

Chapter 12: Describing Distributions with Numbers

Overview: In Chapter 11, we saw additional methods of visualizing the distribution of numerical variables with histograms and stemplots. 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 the 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 **the midpoint of the distribution. It is the number such that half (50%) of the observations are smaller and half are larger.**

**The median is one way to measure the centre of the distribution.
(sample mean is another measure of the centre)**

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:

61 62 63 74 79 | 82 83 83 85
 └─┘

$$\tilde{x} = \text{median} = \frac{79+82}{2} \simeq 80.5$$

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** Q_1 of a distribution is the median of the observations which fall to the left of the median

Q_1 is the value such that 25% of the observations fall below it.

- The **third quartile** Q_3 of a distribution is the median of the observations which fall to the right of the overall median

Q_3 is the value such that 75% of the observations fall below it.

Note: It is important that you do not include the median M in your calculation of the quartiles Q_1 and Q_3 .

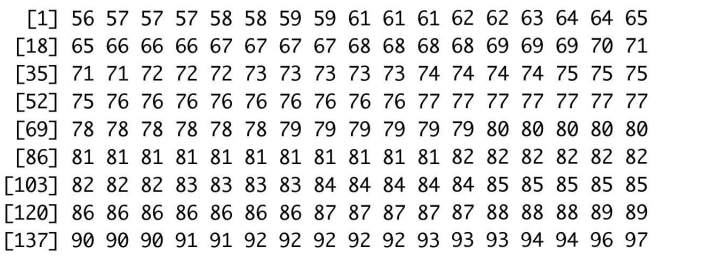
Example: Determine the quartiles Q_1 and Q_3 for the previous sample of 10 temperature observations.

61 62 63 64		83 83 83 85
$Q_1 = \frac{62+63}{2} = 62.5$		$Q_3 = \frac{83+83}{2} = 83$

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:

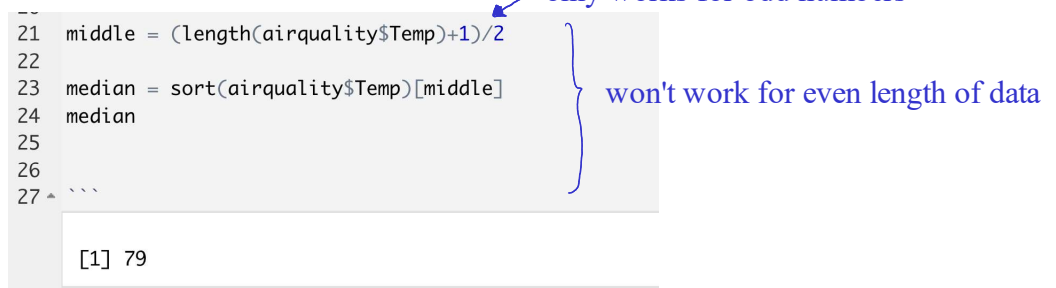
1. The most manual way to do this is sort the observations using R and then hard-code the values you are looking for:

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



We could then determine the middle value of the observations:

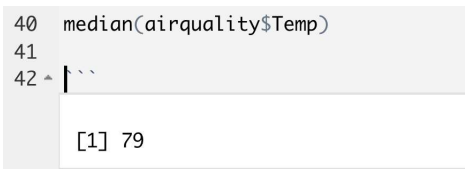
```
21 middle = (length(airquality$Temp)+1)/2
22
23 median = sort(airquality$Temp)[middle]
24 median
25
26
27 ^ ` ` `
```



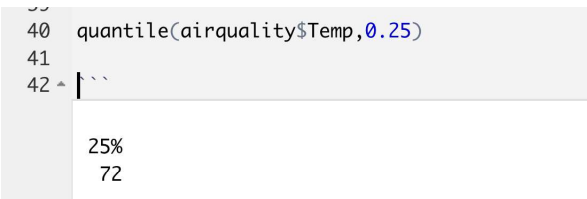
Similarly, you could determine the quartile values. However, this is not a smart way to do this. There are built-in functions in R which will compute these values for you in far fewer steps with far less thinking required.

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

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



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



```
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

min Q1 median Q3 max

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.

1. Using whatever method you would like, determine the median wind speed.

(A) 7.4 (B) 9.7 (C) 9.958 (D) 11.5
2. 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
3. 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

1. min
2. Q_1
3. median
4. Q_3
5. max

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

(A) median()

(B) ☒ quantile()

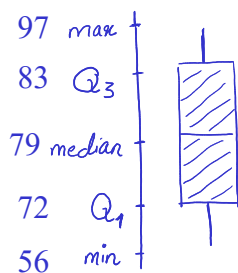
(C) summary()

Definition: A **boxplot** is **graph of visualizing grouped data**

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 Q_1 and the top at y -value Q_3 .
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.



or



does not tell you where the mean is.

box and whisker plot

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** measures the "average" distance of the observations from the mean.

For example: if we have a mean of 10 and a standard deviation of 3, then we were to compute the distance from each observation to 10 (the mean), then the average of those distances, you would get roughly 3.

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

$$\textcircled{1} \quad \bar{x} = \frac{20+15+23+10+18}{5} = 17.2 \qquad \textcircled{3} \quad \frac{7.84+4.84+33.64+51.84+0.64}{5-1} = 24.7 \equiv s^2$$

$$\textcircled{2} \quad \begin{aligned} (20-17.2)^2 &= 7.84 \\ (15-17.2)^2 &= 4.84 \\ (23-17.2)^2 &= 33.64 \\ (10-17.2)^2 &= 51.84 \\ (18-17.2)^2 &= 0.64 \end{aligned} \qquad \textcircled{4} \quad s = \sqrt{s^2} = \sqrt{24.7} \approx 4.97$$

Sample Standard
deviation

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.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \left| \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right.$$

N = population size n = sample size

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) ← sample variance
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) ← sample standard deviation
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 $(n-1)/n$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

[1] 89.00577

$$\sigma^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \right] \frac{n-1}{n}$$

$s^2 \frac{n-1}{n} = \sigma^2$

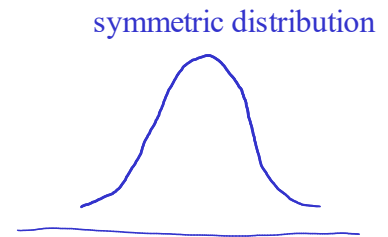
Note: Since standard deviation is computed using the mean \bar{x} we can only use standard deviation to measure variability if we are using the mean to describe the center. If we were using the median to describe the center then we would use quartiles to describe the variability.

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: All observations are exactly the same



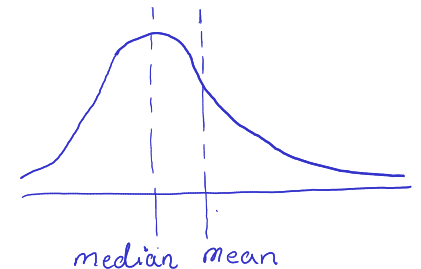
When to use Median vs When to use Mean:

- 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
```



better center

- 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.