Introduction to Biostatistics 4120
Breheny

Lab #4

In today's lab, we'll be starting off with some summary statistics, then using the last half of discussion to review for the quiz on Thursday. We were introduced to the "Tips" dataset in Lab #2, but today we'll look more in depth at some its variables by looking at summary statistics, plots and histograms. Along the way we'll learn tools for describing, graphing, and exploring the distribution of continuous variables as well as relationships between two continuous variables, and between continuous and categorical variables.

Recall, the data set that we will use comes from the efforts of a waiter who recorded information about 244 tips he received over a period of a few months working at a restaurant (`tips.txt`). He recorded several variables:

- `Tip`: gratuity, measured in dollars
- `TotBill`: cost of meal including tax
- `Sex` of the person paying for the meal
- `Smoker`: indicates whether smoking or not
- `Day`: (Thur, Fri, Sat, Sun)
- `Time`: (Day, Night)
- `Size`: Size of the party

The important continuous variable in this dataset are `TotBill` and `Tip`.

1 Summary Statistics

In R, summary statistics for continuous data can be obtained using the functions below:

```
tips <- read.delim("http://myweb.uiowa.edu/pbreheny/data/tips.txt")
attach(tips)

## Summary Statistics
mean(Tip)
mean(TotBill)
median(TotBill)
sd(Tip)
quantile(Tip, c(0, .25, .5, .75, 1))
```

These functions, along with the `max, min` functions from lab 2, provide us with many of the summary statistics we talked about in lecture. If you are looking for one not listed here, chances are has a function for it. To obtain group specific summaries, we can use brackets:

```
mean(TotBill[Time == "Night"])
sd(TotBill[Time == "Night"])
mean(TotBill[Time == "Day"])
sd(TotBill[Time == "Day"])
```

What do you notice about the differences between the total bills during the day and the total bills at night? Why might this be?

2 Histograms

What does this look like when we plot it? Let's start with histograms. Histograms are created with either the `hist()` function or the `histogram()` function. The latter requires the `lattice` package, and produces slightly better looking histograms.

```
## Histogram
## gives counts
hist(TotBill)
## gives percentage
hist(TotBill, freq = F)

require(lattice)
## gives percentage
histogram(TotBill)
```

We can see that most bills are around $15, but that some were as high as $50. Besides looking nice, another advantage of the `histogram` function is that it allows us to break down the plot by conditioning:

```
histogram(~ TotBill | Time)
```

We can see slightly more clearly now that a rather high percent of lunches tend to be between about $10 and $20, whereas dinners are more spread out through the $20-$30 range as well. Note that this observation agrees with the mean and SD that we got earlier.

3 Box Plots

Box plots are pretty straightforward:

```
boxplot(TotBill ~ Time, ylab = "Total Bill")
points(1, mean(TotBill[Time == "Day"]), pch = 5)
points(2, mean(TotBill[Time == "Night"]), pch = 5)
```

Note you can see the skew of the data by determining where the mean lies in relation to the median. Is this data skewed left or right?

4 Scatter Plots

When we are interested in the connection between two continuous variables, we'll have to use something new; this is where scatter plots are useful.

```
plot(TotBill, Tip) ## Or:
xyplot(Tip ~ TotBill)
```

The plot illustrates several trends:

- As we would expect, there is a positive association between total bill and tip
- There is plenty of variation, however (big tips on small bills, small tips on big bills)
- There are more points in the lower right of the plot than the upper left – does this mean that cheap tippers are more common than generous tippers?
- There seem to be some horizontal stripes in the plot. Why is that?

We can add the regression line by using the `type` parameter to declare that we want both points (”p”) and the regression line (”r”):

```
xyplot(Tip ~ TotBill, type = c("p","r"))
```

This is the line that minimizes the residual sum of squares for predicting tip from total bill. Notice that we get a different line if we change x and y.

```
xyplot(TotBill ~ Tip, type = c("p","r"))
```

Finally, recall that conditioning helps us see how the relationship between bill and tip differs for different subcategories of dining parties. For example, let's compare smokers and nonsmokers:

```
xyplot(Tip ~ TotBill | Smoker)
```

The correlation between tip and bill seems to be much stronger in the nonsmoking section than in the smoking section. More on correlation next time, for now let's begin the review for Thursday's quiz.

5 Quiz Review

A: Selection Bias: bias resulting from when your sample has not been properly randomized from your population.

B: Nonresponse bias: bias that results when respondents differ in meaningful ways from nonrespondents.

C: Perception bias: Merely knowing you're receiving treatment can have an effect (the placebo effect). In a properly controlled and blinded study, this is not an issue.

D: Diagnostic bias: Doctors may change their diagnosis if they know whether a patient has been given treatment/placebo. In double blind studies, this is not an issue.

In each of the following examples, determine which bias(es) may be present. If possible, determine which direction the bias may skew the results. Then, state the null and alternative hypotheses.

- A doctor wanted to investigate whether Tylenol is better than Ibuprofen in curing head-aches, so he designed an experiment in which he randomly selected which treatment he would give people and blinded them to which one they got. He then noted how much their condition improved in either case.

- A statistician who was also a Subway enthusiast was heartbroken to find out that his footlong sandwich was only 11 inches long. He sets out to determine what the true mean sandwich length is by measuring his Subway lunch every day for a month. He hopes to gather enough evidence to prove false advertising.

- A parent-teacher association for schools in Austin, Minnesota were wondering how pervasive drug culture was among their high school students, compared to the national average. In order to gain a handle on the situation, they handed out a survey to the students at a school assembly during homecoming week.

- In a randomized controlled double blind study, 12 people in the treatment group died before receiving the treatment, so the researchers decided to omit them from the data analysis.

Bottom line: Know which biases show up and when, particularly how the randomized controlled double blind studies can eliminate ALL sources of them.

Know your errors

Type I Error: This is when the null hypothesis is true, but we reject it nonetheless. Sometimes this is called a false positive.

Type II Error: This is when the null hypothesis is false, but we fail to reject it. Sometimes this is called a false negative.

Studying suggestions

Make sure to review the homework solutions, and try to go through the practice quiz on Dr. Breheny's website without the solutions, then check how well you did. Under the notes section of the website, there is a column denoting which lecture each quiz will cover. You should expect at least one question per lecture.

Any remaining time will be for questions. This concludes lab 4.