



# Crackers

Getting hungry just thinking about learning some statistical analysis? If you are like many, you might reach for a box of crackers as a “healthy alternative” to more sugar-laden foods. But just how healthy are crackers? For instance, the Los Angeles Unified School District Obesity Prevention Motion identified Pepperidge Farm’s Cheddar Goldfish as an unapproved snack food. <sup>1</sup> To investigate what is contained in a crackers box, we learn the tools of exploratory data analysis.

## 1 The crackers data set

We’ll use a data set containing various nutritional information (see Table 1) gleaned from the website [www.dietfacts.com](http://www.dietfacts.com) and the sides of cracker boxes at a supermarket in Oberlin Ohio. <sup>2</sup>

To access the data we download it from a website by issuing the command:

```
> crackers = read.csv("http://www.math.csi.cuny.edu/st/R/crackers.csv")
```

The variable `crackers` contains the data. The data set contains measurements on 11 variables with 92 different types. How might we view this data? We can see all the values at once by typing the name of the data set, `crackers`. However, the values will quickly scroll by. The `edit()` function provides a better alternative:

```
> edit(crackers)
```

A basic spread-sheet like window should open showing the data set. A warning: *you will be unable to proceed until this window is closed.*



Question 1: Scroll through the `Company` variable and note any companies you do not recognize.



Question 2: Which company seems to have the most cracker products on the shelves? Guess how many different products are on the shelves?

What we would like to be able to do is access the data in the variables. In order to do this easily, we `attach()` the data set:

```
> attach(crackers)
```

---

<sup>1</sup><http://cafe-la.lausd.k12.ca.us/usnacks.htm>

<sup>2</sup> This data set has been contributed to the *Journal of Statistical Education*, <http://www.amstat.org/publications/jse/>, by Professor Carolyn Cuff of Westminster College. Many of the questions in this project follow those outlined in her accompanying paper. Some of the variable names were shortened from the original. The data is census data for all crackers sold at this supermarket. Can you tell what the supermarket was?

Now we can refer to the variables by name, such as `Product` (case is important).

The numeric variables are all per-serving measurements. To see what variables are available, the variable names can be read off from the spread sheet, or printed out using the `names()` function:

```
> names(crackers)
```

```
[1] "Company"           "Product"           "Crackers"
[4] "Grams"            "Calories"          "Fat.Calories"
[7] "Fat.Grams"        "Saturated.Fat.Grams" "Sodium"
[10] "Carbohydrates"    "Fiber"
```

For instance, the variable `Crackers` records the crackers per serving and the variable `Fat.Calories` records the calories due to fat per serving.

<b>Nutrition Facts</b>	
Serving Size 55 crackers (55g)	
Servings Per Container 12	
Amount Per Serving	
<b>Calories</b> 150	Calories From Fat 60
<b>%Daily Values*</b>	
<b>Total Fat</b> 6g	<b>11%</b>
Saturated Fat 1.5g	<b>7%</b>
<b>Sodium</b> 250mg	<b>10%</b>
<b>Total Carbohydrates</b> 19g	<b>6%</b>
Dietary Fiber 0g	<b>0%</b>
* Percent Daily Values are based on a 2,000 calorie diet.	

Table 1: Example of a nutritional label. This one for Pepperidge Farm's Cheddar Goldfish.

## 2 Numeric variables

Most of the variables in the data set are numeric. For such data, we have numeric summaries and graphical summaries available. Each has their place. Graphical summaries allow us to quickly see several features of a distribution at once, whereas numeric summaries allow us to quantitatively compare our data to some other data set or pre-conceived notion about our data.

## 2.1 Numeric summaries

Numeric summaries we use summarize the center (`mean()` and `median()`), the spread (`range()`, `sd()` and `IQR()`), or even the position within a data set (`quantile()` and `scale()`).

For example, looking at the `Crackers` variable, we can compare centers with

```
> mean(Crackers)
[1] 13.35870
> median(Crackers)
[1] 6.5
```

The difference leads us to believe the data set is not symmetric. The spread is summarized with


```
> sd(Crackers)
[1] 14.18603
> IQR(Crackers)
[1] 11
```


The IQR is computed from the 0.25 and 0.75 quantiles, which may be found with


```
> quantile(Crackers, c(0.25, 0.75))
25% 75%
 5  16
```

The 95th percentile would be returned with

```
> quantile(Crackers, 0.95)
95%
46.35
```

 Question 3: Apply the functions `mean()`, `median()`, `range()`, `IQR()`, and `sd()` above to the `Calories` data set. What values do you get?

 Question 4: Of the five numbers found in the previous exercise, which one is not available from the output of `summary()` applied to the data set `Calories`.


 Question 5: The grams per serving variable, `Grams`, has missing data. You can verify this by typing the command

```
> Grams
```

The missing values are coded `NA`, read “not available.” When there are missing values, the extra argument `na.rm=TRUE` is often needed (this removes `NA` values). Verify that `mean(Grams)` is not what is wanted, but

```
> mean(Grams, na.rm = TRUE)
```

computes the desired answer.

 Question 6: For the `Calories` data, find the 5th and 95th percentiles.

## 2.2 Graphical summaries of numeric data

There are a number of graphical means to summarize a data set. For instance stem-and-leaf diagrams, dotplots, histograms, densities, and boxplots. All of these devices allow us to quickly visualize the following: the center of the distribution, a sense of spread, the minimum value of the data, the maximum value, the range, where the bulk of the data sits, if there are any values far from the bulk, and the general shape of the data.

The R functions used to produce these graphics are `stem()`, `stripchart()`, `hist()`, `density()`, and `boxplot()`

## 2.3 Stem and leaf diagrams

Stem and leaf diagrams are produced in R using `stem()`.<sup>3</sup>

For instance, a stem-and-leaf diagram of the calories per serving is produced with:

```
> stem(Calories)
```

```
The decimal point is 1 digit(s) to the right of the |
```

```

4 | 00
6 | 00000000000000000000000000000000
8 | 000000000000
10 | 00
12 | 00000000000000
14 | 000000000000000000000000000000
16 | 000

```

From the diagram we can see that the smallest recorded number is 40, the largest 160, the range is 120. The “shape” of this data set is *bimodal*—that is, there are two distinct “peaks” (around 60 and 140).



Question 7: Issue the command

```
> stem(Grams)
```

To make a simple stem-and-leaf diagram of the grams per serving. What is the range of values, as read from the stem-and-leaf diagram? Describe the shape of the data set.



Question 8: Verify, that the command `stem(Calories)` actually truncates some of the data, by comparing the diagram with the output of `range(Calories)`. Explain.





Question 9: The choice of stem is automatically determined by `stem()`. It may not always produce the best results. However, you can override the choice by using an extra `scale=` argument. For instance, try the command

<sup>3</sup>An alternative to this is to use the `stem.leaf()` function in the `Rcmdr` package. If present, this is loaded with the command `library(Rcmdr)`. This package provides a GUI to the R workspace. Alternatively, you can download a copy of the function with the command `source("http://www.math.csi.cuny.edu/st/R/stem.leaf.R")`.

```
> stem(Grams, scale = 1/2)
```

Find the range of the data and compare to your previous answer. Does this make a better stem and leaf diagram than before? Explain why?

 Question 10: Remake a stem and leaf diagram of **Calories** finding a value for **scale=** that uses a stem recording the tens digits.

 Question 11: Make a stem-and-leaf diagram of the crackers per serving variable **Crackers**. What is the range of the data? This shape of this data set is skewed right. What type of cracker do you think has 55 or more per serving? Check your answer by quickly scrolling through the data set.

## 2.4 Dotplots

A dotplot can be produced with the `stripchart()` function.<sup>4</sup>

The simplest use of `stripchart()` will not stack points when there are ties, you must ask for this behavior. For example, to make the dotplot (Figure 1) of the data in **Grams** (grams per serving), we issue the following command:<sup>5</sup>

```
> stripchart(Grams, method = "stack")
```

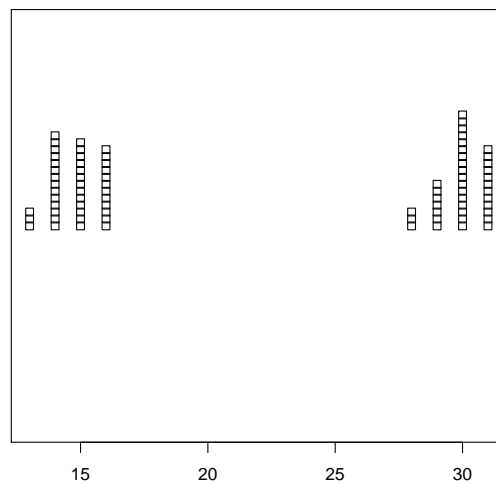


Figure 1: Dotplot of grams per serving

<sup>4</sup> Alternatively, you can use the `DOTplot()` function that may be installed with the command `source("http://www.math.csi.cuny.edu/st/R/DOTplot.R")`

<sup>5</sup> If you use `DOTplot()` the graphic is produced by `DOTplot(Grams)`.

Some things we can quickly see from Figure 1 are that the range is roughly 13 to 31, the mean is around 20 (balance point), the median is somewhere in the left cluster of values, and the shape is bimodal. To check our guesses on the mean and median we have:

```
> summary(Grams)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
13.00	15.00	16.00	22.11	30.00	31.00	11.00



Question 12: Make a stripchart of the data in the variable **Calories**.

1. What is the range of the data?
2. What is the shape of the data?
3. Estimate the mean of the data using the idea of balancing.
4. Are there any crackers that are “average” by this measure?
5. Check your estimate, by computing the mean.

## 2.5 Histograms

The dotplots made with this much data are pretty busy, a histogram may better show the key features of the data set.

Histograms are made using the function `hist()`<sup>6</sup>, as in its use to produce a histogram of the amount of sodium per serving:

```
> hist(Sodium)
```



Question 13: Based on the histogram in Figure 2 do the following:

1. Estimate the range of the data
2. Estimate the mean of the data set
3. Estimate the median of the data set
4. Describe the shape of the data

Check your numeric answers using the appropriate function.



Question 14: Produce a histogram of the number of crackers per serving. Based on the histogram do the following

1. Estimate the range of the data

---

<sup>6</sup>Or, one can use the function `truehist()` from the MASS package. This is available after loading the package, which can be done with the command `library(MASS)`

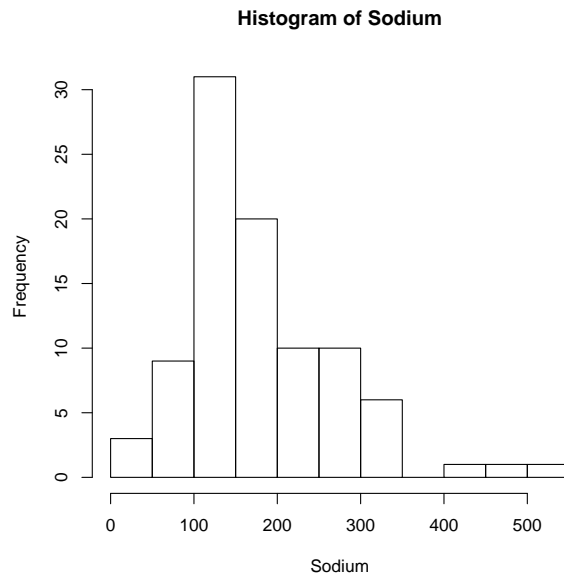


Figure 2: Histogram of sodium amount in crackers.

2. Estimate the mean of the data set
3. Estimate the median of the data set
4. Describe the shape of the data

Check your numeric answers using the appropriate function.

## 2.6 Boxplots

Boxplots succinctly describe a numeric data set in a manner that lends itself to multiple comparisons. From a boxplot we can quickly identify all of the following: the center, the spread, the range, symmetry or skew, and tail length.

Boxplots are produced with the `boxplot()` function. For example, a boxplot of the number of crackers per serving (Figure 3) may be made with

```
> boxplot(Crackers)
> title("Number of crackers per serving")
```



Question 15: From Figure 3 answer the following:

1. What is the “center” of the variable?
2. What is the spread?
3. Is the data set skewed?

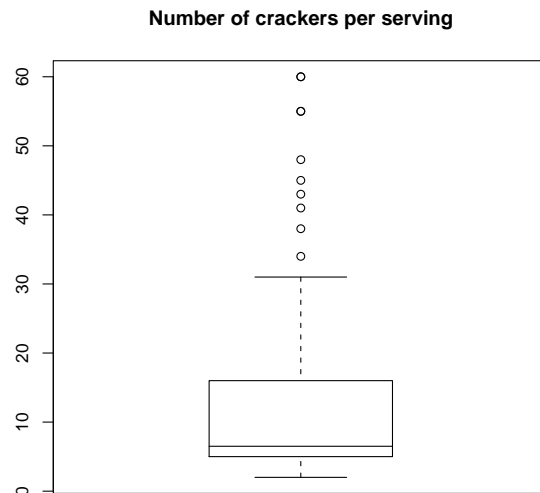




Figure 3: Boxplot of number of crackers per serving

4. Do you expect the mean or median to be the largest? Why? Check it.
5. Are there any “outliers?” If so, how many?

 Question 16: Make a boxplot of the `Fat.Grams` variable. Compare the values plotted in the boxplot with those produced by the command <sup>7</sup>

```
> summary(Fat.Grams)
```

 Question 17: If there are two variables, one numeric and one categorical (numeric will also work) of the same size, it is easy to produce parallel boxplots for each level of the categorical variable. The syntax is to use `numeric ~ categorical` for the argument. For instance: <sup>8</sup>

```
> boxplot(Calories ~ Company, las = 2)
```

Based on your plot, answer the following:

1. Which company has the widest range?
2. Which company has the largest median?
3. Which company has the largest IQR?
4. Why do some companies have just a short horizontal line?

<sup>7</sup>The `summary` command finds the actual quantiles, whereas the boxplot technically uses the related—but sometimes different—hinges.

<sup>8</sup>The extra argument `las=2` is optional. It turns the labels so that they are easier to read. You can also enlarge the plot window to see more text.

### 3 Bivariate analysis

What can we glean from the crackers data set when we look at two numeric variables simultaneously. Can we see what determines the calories per serving? Is there a relationship between `Sodium()` and `Fiber()`? etc. To answer these it helps to look at bivariate relationships.

#### 3.1 Scatterplots

Looking at two numeric variables simultaneously is often done using a *scatterplot*. These are produced in R with the `plot()` function. This function can be used several ways. To make a scatterplot of the variables `x` and `y`, we'll use it with an argument like:

```
plot(y ~ x).
```

If only a subset of the data is desired, you can specify this by the indices, or using a logical expression using syntax like

```
plot(y ~ x, subset= ...).
```

Simply replace `...` by a logical expression, or a data vector of indices.

For example, a scatterplot of grams per serving on the `x` axis and calories per serving on the `y` axis is done, as follows (Figure 4):

```
> plot(Calories ~ Grams)
```

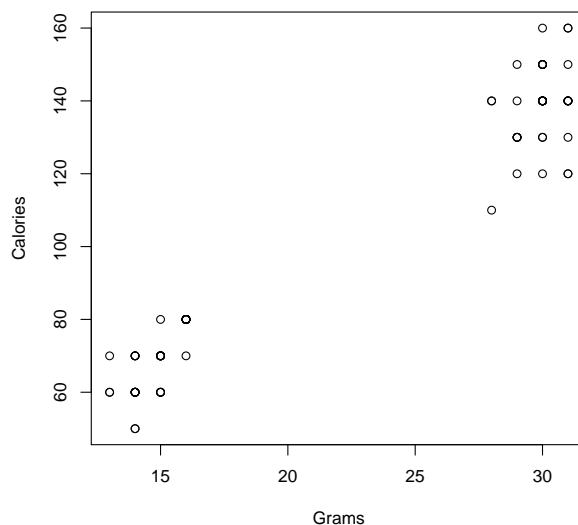


Figure 4: Scatterplot of grams per serving predicting calories per serving


The scatterplot shows two distinct clusters of data. Within each cluster there appears to be very little trend.




Question 18: Verify that the command

```
> plot(Calories ~ Grams, subset = Calories >= 100)
```

plots just the upper cluster. What subset command using `Grams` instead of `Calories` will produce this same graphic?


 Question 19: Make a scatterplot with `Crackers` on the  $x$  axis and `Calories` on the  $y$  axis. Does there appear to be any trends?

 Question 20: From your graph of `Crackers` versus `Calories` there is an “outlier” around (45,90). To identify that point in the data set can be done using the mouse. Type the command

```
> identify(Crackers, Calories, labels = Product)
```


Now click on a point, and the point will be labeled with the corresponding product name. Right click to stop the process.<sup>9</sup>


What product is the outlier?

 Question 21: As an exploratory device, multiple scatterplots can be made at once in a graphic called a scatterplot matrix. The `pairs()` function will do so. For example, to make all pairs of scatterplots of the variables 5 through 11 we have the command

```
> pairs(crackers[5:11])
```

Which relationship is closest to a straight line?

 Question 22: Make a scatterplot with `Carbohydrates` on the  $x$  axis and `Calories` on the  $y$  axis. Does the amount of carbohydrates per serving seem to affect the number of calories per serving?

 Question 23: Make an indicator variable, `low.carb`, which is `TRUE` if the number of carbohydrates per serving is less than 15. Make parallel boxplots of the number of calories per serving broken up by the values of `low.carb`. Explain the differences in the boxplot.

## 3.2 Correlation

The Pearson correlation coefficient is a numeric summary of the strength of a linear relationship between two variables. The Spearman correlation coefficient is a numeric summary of the *monotonic* relationship between two variables. They are both computed by the `cor()` function. The default is to return the Pearson coefficient. When the extra argument `method="spearman"` is used the Spearman coefficient is returned.


For instance the correlation between the calories per serving and the carbohydrates per serving is computed with


```
> cor(Calories, Carbohydrates)
```


```
[1] 0.8670158
```


---

<sup>9</sup>In Mac OS X you may need to hit the ESC key.

 Question 24: What is the correlation between `Fat.Grams` and `Fat.Calories`? Make a scatterplot and guess the correlation first.

 Question 25: What is the correlation between the crackers per serving (`Crackers`) and calories per serving (`Calories`)?

 Question 26: Make a scatterplot of crackers per serving (`Crackers`) and calories per serving (`Calories`). Does the relationship appear to be linear? Monotonic? If you said yes to monotonic, compute the Spearman correlation coefficient and compare to the Pearson correlation coefficient found in the previous question.

 Question 27: The correlation between `Calories` and `Carbohydrates` is positive. However, a scatterplot shows two distinct clusters. These are indicated by the indicator variable defined as

```
> low.carb = Carbohydrates < 15
```

Find the correlation for just the low-carbohydrate data and compare to that for the non-low-carbohydrate data. Then compare to the value of 0.867 to the two just found. This is an example of data with two clusters throwing off the interpretation of correlation.

(This can be done with syntax like `Calories[low.carb]` or `Calories[!low.carb]`.)