Personal Expenses

1 Personal Expenses

Most every one lives within a budget. Yet somehow we manage to pay for all the things we want to do (hopefully). The U.S. Department of Labor collects data on consumer expenses and publishes the data for others to use. The data was use by the New York Times in May 2003 to support the headline "New York Pays for the Steak, Boston the Cigar". We'll look at some of this data in a bit, but first, let's get honest and answer some questions about ourselves tracking the data presented in the `cx` dataset. Only answer those you feel comfortable answering.

Problems

? 1. What are your yearly expenses on alcohol? How do you calculate this? For example, did you estimate a monthly *average* and multiply by 12?

? 2. What are your yearly expenses on apparel

? 3. What are your yearly expenses on shoes?

? 4. What are your yearly expenses on entertainment?

? 5. What are your yearly expenses on food?

? 6. What are your yearly expenses on housing?

? 7. What are your yearly expenses on transportation?

? 8. What are your yearly expenses on phone services?

? 9. What are your yearly expenses on utilities?

1.1 Entering data into R

Entering data into `R` is easy and can be done several ways. For instance, if you wanted to enter this data

```
500 500 200 1000 3000 12,000 2000 1200 2500
```

we can use the `c()` function to combine them into a data vector. For example, to store the data into a variable `my.expenses` we would do

```
> my.expenses = c(500, 500, 200, 1000, 3000, 12000, 2000, 1200, 2500)
> my.expenses
[1]   500   500   200  1000  3000 12000  2000  1200  2500
```

We could use the `sum()` function to total these.

Problems

10. Use the `c()` function to make a data vector with all of your expenses above except for shoes and phone services. Now add them up and compare to your yearly income.

11. Is your income more than your expenses above? Are there other large expenses in you budget not appearing above?

2 Where you stand

We want to understand the "shape" of the distributions of expenses for these categories. Before we look at the real data, lets try to think of what the shape should look like.

For example, the amount people spend on alcohol. For starters, we expect a bunch of people who don't drink and therefore would spend close to nothing on alcohol. We would also expect social drinkers to spend relatively little. Finally, a daily drinker will probably spend alot. How much. A quick guess may be $5 to $10 per day or from $1800 to $3600. It would be hard to spend much more, so we expect the distribution to not have too long of a tail.

Problems

12. Does the above sound reasonable? Do a similar *ad hoc* analysis to predict what the shape of the `phone` variable will be.

Now to put our theory to the test, First we read in the data, then we make a histogram of the `alcohol` variable.

```
## do all this on one line, no spaces
> my.url = "http://www.math.csi.cuny.edu/CCLI/Projects/ \
    PersonalExpenses/expenses.R"
> source(file=url(my.url))
> attach(cx)
> hist(alcohol)
> boxplot(alcohol)
```

We see that the tail is not as long as perhaps we guessed. The data we have is already averaged over groups and consequently the extreme values are not well represented. Or, maybe people are

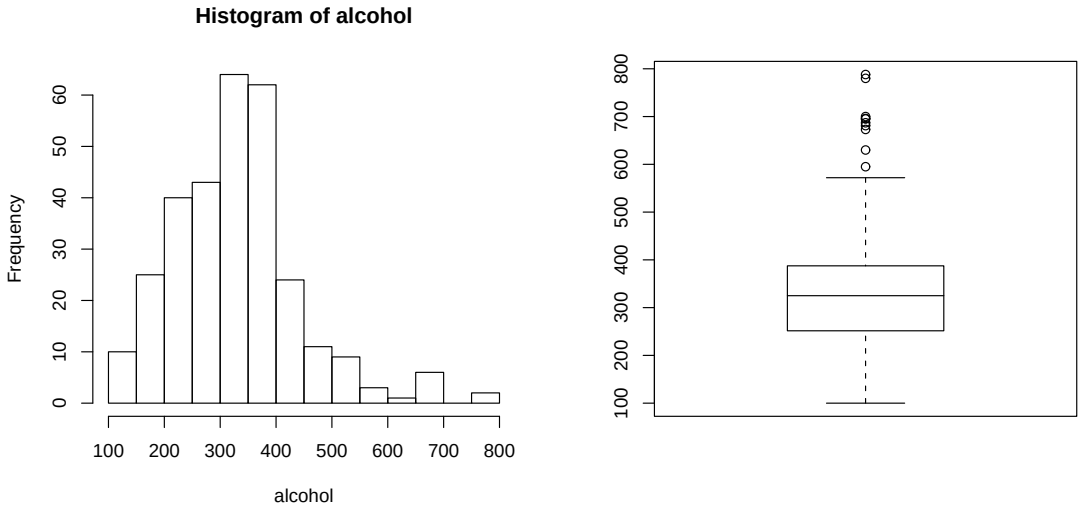

Figure 1: histogram and boxplot of `alcohol`

reluctant to admit how much they really spend on alcohol. Or maybe we overestimated. However, we do see that the distribution is skewed right and is not symmetric.

Problems

? 13. Explain why you might expect the amount spent on `entertainment` to be skewed. Make a histogram and discuss if you were correct.

2.1 Some statistical functions

The computer provides functions to find many familiar statistics such as the functions `mean()` and `median()` which find these familiar measures of center. For example, for the `alchohol()` variable these are

```
> mean(alcohol)
[1] 330.6067
> median(alcohol)
[1] 325
> summary(alcohol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  100.0   252.3   325.0   330.6   387.3   788.0
```

The last line illustrating the useful `summary()` command which finds both of these and more.

Another useful summary is the proportion less than some value. In `R` the count or number less than a value is generically found using `sum()` with a logical comparison. For example

```
> sum(alcohol < 200)
[1] 35
```

says that 35 data points are less than 200. What proportion is this? We need to divide by the number of data points to get the relative frequency or proportion

```
> sum(alcohol < 200)/ length(alcohol)
[1] 0.1166667
```

Which says only 11 percent spend less than 200 dollars.

Problems

? 14. Do you think people spend more or less on phone services than they did 5 years ago. Explain your reasoning.

15. Make a histogram for the `phone` variable. Visually estimate the mean and median. Compare your answers to the calculated values.

16. What percent of the data spends less that $200 on alcohol?

17. What percent spends less than the mean for `alcohol`?

18. What percent spend less than you do on alcohol?

19. For the variables, `housing`, `transportation` and `entertainment`, find out what percent spend less than you.

3 Long tailed distributions

Have a look at a histogram and boxplot of the `income` dataset. It appears to have a skew to the right and a long tail, but the maximum amount is only 107,027 (`max(income)`). Surely, this isn't a very good reflection of the true maximum salary. This is because this dataset "averages" the data over various classes thereby not giving an accurate picture of the true range of the data.

There is another data source which gives a more complete picture of many things. In particular income. The Survey of consumer finances (SCF) is a survey conducted by the Federal Reserve Board (`http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html`). A sampling the current data is available in the `cfb` dataset.

```
> names(cfb)
 [1] "WGT"      "AGE"      "EDUC"     "INCOME"   "CHECKING"
 [6] "SAVING"   "NMMF"     "STOCKS"   "FIN"      "VEHIC"
[11] "HOMEEQ"   "OTHNFIN"  "DEBT"     "NETWORTH"
> attach(fb)
```

Now look at the variable `INCOME` with a histogram and a boxplot. Does it look dramatically different? It should, in fact, you will get better results if you take a *log transformation* of the data first. Try this

```
> log.income = log(1 + INCOME)  # why add 1?
> hist(log.income,prob=T)
> lines(density(log.income))    # add a density
> boxplot(log.income)
```

If you typed the commands above, you will get the following pictures We added a `density`

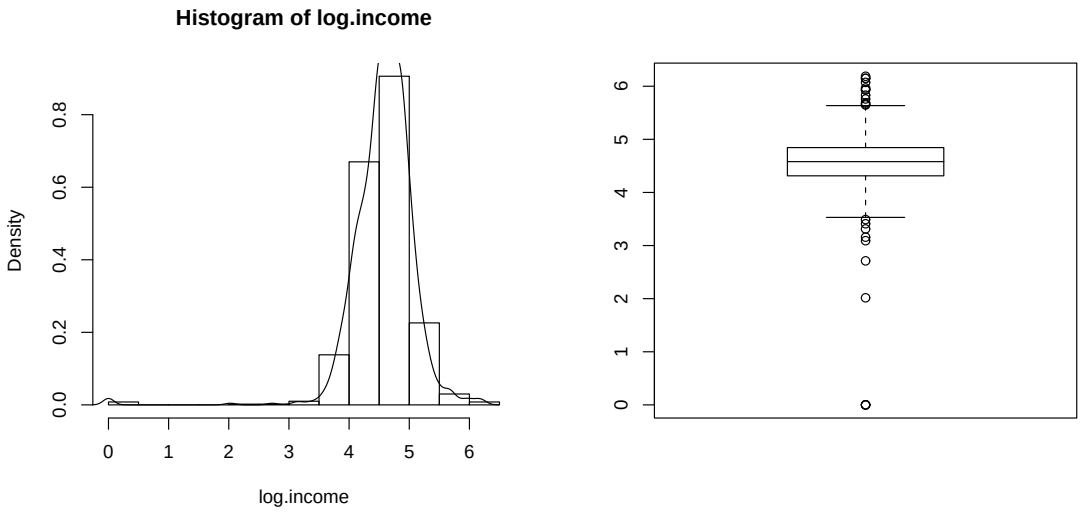

Figure 2: Distribution of the logarithm of income

plot. This is like a frequency polygon, but has more of a mathematical theory. It helps us understand the shape of the distribution. In this case, the data looks bimodal with a small hump near 0 for those with no income.

The raw data is very skewed, that's why taking a logarithm helped. A shift from 5 to 6 is a factor of 10 times more income.

It doesn't make a whole lot of sense to summarize really skewed data with means. Let's see why.

Problems

20. Compare the values of the `mean` and `median` on the `INCOME` data. What do you see. Is this expected. What quantile is the mean? What is the value of the .90 quantile?

21. The `DEBT` variable contains the amount of debt the person has. It is a non-negative number. What percent of this survey have no debt?

22. Is the debt also skewed right? Make a histogram of `DEBT` and one of the logarithm of `DEBT`.

23. Which is more the maximum income or the maximum debt?

4 correlations

The Pearson correlation coefficient, R, is found from the formula

$$R = \frac{\mathbf{COV}(x,y)}{\mathbf{SD}(x)\mathbf{SD}(y)} = \frac{\sum(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

This is found in `R` with the `cor()` function.

If two variables are in a linear relationship, then the (Pearson) correlation coefficient measures in some way the strength of that relationship. However, two variables may be strongly related and yet have no correlation. A simple example is found with a parabola:

```
> x = -2:2
> cor(x,x^2)
[1] 0
```

Clearly `x` and `x^2` are related but the correlation is 0.

The Spearman correlation coefficient relaxes this problem somewhat, by first ranking the data, and then applying the correlation coefficient. If the two data sets are in a monotonic relationship, then the Spearman correlation coefficient will be close to 1 or -1. This is found in `R` by first using `rank()` and then `cor`. A function to do so is

```
scor = function(x,y) cor(rank(x),rank(y))
```

Before we look at some real data, it may help us to look at some artificial data to see what is going on. A scatterplot should be made to help your understanding. Problems

24. What is the Pearson correlation for `x = rnorm(100)`, `y = rnorm(100)`?

25. What is the Pearson correlation for `x = rnorm(100)`, `y = x + rnorm(100)`? Explain why this is different from the previous.

26. Let `x = 0:5`. What are the Pearson and Spearman coefficients for `y = x^5`. Explain.

For the `cfb` dataset, many of the variables are correlated. This makes sense, if for example, a person has above average home equity then they will likely have above average net worth. Let's see the difference Problems

27. Make a scatterplot of `HOMEEQ` and `NETWORTH`. Find both the Pearson and Spearman correlation coefficients and compare. A plot of the logarithm of `HOMEEQ` versus the logarithm of `NETWORTH` reveals a strong linear trend.

28. Find the Pearson and Spearman correlations for the two variables `INCOME` and `VEHIC` (the amount of value in vehicles). Comment on their values based on a scatterplot. Does a transformation make sense here?