ISTA 116 Lab: Week 4

Colin Dawson

Last Revised September 12, 2011

1 Warmup Exercises

1.1 Executive Pay

- The `exec.pay` data set in the `UsingR` package contains salaries (in $10K units) of U.S. executives in the year 2000.
- Make a histogram and overlay a density curve on this data.
- What would be a good measure of central tendency? Variability?

1.2 Diabetes Revisited

- The `Pima.te` data set in the `UsingR` package contains data about occurrences of diabetes in Pima women.
- Separate the data into 2 sets according to type (whether the subject has diabetes or not). Remember that `Pima.te$type` is equivalent to `Pima.te["type"]`
- Plot the histograms and density curves of `age` for each set. To plot 2 histograms on the same page, use the command `par(mfcol=c(2,1))` and be sure to use `xlim` and `ylim` to set the range of the axes to be the same in both histograms.
- What to the histograms tell us about age and diabetes?

2 Review of Variability

2.1 Interquartile Range

- The IQR is often paired with the median, because it is "robust" to outliers (on the flip side, it "ignores" half the data).

```
> library(UsingR)
> data(exec.pay)
> par(bg = "cornsilk1")
> hist(exec.pay,  #Exec Income in $10K
         main = "", #No title
         prob = TRUE, #We will overlay a density
         breaks = 40,
         xlab = "Income",
         ylim = c(0,0.02), #Need to extend range for density curve
         col = "forestgreen"
         )
> lines(density(exec.pay),
          col = "darkblue",
          lwd = 2
          )
> medianincome <- median(exec.pay)
> abline( v = medianincome,
            col = "magenta",
            lwd = 2)
> IQRincome <- IQR(exec.pay)
> abline( v = c(medianincome - 0.5*IQRincome,
                  medianincome + 0.5*IQRincome),
            col = "orange"
           )
> #Actually, the "semi-IQR" is more comparable to the standard deviation
>
> #What will we get?
> (pctiles <-
     ecdf(exec.pay)(c(medianincome-0.5*IQRincome,medianincome+0.5*IQRincome)))
```

- In a skewed distribution, Q_3 is farther from Q_2 than Q_1 is from Q_2.

```
> data(faithful)
> 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
          )
> medianerup <- median(erup)
> abline( v = medianerup,
            col = "red",
            lwd = 2
           )
> IQRerup <- IQR(erup)
> abline( v = c(quantile(erup,0.25),
                  quantile(erup,0.75)
                  ),
            col = "orange"
           )
```

- In a bimodal distribution, Q_1 and Q_3 are very far apart

2.2 Practice Exercises

- Plot 2 histograms with density curves of the baby weight data set (`babies$wt`): On the first, overlay the mean and standard deviation, and on the second show the median and IQR.
- How do the two measurements compare?
- Now do the same for the income data discussed above.
- When is it better to use median & IQR vs. mean & sd?

3 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()`

```
> data(babies)
> wt <- babies$wt
> fivenum(wt)
> fivenum(exec.pay)
```

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(exec.pay,xlab="Incomes")
> boxplot(erup,xlab="Eruption Durations")
```

3.1 Practice Exercise

- How does a box and whisker plot relate to a histogram?
- Set R to plot 2 plots vertically with `par(mfcol=c(2,1))`
- Plot the weights in a `boxplot`, but set `horizontal=TRUE` as an additional argument.
- Plot a histogram below the box plot and overlay the median, Q1, and Q3 values.

4 Transforming Variables (Review)

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
> hist(exec.pay,
         main = "", #No title
         prob = TRUE, #We will overlay a density
         breaks = 40,
         xlab = "Income",
         ylim = c(0,0.02), #Need to extend range for density curve
         col = "forestgreen"
         )
> lines(density(exec.pay),
          col = "darkblue",
          lwd = 2
          )
> meanincome <- mean(exec.pay)
> abline( v = meanincome,
            col = "red",
```

```
           lwd = 2)
> sdincome <- sd(exec.pay)
> 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(exec.pay[exec.pay!=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 HW2 Questions?