

Chapter 7: Describing Distributions with Numbers

Overview: In Chapter 6, we saw methods of visualizing the distribution of numerical variables with histograms. We also briefly discussed characteristics of a distribution such as variability, symmetry and skewness. With a concept such as variability, we have not seen how to precisely capture this with a numerical value. In this chapter, we will learn not only how to compute variability, but several numerical values that help us describe a distribution.

Motivating Example: Consider the *airquality* data set that is built into R. This data set consists of daily readings of air quality values from May 1, 1973 to September 30, 1973 in New York. The variables in this data set are Mean Ozone (in parts per billion), Solar radiation, Average Wind speed, and Maximum daily Temperature (measured in degrees Fahrenheit). All variables are numerical and their distributions can be visualized by creating histograms. How else can we quantify their distributions?

We begin with some definitions:

Definitions of Median and Quartiles:

- The **median** M is

How to calculate the median:

1. Arrange all observations from smallest to largest.
2. Select the middle observation. If there is no middle (because the number of observations is even) then take the average of the two middle values. This is the median.

Example: Consider a random sample of 10 observations of the Temperature variable from the *airquality* data set:

```
12 temp.samp = sample(airquality$Temp,10)
13 temp.samp
14
15 # | ``
    [1] 83 83 63 74 83 79 61 82 85 62
```

Determine the median of this sample:

Note: What we just computed was a **sample median** (thus it was an observed value of a statistic \tilde{x}). If we had the median of the entire population then that would be the value of the **population parameter** M .

- The **first quartile** $Q1$ of a distribution is

- The **third quartile** $Q3$ of a distribution is

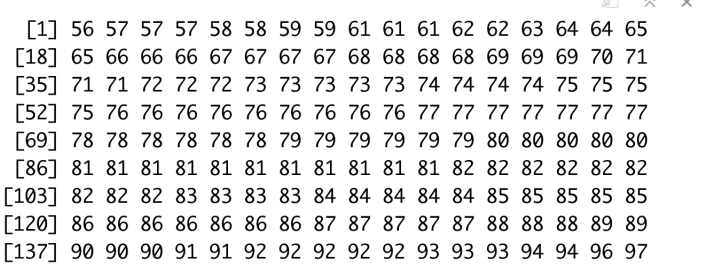
Note: There are more than one way to compute quartiles (more on that later). For our purpose, we will not include the median M in your calculation of the quartiles $Q1$ and $Q3$.

Example: Determine the quartiles $Q1$ and $Q3$ for the previous sample of 10 temperature observations.

How to compute median and quartiles in R: There are many ways to determine the median and the quartiles of a distribution in R. We will now see a few of those ways:

1. The most manual way to do this is sort the observations using R and then hard-code the values you are looking for:


```
21 sort(airquality$Temp)
22
23
24 ^ ` ` `
```



```
[1] 56 57 57 57 58 58 59 59 61 61 61 62 62 63 64 64 65
[18] 65 66 66 66 67 67 67 67 68 68 68 68 69 69 69 70 71
[35] 71 71 72 72 72 73 73 73 73 73 74 74 74 74 75 75 75
[52] 75 76 76 76 76 76 76 76 76 76 77 77 77 77 77 77 77
[69] 78 78 78 78 78 78 79 79 79 79 79 79 80 80 80 80 80
[86] 81 81 81 81 81 81 81 81 81 81 81 81 82 82 82 82 82
[103] 82 82 82 83 83 83 83 84 84 84 84 84 85 85 85 85 85
[120] 86 86 86 86 86 86 86 87 87 87 87 87 88 88 88 89 89
[137] 90 90 90 91 91 92 92 92 92 92 93 93 93 94 94 96 97
```

We could then determine the middle value of the observations:

```
21 middle = (length(airquality$Temp)+1)/2
22
23 median = sort(airquality$Temp)[middle]
24 median
25
26
27 ^ ` ` `
```




```
[1] 79
```

Similarly, you could determine the quartile values. However, this is not a smart way to do this. There are built-in functions in R which will compute these values for you in far fewer steps with far less thinking required.

2. Use the median() and quantile() functions:


```
40 median(airquality$Temp)
41
42 ^ ` ` `
```



```
[1] 79
```



```
40 quantile(airquality$Temp,0.25)
41
42 ^ ` ` `
```



```
25%
72
```

```

39
40 quantile(airquality$Temp,0.75)
41
42 ^ ` ` `

```

75%
85

3. It turns out, the `quantile()` functions default (without specifying which quartile you are looking for) is even more useful:

```

40 quantile(airquality$Temp)
41
42 ^ ` ` `

```

0%	25%	50%	75%	100%
56	72	79	85	97

4. One other function that is very useful is the `summary()` function:

```

40 summary(airquality$Temp)
41
42 ^ ` ` `

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
56.00	72.00	79.00	77.88	85.00	97.00

Practice Question: Consider the *airquality* data set and the *wind* variable.

- Using whatever method you would like, determine the median wind speed.
 (A) 7.4 (B) 9.7 (C) 9.958 (D) 11.5
- Using whatever method you would like, determine the *Q1* value for wind speed.
 (A) 7.4 (B) 9.7 (C) 9.958 (D) 11.5
- Using whatever method you would like, determine the *Q3* value for wind speed.
 (A) 7.4 (B) 9.7 (C) 9.958 (D) 11.5

Definition: The **five number summary** of a distribution consists of

Practice Question: Which R function gives exactly the five-number summary (and no additional information):

(A) median()

(B) quantile()

(C) summary()

Definition: A **boxplot** is

How to make a Boxplot:

1. Draw a y axis that has range spanning at least from the min value to the max value.
2. Draw a box which has a base at y -value $Q1$ and the top at y -value $Q3$.
3. Draw a horizontal line in the box at y -value M .
4. Draw vertical lines extending out of bottom of the box down to the min value and out of the top of the box up to the max value.

Example: Create a boxplot of the five-number summary for the temperature variable in the *airquality* data set.

While we can use median to describe the center of a distribution, and the quartiles (along with the min and max values) can be used to describe the variability of the distribution, this is not typically how statisticians go about describing center and variability.

Typically, we describe the center of a distribution by calculating the **mean** and we describe the variability of a distribution by calculating the **standard deviation**.

We've already seen how to compute the mean (both by hand and using R). Now we will see how to compute the standard deviation.

Definition: The **standard deviation**

Notation: The population standard deviation (which is a parameter) is represented by the greek letter sigma: σ

The sample standard deviation (which is a statistic) is represented by: s

How to Compute s (by hand):

1. Compute the sample mean \bar{x} .
2. Find the distance of each observation from the mean and square each of these distances.
3. Add up all of the squared distances found in step 2, and divide them by $n - 1$ (where n is the sample size). Note: This value is called the **sample variance** (which is denoted by s^2).
4. Take the square root of the value found in step 3. This is the sample standard deviation s .

Example: Compute the sample standard deviation for a sample consisting of the values:

20, 15, 23, 10, 18

Note: If you had access to observations from the entire population, then you could compute the population mean μ and use that to compute the population standard deviation σ . However, the formula is slightly different than that for the sample standard deviation.

Computing Sample standard deviation in R:

There is a function `var()` which computes the sample variance s^2 :

```
53 # Sample variance s^2
54 var(airquality$Temp)
55
56 ▶ ```
[1] 89.59133
```

There is a function `sd()` which computes the sample standard deviation s :

```
53 # Sample standard deviation s
54 sd(airquality$Temp)
55
56 ▶ ```
[1] 9.46527
57
```

In the (extremely) rare case where you have access to the entire population and want to compute the population variance, then you can multiply the sample variance by $(n - 1)/n$:

```
52
53 n = length(airquality$Temp)
54
55 # Population variance sigma^2
56 var(airquality$Temp)*(n-1)/n
57
58 ▶ ```
[1] 89.00577
```

Note: Since standard deviation is computed using the mean \bar{x} we can only use standard deviation to measure variability if we are using the mean to describe the center. If we were using the median to describe the center then we would use quartiles to describe the variability.

Interpreting Standard Deviation: The standard deviation is a measure of variability in the population. It tells us how spread apart the values of the variable are.

Recall: If the population has large variability, we need to increase the size of our sample in order to accurately represent the population.

Question: What does it mean to have $s = 0$?

Answer:

When to use Median vs When to use Mean:

- If your data contains outliers, then you should use the median to describe the center (and thus quartiles to describe the variability).

The reason why is that the mean is heavily influenced by outliers whereas the median is not. For example, consider the following sample of observations:

100, 107, 98, 20, 105

```
64 sample = c(100,107,98,20,105)
65 mean(sample)
66 median(sample)
67
68 ^ ```
    [1] 86
    [1] 100
```

- If your distribution is reasonably symmetric (no strong skew to the left or right) then you should use the mean to describe the center (and thus use the standard deviation to describe variability) as many more statistical results rely on those values.