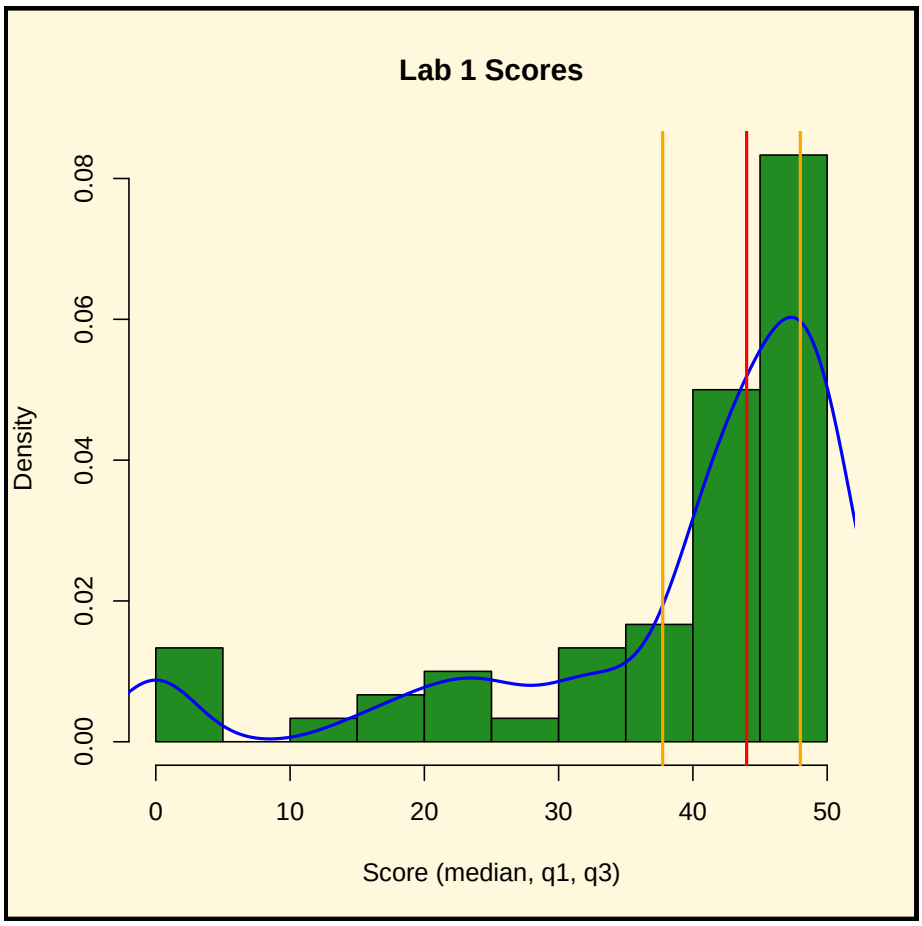
ISTA 116 Lab: Week 5

Colin Dawson

Last Revised September 19, 2011

1 HW1 Scores



2 HW2

- **Go over HW2**

3 Bivariate Categorical Data

- What does bivariate mean?
- What's the difference between one bivariate data set and two univariate data sets?
- With a single categorical variable, we summarized it using counts for each category, creating a *table*. We also used *prop.table* to convert this to proportions.
- When *both* variables in a bivariate data set are categorical, we could just create two tables. But what do we lose by doing this?

3.1 Contingency Tables

- Tables that give counts for *combinations* of variables are called *contingency tables* (e.g. how often do children have brown eyes, *contingent* on their mother having brown eyes?).
- Create them with `table()`, just include more than one variable in the arguments.
  - Note: The variables should "line up" (the first element goes with the first element, etc.), or the table won't make sense. Why?

3.2 Interactive Experiment!

- Do you prefer coffee, tea or neither (c,t,n)?
- More often, do you prefer salsa that's red, green, or nonexistent (r,g,n)?

Use `rbind` to "bind" data together as rows:

```
> preferences <- rbind(c("c", "r"), c("c", "r"), c("c", "g"), c("t",
+     "r"), c("t", "n"), c("c", "g"), c("t", "g"), c("c", "r"),
+     c("t", "r"), c("t", "r"), c("n", "g"), c("n", "n"), c("c",
+         "r"), c("c", "r"), c("t", "r"), c("t", "g"), c("c", "r"),
+     c("c", "r"), c("t", "r"), c("c", "r"), c("n", "r"))
```

Convert the result to a data frame:

```
> preferences <- as.data.frame(preferences)
```

Give it names and make a table:

```
> names(preferences) <- c("beverage", "salsa")
> (preferencesFreqTable <- with(preferences, table(beverage, salsa)))
```

Now that we have a table, we can ask questions easily:

- Are coffee drinkers more likely to go for red salsa than green?
  - How would we answer this?
- Is the overall preference for red salsa stronger among coffee drinkers?
  - What about this? Can we (easily) answer it just by looking at counts?
- Do salsa preferences differ across beverages?
- Do beverage preferences differ across salsas?

3.3 Joint and Marginal Frequencies and Proportions

- Joint Frequency: "How many people prefer coffee *and* red salsa?"
  - Just look at the frequency table
- Joint Proportion: "What proportion of people prefer coffee *and* red salsa?"
  - Divide a cell count by the sum of *all* counts
  - One value for each cell in the table
  - These sum to 1 across _____.
  - In `R`, use `prop.table(someFrequencyTable)`. Note: the argument is already a *table*.

```
> (preferencesJointPropTable <- prop.table((preferencesFreqTable)))
```

- Marginal Frequencies: "How many people prefer each salsa (regardless of beverage choice)?"
  - Sum down the columns (in this case) of the frequency table

  - Use `margin.table(someFrequencyTable, margin = 2)`
  - Set `margin = 1` for questions about rows (e.g. about beverage preference totals)
  - **Rows always come before columns!**

```
> margin.table(preferencesFreqTable, margin = 2)
```

- Marginal Proportions: "What proportion of people prefer each salsa (regardless of beverage choice)?"
  - How would you do this?
  - How would you calculate this without R?

3.4 Conditional Proportions

- Do salsa preferences differ across beverages?
- Do beverage preferences differ across salsas?

To answer these, we need *conditional proportions*.

- "*Given* people who prefer coffee, what proportion prefer red salsa?"
- In other words, "*Conditioned on* having a preference for coffee, what proportion prefer red salsa?"
- Conditioning means we are restricting our attention to a particular subset of the data (or section of the table).
  - This affects which total we care about.
- In `R`, we can do `prop.table(someFrequencyTable, margin = 2)` to get proportions *conditioned* on the column variable (set `margin = 1` to condition on rows).
- If we condition on columns, the proportions should sum to 1 across each _____.

```
> (preferencesCondPropTable <- prop.table(preferencesFreqTable,
+     margin = 2))
```

3.5 Pre-summarized data

Sometimes our data comes in already summarized into counts, rather than individual observations, but not necessarily in the right format.

- Bring in the `Simonoff07.csv` data set, on causes of power plant failures in the U.S. and Canada.
- First column is the levels of one variable; remaining columns are counts at levels of the other variable.
- Not straightforward to create a contingency table from this format.
- Instead, tell `R` that the first column is special, by specifying `row.names = "Nation"` when you import the data. Now it will look like a table.
- We need to turn it into a proper table instead of a data frame, though, using `as.matrix()`

```
> pplants <- read.csv("Simonoff07.csv", header = TRUE, row.names = "Nation")
> pplantsTable <- as.matrix(pplants)
> (pplantsJointProbTable = prop.table(pplantsTable))

> sum(pplantsJointProbTable[1, ])

[1] 0.8016194

> sum(pplantsJointProbTable)

[1] 1

> (pplantsCondProbTable <- prop.table(pplantsTable, margin = 1))

> sum(pplantsCondProbTable)

[1] 2

> sum(pplantsCondProbTable[1, ])

[1] 1

> (pplantsCondProbTable <- prop.table(pplantsTable, margin = 2))

> sum(pplantsCondProbTable[1, ])

[1] 7.517581

> sum(pplantsCondProbTable[, 1])

[1] 1
```

Exercise: Find out whether a failure is more or less likely to be due to equipment failure in the U.S. vs. Canada.

3.6 Visualizing Bivariate Categorical Data

If there are more than 3 or 4 rows/columns, it's hard work to read through a contingency table and see easily what's going on. We'd like some sort of graphical depiction of the data.

3.6.1 Grouped and Stacked Bar Plots

Just as we used `barplot()` to create a bar plot of a univariate table of counts, we can use it on a two-way table.

- By default, `barplot(myTable)` will draw one bar for each *column*, whose height is the marginal frequency for that column.
- Each bar is subdivided into stacked pieces whose size corresponds to the cell counts in that column.

```
> par(bg = "cornsilk1")
> beverages <- c("Coffee", "Neither", "Tea")
> salsas <- c("Green", "Neither", "Red")
> beveragecolors <- c("saddlebrown", "skyblue", "sienna2")
> barplot(preferencesFreqTable, names.arg = salsas, xlab = "Salsa Preference",
+     ylab = "Number of Students", col = beveragecolors, legend.text = beverages)
```

If we want to draw the plot the other way around, the easiest thing to do is to *transpose* the table, using the `t()` function.

```
> salsacolors <- c("yellowgreen", "ivory", "firebrick")
> prefTable2 <- t(preferencesFreqTable)
> barplot(prefTable2, names.arg = beverages, xlab = "Beverage Preference",
+     ylab = "Number of Students", col = salsacolors, legend.text = salsas)
```

Another option, instead of stacking the bars, is to group the bars. To do this, just specify `beside = TRUE`.

```
> barplot(prefTable2, names.arg = beverages, xlab = "Beverage Preference",
+     ylab = "Number of Students", col = salsacolors, legend.text = salsas,
+     beside = TRUE)
```

By default, the table is created in alphabetical order along each axis. We might want to reorder it for plotting purposes:

```
> prefTable3 <- prefTable2[c("r", "g", "n"), c("c", "t", "n")]
> beverages <- c("Coffee", "Tea", "Neither")
> salsas <- c("Red", "Green", "Neither")
> salsacolors <- c("firebrick", "yellowgreen", "ivory")
> barplot(prefTable3, names.arg = beverages, xlab = "Beverage Preference",
+     ylab = "Number of Students", col = salsacolors, legend.text = salsas,
+     beside = TRUE)
```

Exercise: Show frequencies of power outages in the U.S. and Canada, grouped by cause.

We may not care about absolute numbers, but instead we want to focus on *conditional* proportions. Just plot the result of `prop.table()`.

Condition and Group on Beverage:

```
> prefProps <- prop.table(prefTable3, margin = 2)
> barplot(prefProps, names.arg = beverages, xlab = "Beverage Preference",
+     ylab = "Number of Students", col = salsacolors, legend.text = salsas,
+     beside = TRUE)
```

Exercise: Show distributions of power outage cause, conditioned on country but grouped by cause.

4 Questions?