STA 4033
Mathematical Statistics with Computer Applications
Lecture 3*

Brett Presnell
Department of Statistics
University of Florida

16 May 2001

*Copyright ©2001 by Brett Presnell
2. Exploratory Data Analysis: Graphics and Descriptive Statistics

2.1. Case Study: The 2000 Election in Florida

There were many questions about the 2000 presidential election in Florida.

- Did Buchanan get too many votes in Palm Beach Co.?
- Did punch card ballots cause uncounted votes?
- Were most of the uncounted ballots for Gore or Bush?
- Or did voters intend not to vote in the presidential election?
- Etc.

Will use this as an illustrative example.
2.2. Scatterplots

With two numerical variables it is useful to look at a *scatterplot* of the observations. In other words, assume we observe $X$ and $Y$ on $n$ cases, resulting in data $(x_1,y_1),\ldots,(x_n,y_n)$. A scatterplot is just a plot of all $n$ pairs as points in the Cartesian plane.

2.2.1. Example: the Buchanan Vote

- Was the Buchanan vote in Palm Beach Co. “unusual”? How to tell?
- One approach (Alan Agresti): consider relationship between Buchanan vote in 2000 and Perot vote (same party) in 1996.
  - Are they closely related for most counties?
  - Is Palm Beach different?
- Similar approaches taken by others (more later).

```r
> aa <- read.table("aa.txt")
> attach(aa)
> plot(perot, buchanan, xlab="Perot 1996", ylab="Buchanan 2000")
> identify(perot, buchanan, labels=row.names(aa))
> ## Now left-click on interesting points. Right-click to end.
> detach(aa)
```
Figure 1: Total vote, by county, for Reform Party candidates Buchanan in 2000 and Perot in 1996.
2.3. Boxplots, Barcharts, and Histograms

The boxplot is especially useful when comparing two or more groups on some numerical variable. How is it drawn?

- Ends of box represent first and third quartiles.
- Line across middle of box represents median.
- Whiskers go out to minimum and maximum.
  - “Outliers” may be plotted as dots beyond whiskers.
2.3.1. Undervote in Florida Counties

- Did punchcard ballots cause votes not to be counted?

- Figure 2 shows a replica of a barchart presented by Nicholas Hengartner (statistician for Gore legal team) in Leon Co. circuit court.

```r
> nonvotingrates <- c(0.00343, 0.0151, 0.0223)
> names(nonvotingrates) <- c("Optical Ballot Average", "Punch Card Average", "Palm Beach")
> barplot(100*nonvotingrates, ylim=c(0,2.5), col='gray', main="Non-voting Rates (\%)")
```
Figure 2: Statewide undervote rates for optical and punch card ballots. A replica of a plot presented by Prof. Nicholas Hengartner in the Leon County Circuit Court on Dec. 2, 2000.
• Figure 3: boxplots of county undervote rates is more informative (but confusing to lawyers?).

```r
> ### Boxplot of undervote by ballot type.
> fl2000 <- read.table("fl2000.txt")
> attach(fl2000)
> plot(100*Under/TotalVotesCast ~ Method)
> plot(100*Under/TotalVotesCast ~ Method, names=c("Optical","Punchcard"))
```
Figure 3: County undervote rates in Florida in the 2000 presidential election.
• Boxplots are useful for comparing different groups on a numerical variable, but for viewing a single distribution the histogram is generally superior. Figure 4 presents a histogram of county undervote percentages.

```r
# # # Histogram of county undervote rates.
# # Attach fl2000 if it is not already.
hist(100*Under/TotalVotesCast, xlab="Percentage Undervote", main="Undervote Rate for Florida Counties")
```
Undervote Rate for Florida Counties

Figure 4: County undervote rates in Florida in the 2000 presidential election.
2.4. Descriptive Statistics

Descriptive statistics is a term indicating summary measures of a set of data that are calculated in order to describe the data in some way. This terminology is mostly to distinguish intent. Here statistics are presumably calculated for descriptive purposes, whereas in a more formal statistical inference the intent is usually to estimate some unknown parameter or to make a decision, generally while providing measures of the goodness of the estimate and/or the reliability of the decision. In reality the line dividing descriptive and inferential statistics is indistinct, with one gradually fading into the other with much overlap.

The most common descriptive statistics are relative frequencies or proportions (often presented as percentages), measures of location (such as the mean, the median, and the trimmed mean), measures of variability or spread (such as the standard deviation and the interquartile range (IQR)), and measures of association (such as the correlation coefficient, to be discussed later).

Examples showing the calculation of these statistics in R are given in the transcript file 2001-05-17.Rt.