

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 used 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 entertainment?
- ? 3. What are your yearly expenses on food?
- ? 4. What are your yearly expenses on housing?
- ? 5. What are your yearly expenses on transportation?
- ? 6. What are your yearly expenses on phone services?

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
```

Once data is entered into a variable, we can apply functions to it. For example, `sum()`, `mean()` and `median()` to find the sum, the mean and median.


```
> sum(my.expenses)
[1] 22900
> mean(my.expenses)
[1] 2544
> median(my.expenses)
[1] 1200
```

The big difference between the mean and median is due to the one large value of 12,000.

Problems

 7. 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.

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

 9. If you spent 200, 175, 167, 96, 229, 183 on entertainment for 6 months, what would your monthly average be? What about an estimated yearly amount?

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, let's 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 a lot. 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

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

2.1 Reading in a dataset

Now to put our theory to the test, First we read in the data using the `url()` function to read a file from a webpage.

```
## do all this on one line, no spaces
> f = "http://www.math.csi.cuny.edu/st/R/PersonalExpenses.R"
> source(url(f))
```



This loads a dataset `cx`. It's big. To see what variables are included in it we use `names()`

```
> names(cx)
[1] "housing"           "transportation" "alcohol"
[4] "entertainment"    "income"         "phone"
[7] "utilities"        "apparel"        "food"
[10] "shoes"
```

So there are 10 variables. To look at any one of them it is easiest to first attach the dataset so the names are available.

```
> attach(cx)
```

Now we can access the variables by their name.

For example, we make a histogram and boxplot of the `alcohol` variable as follows.

```
> hist(alcohol)
> boxplot(alcohol)
```

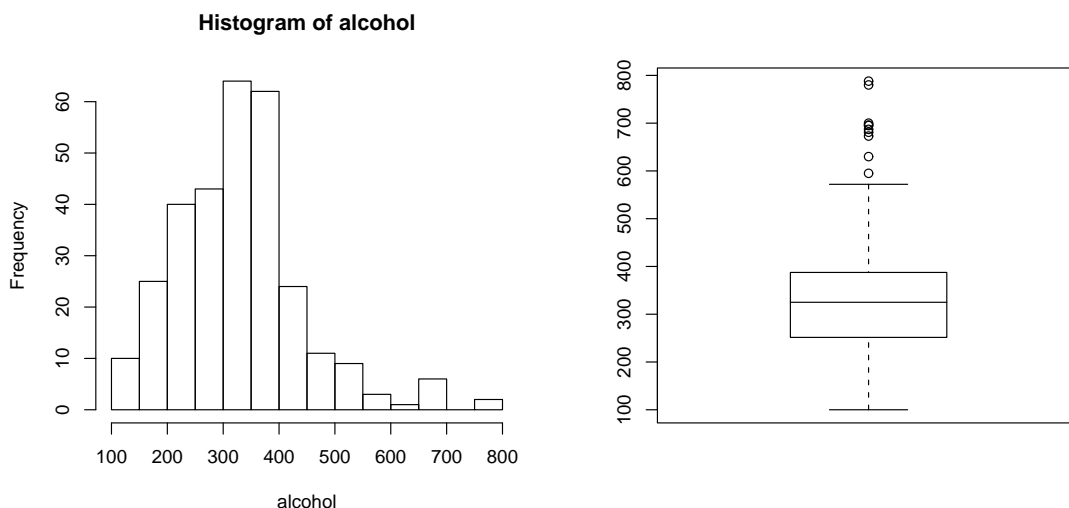


Figure 1: histogram and boxplot of `alcohol`

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 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

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

🖥️ 12. Make a histogram of `housing` and `food`. Which has a wider range? Do they have similar shapes?



2.2 Position in a dataset

The mean and median measure the center of a dataset. The `summary()` command will return both.

```
> summary(alcohol)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
100.0  252.3   325.0   330.6   387.3   788.0
```

Notice it also returns the min, max and quartiles. These are values that split the dataset into quarters, roughly speaking. The 1st quartile is where 25% of the data is to the left and 75% to the right. Quartiles can be found directly using the `quantile()` function.

```
> quantile(alcohol, .25)
25%
252.2
```

The quantile is related to a percentile. A percentile is a data value that splits the data by a certain percent. The quantile is the same only using a scale of 0 to 1 instead of 0 to 100. For example, the command

```
> quantile(alcohol, .9)
90%
461.7
```

says that 461.7 is the value with 90% of the data less and 10% more.

Quantiles, answer which data point has a given amount less. The reverse is to find what percent or proportion is less than some value. This can be found by adding all the values less and dividing by the total number of values.

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.

Alternately, you may want to include the value of 200. That is you want less *or equal*. This is done using `<=`


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




Which in this case gives the same answer.

Problems

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

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

 15. What percent of the data spends less than \$200 on `entertainment`?

 16. The quintiles split the data into 5 pieces. They can be found on the variable `x` by `quantile(x, c(.2, .4, .6, .8))`. Find these for the `housing` variable.

 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 of 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(fcb)
```

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, 10) # 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`

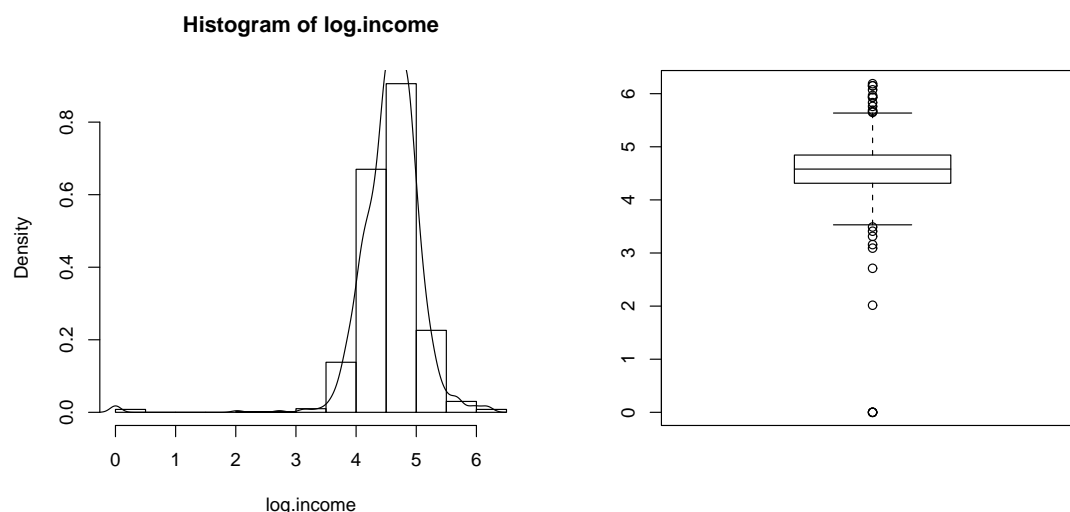


Figure 2: Distribution of the logarithm of income

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. The `log()` function with the extra argument 10 means base 10 for the logarithm. This means 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+1`.
23. Which is more the maximum income or the maximum debt?

