Summary statistics and graphics: Continuous Data

Patrick Breheny

September 8, 2016

In last week's lab, we explored the Titanic data set and how to describe, graph, and explore categorical data. In this lab, our data set will have both continuous and categorical variables, and we'll learn the tools in `R` for describing, graphing, and exploring the distribution of continuous variables as well as relationships between two continuous variables, and between continuous and categorical variables.

Our 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 in a restaurant (`tips.txt`).

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

1 Summary statistics

Total bill (`TotBill`) is an important continuous variable in our data set. As we discussed in class, a common approach to summary statistics for continuous variables is the two-number summary mean ± SD:

```
> mean(tips$TotBill)

[1] 19.78594

> sd(tips$TotBill)

[1] 8.902412
```

We can get percentile-based summaries as well:

```
> median(tips$TotBill)

[1] 17.795

> IQR(tips$TotBill)

[1] 10.78

> fivenum(tips$TotBill)

[1]  3.070 13.325 17.795 24.175 50.810

> quantile(tips$TotBill) ## Same thing

     0%     25%     50%     75%    100%
 3.0700 13.3475 17.7950 24.1275 50.8100
```

```
> quantile(tips$TotBill, seq(0,1,.1)) ## By tenths

    0%    10%    20%    30%    40%    50%    60%    70%    80%    90%
 3.070 10.340 12.636 14.249 16.222 17.795 19.818 22.508 26.098 32.235
  100%
50.810
```

Now, these summary statistics are for the entire data set. We might be interested in summaries for various subsets instead. This can accomplished either directly with brackets or by using the `by` function:

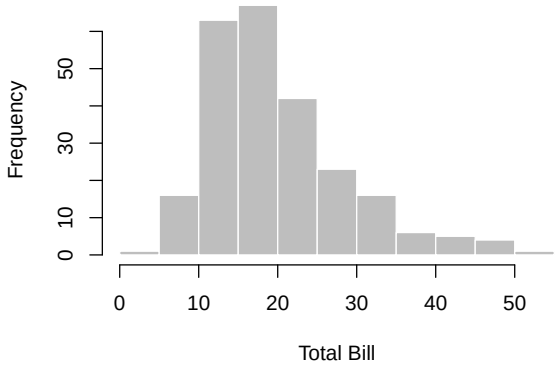
```
> with(tips, mean(TotBill[Time=="Night"]))

[1] 20.79716

> with(tips, by(TotBill, Time, mean))

Time: Day
[1] 17.16868
------------------------------------------------------------
Time: Night
[1] 20.79716

> with(tips, by(TotBill, Time, sd))

Time: Day
[1] 7.713882
------------------------------------------------------------
Time: Night
[1] 9.142029
```

Note the double equal sign (`Time=="Night"`); this tests whether `Time` is equal to the string `"Night"`, as opposed to the single equal sign (`Time="Night"`), which assigns the value `"Night"` to `Time`, which is not what we want to do. Note that nighttime meals have a higher average bill, which makes sense – dinner is usually more expensive than lunch at American restaurants.

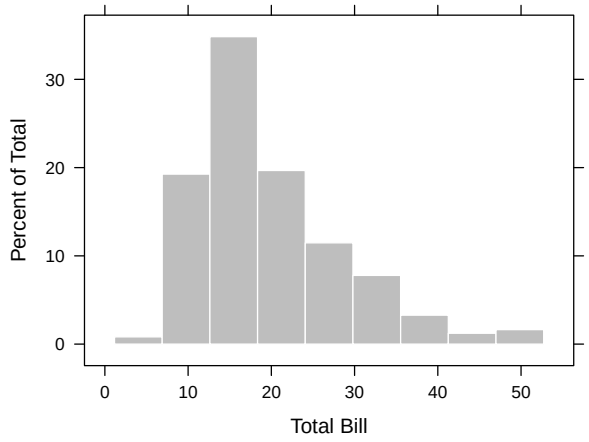
2 Histograms

What does this look like when we plot it? Let's make histograms first. As we've already seen, histograms are created with `hist`. They can also be constructed in `lattice` using `histogram`:

```
> hist(tips$TotBill, col="gray", border="white", main="", xlab="Total Bill")
```

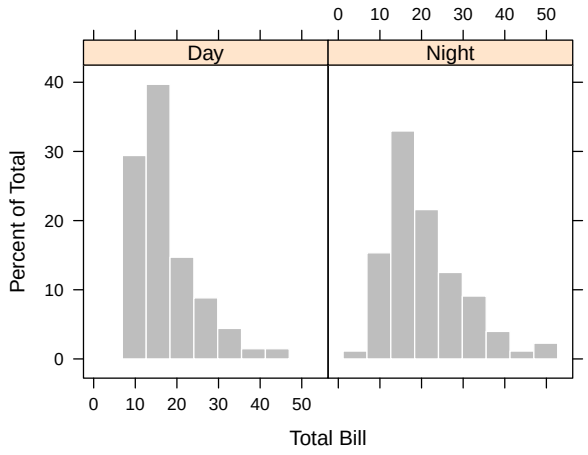



```
> require(lattice)
> histogram(~TotBill, data=tips, col="gray", border="white", xlab="Total Bill")
```

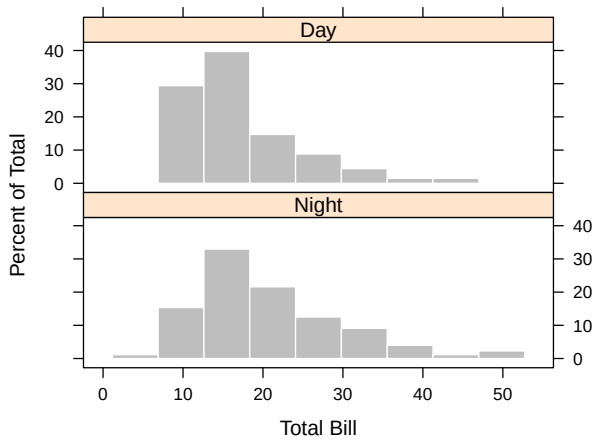


We can see that most bills were around $15, but that some were as high as $50. As we did last week, we can break this plot down by conditioning:

```
> histogram(~TotBill|Time, data=tips, col="gray", border="white", xlab="Total Bill")
```




```
> histogram(~TotBill|Time, data=tips, col="gray", border="white", xlab="Total Bill", layout=c(1,2))
```

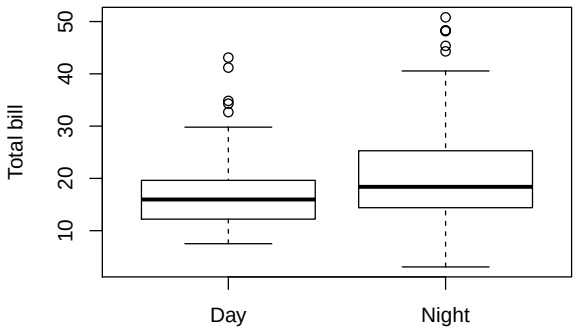


Note that dinners tend to be more expensive and more highly variable; this agrees with our numerical summaries from earlier.

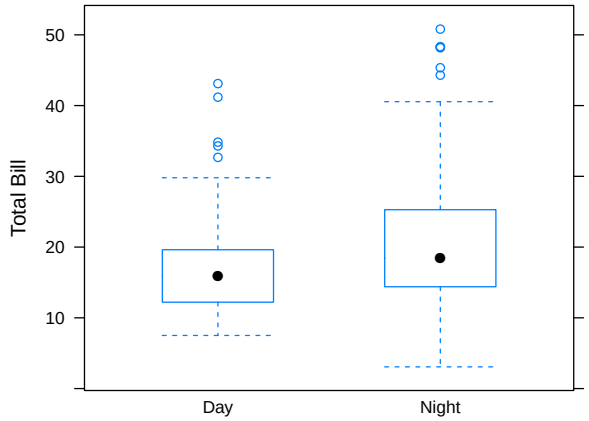
3 Box plots

Box plots are pretty straightforward (again, here are base graphics and lattice versions):

```
> boxplot(TotBill~Time, data=tips, ylab="Total bill")
```




```
> bwplot(TotBill~Time, data=tips, ylab="Total Bill")
```

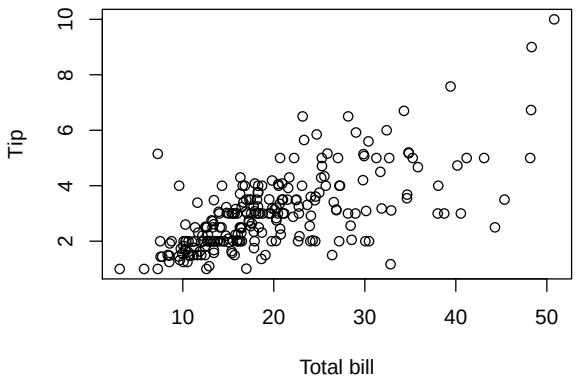


Once again, dinner bills are a little higher and more spread out than lunch bills. Note that `lattice` draws a dot for the median instead of a line.

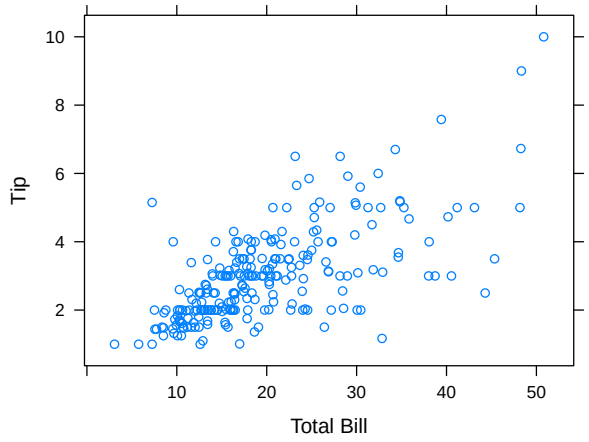
4 Scatter plots

Scatter plots (which we'll discuss more in class when we get to regression and correlation) are made using, simply `plot` (base graphics) or `xyplot` (`lattice`):

```
> with(tips, plot(TotBill, Tip, xlab="Total bill"))
```




```
> xyplot(Tip~TotBill, data=tips, xlab="Total Bill")
```



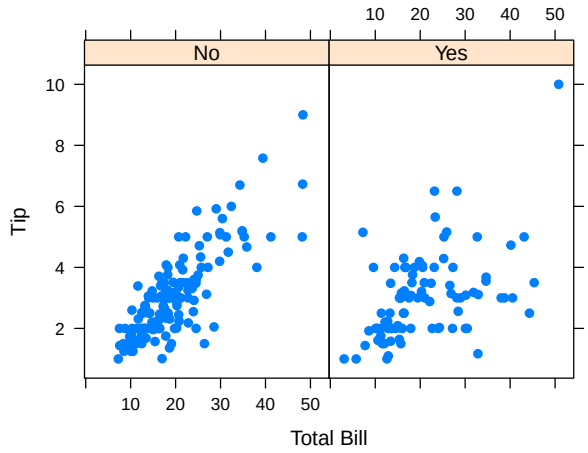
The plot illustrates several trends:

- As we would expect, there is a positive association between 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 – cheap tippers are more common than generous tippers?
- There seem to be some horizontal "stripes" in the plot – why?

Note that none of the earlier summaries showed these stripes – all summaries risk concealing features of the distribution, and each plot illustrates something new about the data.

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, data=tips, xlab="Total Bill", pch=19)
```



The relationship between tip and bill seems to be much stronger in the nonsmoking section than in the smoking section.

5 Tip rate

The most interesting "outcome" in the `tips` data set is probably the tipping rate. In the United States, tip rates usually vary between 10% and 20%, depending on factors such as the quality of the service and the generosity of the customer. Here, since all tips involve the same waiter and restaurant, it would be reasonable to assume that variations in tipping behavior primarily reflect differences in the customers' attitudes.

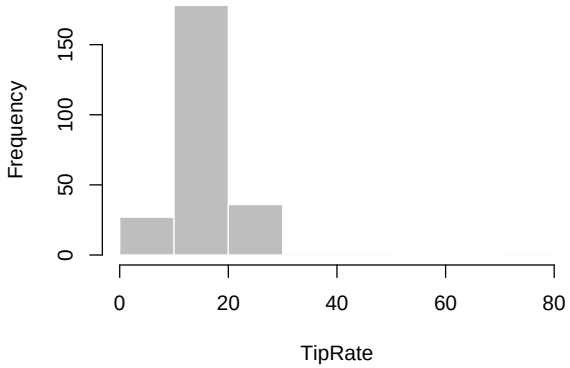
To analyze tip rate, however, we first need to calculate it. You could either create a new variable outside of the `tips` object in `R`, or a new column in the `tips` data frame. I'll create – either way is fine.

converted it to a percent when I multiplied by 100, but you can call it whatever you like and leave it as a fraction if you wish.

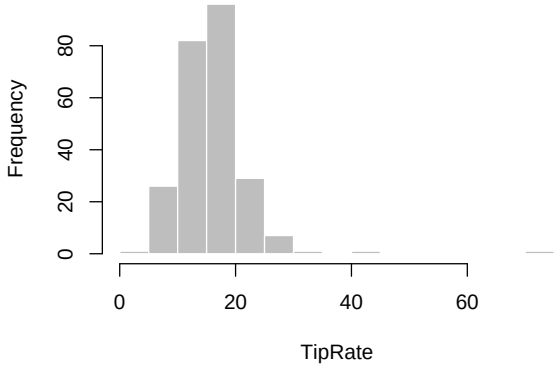
```
> TipRate <- with(tips, 100*Tip/TotBill)
```

Let's check out what our new variable looks like:

```
> hist(TipRate, col="gray", border="white", main="")
```



```
> hist(TipRate, col="gray", border="white", main="", breaks=seq(0,75,5))
```



As we would expect, most tips are between 10 and 20 percent, although there are certainly exceptions.

6 Questions:

Simple questions:

- What percent of tips are above 20%?
- What is the average tip rate that this waiter received?

A more interesting question is how tip rate varies depending on various other factors. In addition, there are many interesting questions one can ask about how the other variables relate to each other. The following list is by no means exhaustive, but contains some questions that I found interesting and looked at. For each question, think about how you would answer the question graphically as well as what numbers you could report that would summarize the trend.

- How does tip rate change with total bill? Do small bills have more variation in tip rate than large bills? Are people proportionally more generous with smaller bills?
- Do smokers tip differently than nonsmokers?
- Suppose that an equal number of men and women dine at the restaurant. Are men more likely to pick up the check than women? Does this depend on whether the meal is lunch or dinner?
- Does tipping behavior change at lunch versus dinner?
- Does tipping behavior differ by days of the week?

There is nothing special about this list; if additional or different questions interest you, feel free to explore them as well.

This is a relatively simple data set, yet provides a wealth of information about a lot of complicated relationships. Just think about how much information is contained in more complicated biomedical studies. Exploring your data to gain an understanding of these relationships is very important!