ISTA 116 Lab: Week 3

Colin Dawson

Last Revised September 5, 2011

1 HW1

- **Go over HW1**
- **Reminder: Please save your HW as a pdf!**
- **Please add your e-mail address and the time of your lab section to the top of future homeworks.**

2 Visualizing Numeric Data

Because the data has an underlying scale, we can think of each point as a location on a number line.

2.1 Strip Charts

- The `DOTchart()` function in the `UsingR` package creates a graphic representation of data points lying on a scale.

```
> data(Pima.te)
> DOTplot(Pima.te$age, main = "Ages", xlab = "Age")
```

2.2 Histograms

- A closely related graphic is the *histogram*
- Similar to a strip chart, but uses bars (that represent ranges, or "bins"), rather than stacks of dots
- Unlike a bar chart (for categorical data), the bars on a histogram are touching to show that there's an underlying numeric scale.
- Using a histogram for categorical data would be misleading. Why?
- Similarly, using a bar chart for continuous numeric data would be confusing.

```
> hist(Pima.te$age,
      main = "Diabetes in Pima Indian Women",
      xlab = "Age",
      ylab = "Number")
```

In addition to the usual plot options, some specific to `hist()` include:

Common histogram options	
`breaks=`	Suggest how many "bins" to use
`prob=`	Set to `TRUE` to use proportions rather than counts
`col=`	Color the bars

```
> par(bg = "cornsilk1") # modify the background color
> hist(Pima.te$age,
     breaks = 20,
     main = "Diabetes in Pima Indian Women",
     xlab = "Age",
     ylab = "Number",
     col = "forestgreen",
     prob = TRUE)
```

2.3 Density Curves

For continuous variables, as we get more and more data, we expect the distribution to approach a "smooth" curve.

- The `density()` function creates a "guess" of what this curve might look like if we had more data.
- How "densely" would the data be packed around a particular value?

```
> ageDensity = density(Pima.te$age)
> plot(ageDensity,
     main = "Density Plot of Age",
     xlab = "Age",
     ylab = "Density")
```

- It's especially useful to see a histogram and a density curve together
- Plot the histogram first, then overlay the density curve with `lines()`
- `lines()` is an example of a "low-level" plotting command that adds to an existing plot instead of creating a new one.
- Others include `abline()`, `points()`, `arrows()`, `rect()`, ...

```
> hist(Pima.te$age,
     breaks = 20,
     main = "Diabetes in Pima Indian Women",
     xlab = "Age",
     ylab = "Number",
     col = "forestgreen",
     prob = TRUE)
> lines(ageDensity, col = "darkblue", lwd = 2)
```

- Notice how the curve is below the high bars but above the low ones.
- Unless we have huge amounts of data, really frequent observations (relative to their neighbors) are likely to be overestimates of the "long run" proportions, and vice-versa.
- This is an example of a phenomenon called "regression to the mean", which we may talk about when we get to probability.

3 Central Tendency

Several measures of the "center" of a distribution

- Mean: “balance point”
- Median: half the data above, half below
- Mode: most common value (or point w/ greatest “density”)
- Midrange: halfway between min and max values

```
> dat = c(0,3,3,5,5,5,7,7,10)
> hist(dat,prob=TRUE)
> lines(density(dat), col = "green",lwd = 2)
> mean(dat)
[1] 5
> abline(v = mean(dat), col = "red", lwd = 2)
> median(dat)
[1] 5
> abline(v = median(dat), col = "blue", lwd = 2)

> dat = c(0,3,3,5,5,5,7,7,100)
> hist(dat,prob=TRUE, ylim = c(0.0,0.15))
> lines(density(dat), col = "green",lwd = 2)
> abline(v = mean(dat), col = "red", lwd = 2)
> abline(v = median(dat), col = "blue", lwd = 2)
```

3.1 The Mean

Advantages

- Easy to understand
- Uses all the data
- Mathematically convenient

Disadvantages

- Sensitive to outliers
- Can misrepresent asymmetric distributions

  ```
  > library(UsingR)
  > data(babies)
  ```

```
> wt <- babies$wt
> hist(wt,
      main = "Birthweights of Newborns",
      prob = TRUE,
      breaks = 40,
      xlab = "Weight (oz)",
      col = "forestgreen")
> lines(density(wt), col = "darkblue", lwd = 2)
> meanWt = mean(wt)
> abline(v = meanWt, col = "red", lwd = 2)
```

This distribution is symmetric, so the mean looks like a pretty good representation of the center.

What about this one (Example 2.5)?

```
> data(cfb)
> income <- cfb$INCOME / 1000
> hist(income,
      main = "U.S. Income in 1000's of Dollars",
      prob = TRUE,
      breaks = 40,
      xlab = "Income (K$)",
      ylim = c(0,0.015),
      col = "forestgreen")
> lines(density(income), col = "darkblue", lwd = 2)
> abline( v = mean(income), col = "red", lwd = 2)
> (meanIncome <- mean(income))  # parens assign and display in one line
```

- Distributions like Income are severely *skewed*: the mean is pulled up above the median by extreme values.

3.2 The Median

Advantages

- Same location regardless of units
- Resistant to skew

Disadvantages

- May not want to ignore extreme values
- Distributions with very different means can have same median

3.3 Mode(s)

What do you think about this distribution?

```
> erup <- faithful$eruptions
> hist(erup,  #How long do eruptions last for Old Faithful?
         main = "", #No title
         prob = TRUE,
         breaks = 20,
         xlab = "Eruption Duration",
         col = "forestgreen"
         )
> lines(density(erup),
          col = "darkblue",
          lwd = 2
          )
> abline( v = c(mean(erup), median(erup)),
            col = c("red","magenta"),
            lwd = 2
           )
```

- Where do you think the mean is?
- The median?

This distribution is *bimodal*: it has two peaks.

- In cases like this, sometimes reporting the *modes* is more informative than the mean or the median.

Another measure is the *midrange*, halfway between the highest and lowest values. This is easy to compute (and visualize) for a “quick and dirty” sense of center, but there’s little point with computers.

4 Transforming Variables

One useful "trick", when a distribution is "badly behaved", is to *transform* the values to a new scale.

- A common transformation for right-skewed "ratio" data is the logarithm.
- Makes intuitive sense when ratios, rather than differences, "feel like" the right unit of comparison.

```
> par(mfcol=c(2,1)) #Here we'll show two plots vertically
> ###The old income plot
> ##Repeated Part
> income <- cfb$INCOME / 1000
> hist(income,  #US Income in Thousands of $
         main = "", #No title
         prob = TRUE, #We will overlay a density
         breaks = 40,
         xlab = "Income (K$)",
         ylim = c(0,0.02), #Need to extend range for density curve
         col = "forestgreen"
         )
> lines(density(income),
          col = "darkblue",
          lwd = 2
          )
> meanincome <- mean(income)
> abline( v = meanincome,
            col = "red",
            lwd = 2)
> sdincome <- sd(income)
> abline( v = c(meanincome - sdincome, meanincome + sdincome),
            col = "purple"
            )
> #####
>
> #Transform with the base 10 log so we can understand the units
> #First drop 0s (another option is to add 1 everywhere)
> logincome <- log10(cfb$INCOME[cfb$INCOME != 0])
```

```
> hist(logincome,  #Income in Log of $
        main = "", #No title
        prob = TRUE, #We will overlay a density
        breaks = 40,
        xlab = "Log Income",
        col = "forestgreen"
        )
> lines(density(logincome),
         col = "darkblue",
         lwd = 2
         )
> meanlogincome <- mean(logincome)
> sdlogincome <- sd(logincome)
> abline( v = c(meanlogincome,
                meanlogincome - sdlogincome,
                meanlogincome + sdlogincome),
          col = c("red", "purple","purple"),
          lwd = c(2,1,1)
         )
```

5 Variability

- Variability is *the key to statistics*

Measures of variability:

- Variance ($\sim$ "average squared deviation" from the mean)
  - `var()` in `R`
- Standard Deviation (square root of variance; same unit as original variable)
  - `sd()` in `R`
- Inter-Quartile Range (range of "middle half" of the data)
  - `IQR()` in `R`

6 The Five Number Summary and Box and Whisker Plot

Can tell a lot about a distribution with five numbers:

- Minimum value (aka Q_0)
- Q_1
- Q_2 (the median)
- Q_3
- Maximum value (aka Q_4)

Collectively, known as the "five number summary"

- Available in Rwith `fivenum()`

```
> fivenum(wt)
> fivenum(income)
```

This info is easily visualized with a *box and whisker plot*.

- `boxplot()` in R
- The box goes from Q_1 to Q_3, with a line at the median.
- The whiskers extend (by default) 1.5 times the IQR from the box edges. "Well-behaved" distributions have almost all the data in here.
- (The multiple can be set with the `range=` argument.)
- "Outliers" plotted individually.

```
> ##Set up 3 plots in one window
> par(mfrow=c(1,3))
> boxplot(wt,xlab="Birthweights")
> boxplot(income,xlab="Incomes")
> boxplot(erup,xlab="Eruption Durations")
```

7 Last Minute HW2 Questions?

8 Quick Glance at HW3