

Topic 2: Describing Distributions with Numbers*

August 30, 2011

We first look at **quantitative data**. Recall that in this case, these data can be subject to the operations of arithmetic. We can add or subtract observation values, we can sort them and rank them from lowest to highest. We will look at two fundamental properties of these observations, some measure of the center, the **median** or the **mean**, and associated to this, some description of how these observations are spread about this given measure of center.

For the median, the central observation if the data is sorted from the lowest to highest observations, we often give the smallest and largest observations as well as the observed value that is 1/4 and 3/4 of the way up this list. For the mean, we commonly use the **standard deviation** to describe the spread of the data.

These concepts will be described in more detail in this section.

1 Measuring Center

1.1 Medians

The **median** take the middle value for x_1, x_2, \dots, x_n after the data has been sorted from smallest to largest,

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}.$$

($x_{(k)}$ is called the k -th **order statistic**. Sorting can be accomplished in **R** by using the `sort` command.)

If n is odd, then this is just the value of $x_{((n+1)/2)}$. If n is even, then the two values closest to the center are averaged.

$$\frac{1}{2}(x_{(n/2)} + x_{(n/2)+1}).$$

1.2 Means

For a collection of numeric data, x_1, x_2, \dots, x_n , the **sample mean** is the numerical average

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

Alternatively, if the value x occurs $n(x)$ times in the data, then use the distributive property to see that

$$\bar{x} = \frac{1}{n} \sum_x xn(x) = \sum_x xp(x), \quad p(x) = \frac{n(x)}{n}$$

So the mean \bar{x} depends only on the proportion of observations $p(x)$ for each value of x .

Example 1. For the data set $\{1, 2, 2, 2, 3, 3, 4, 4, 4, 5\}$,

$$1 + 2 + 2 + 2 + 3 + 3 + 4 + 4 + 4 + 5 = 1(1) + 2(3) + 3(2) + 4(3) + 5(1) = 30.$$

Thus, $\bar{x} = 30/10 = 3$.

*© 2011 Joseph C. Watkins

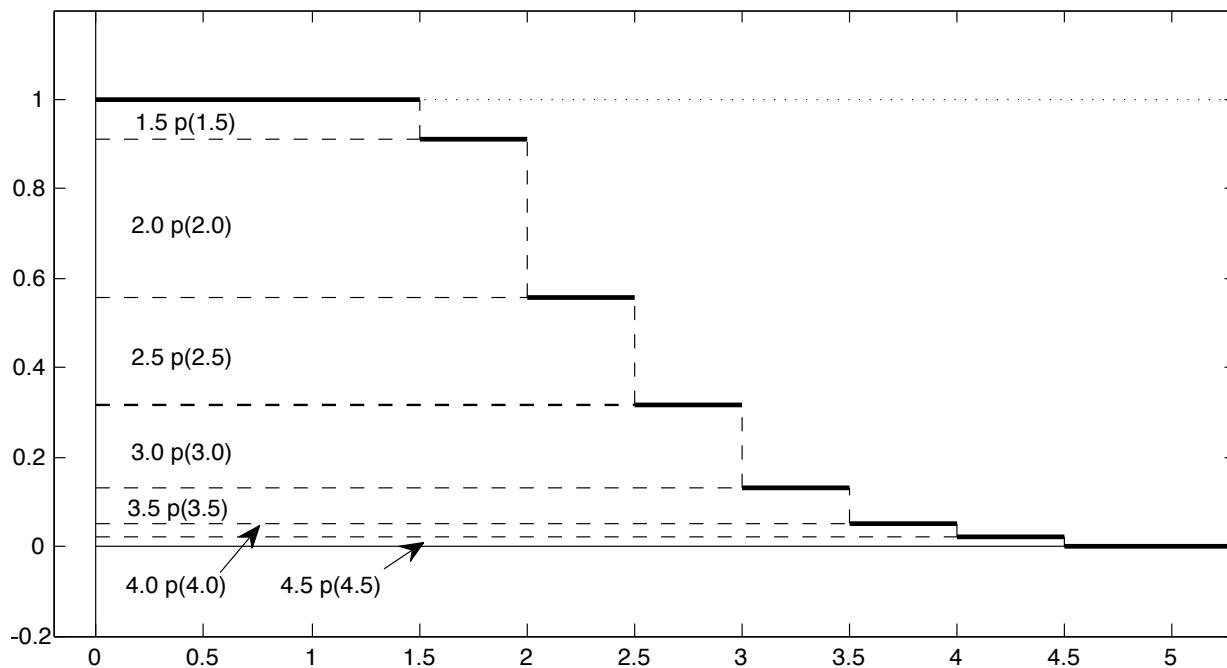


Figure 1: Empirical Survival Function for the Bacterial Data. This Figure displays how the area under the survival function to the right of the y -axis and above the x -axis is the mean value \bar{x} for non-negative data. For $x = 1.5, 2.0, 2.5, 3.0, 3.5, 4.0,$ and 4.5 . This area is the sum of the area of the rectangles displayed. The width of each of the rectangles is x and the height is equal to $p(x)$. Thus, the area is the product $xp(x)$. The sum of these areas are presented in Example 2 to compute the sample mean.

Example 2. For the data on the length in microns of wild type *Bacillus subtilis* data, we have

length x	frequency $n(x)$	proportion $p(x)$	product $xp(x)$
1.5	18	0.090	0.135
2.0	71	0.355	0.710
2.5	48	0.240	0.600
3.0	37	0.185	0.555
3.5	16	0.080	0.280
4.0	6	0.030	0.120
4.5	4	0.020	0.090
sum	200	1	2.490

So the sample mean $\bar{x} = 2.49$.

If we store the data in R in a vector x , we can write `mean(x)` which is equal to `sum(x)/length(x)` to compute the mean.

Exercise 3. Let \bar{x}_n be the sample mean for the quantitative data x_1, x_2, \dots, x_n . For an additional observation x_{n+1} , use \bar{x} to give a formula for \bar{x}_{n+1} , the mean of $n + 1$ observations. Generalize this formula for the case of k additional observations x_{n+1}, \dots, x_{n+k}

Many times, we do not want to give the same **weight** to each observation. For example, in computing a student's grade point average, we begin by setting values x_i corresponding to grades ($A \mapsto 4, B \mapsto 3$ and so on) and giving weights w_1, w_2, \dots, w_n equal to the number of units in a course. We then compute the **grade point average** as a **weighted mean**. To do this:

- Multiply the value of each course by its weight $x_i w_i$. This is called the number of quality points for the course.
- Add up the quality points:

$$x_1 w_1 + x_2 w_2 + \dots + x_n w_n = \sum_{i=1}^n x_i w_i$$

- Add up the weights, i. e., the number of units attempted:

$$w_1 + w_2 + \dots + w_n = \sum_{i=1}^n w_i$$

- Divide the total quality points by the number of units attempted:

$$\frac{x_1 w_1 + x_2 w_2 + \dots + x_n w_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}. \quad (1)$$

If we let

$$p_j = w_j / \sum_{i=1}^n w_i$$

be the **proportion** or **fraction** of the weight given to the j -th observation, then we can rewrite (1) as

$$\sum_{i=1}^n x_i p_i.$$

If we store the weights in a vector w , then we can compute the weighted mean using `weighted.mean(x, w)`

If an extremely high observation is changed to be even higher, then the mean follows this change while the median does not. For this reason, the mean is said to be *sensitive to outliers* while the median is not. To reduce the impact of extreme outliers on the mean as a measure of center, we can also consider a **truncated mean** or **trimmed mean**. The p trimmed mean is obtained by discarding both the lower and the upper $p \times 100\%$ of the data and taking the arithmetic mean of the remaining data.

In R, we write `mean(x, trim = p)` where p , a number between 0 and 0.5, is the fraction of observations to be trimmed from each end before the mean is computed.

Note that the median can be regarded as the 50% trimmed mean.

2 Measuring Spread

2.1 Five Number Summary

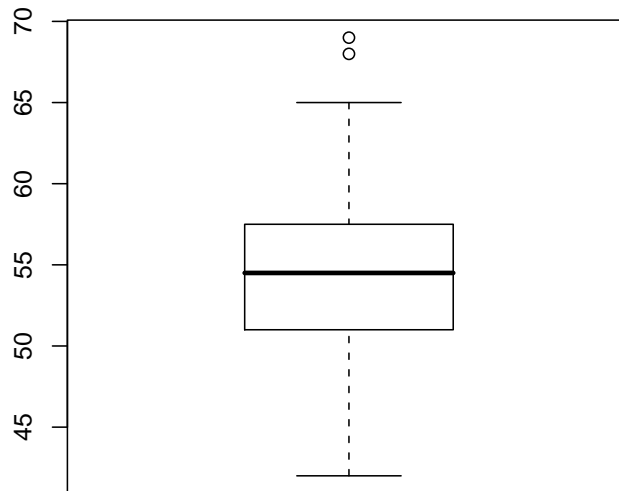
The **first** and **third quartiles**, Q_1 and Q_3 , are, respectively, the median of the lower half and the upper half of the data. The **five number summary** of the data are the values of the minimum, Q_1 , the median, Q_3 and the maximum. These values, along with the mean, are given in R using `summary(x)`. Returning to the data set on the age of presidents:

```
> summary(age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
42.00  51.00   54.50   54.64  57.25   69.00
```

We can display the five number summary using a **boxplot**.

```
> boxplot(age, main = c("Age of Presidents at the Time of Inauguration"))
```

Age of Presidents at the Time of Inauguration

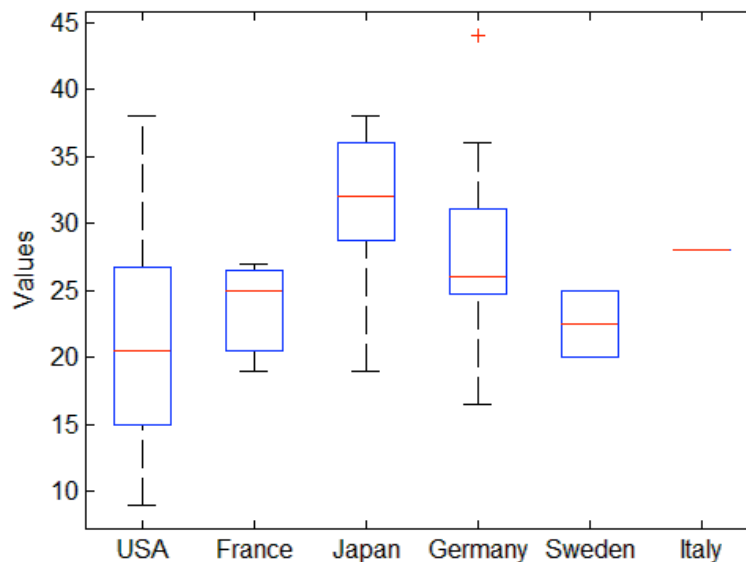


The value $Q_3 - Q_1$ is called the **interquartile range**. It is found in **R** with the command `IQR`. Outliers are somewhat arbitrarily chosen to be those above $Q_3 + \frac{3}{2}IQR$ and below $Q_1 - \frac{3}{2}IQR$. With this criterion, the ages of William Henry Harrison and Ronald Reagan, considered outliers, are displayed by the two circles at the top of the boxplot.

Example 4. Consider a two column data set. Column 1 - `MPH` - gives car gas milage. Column 2 - `origin` - gives the country of origin for the car. We can create side by side boxplots with the command

```
> boxplot(MPG, Origin)
```

to produce



2.2 Sample Variance and Standard Deviation

The **sample variance** averages the square of the differences from the mean

$$\text{var}(x) = s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The **sample standard deviation**, s_x , is the square root of the sample variance. We shall soon learn the rationale for the decision to divide by $n-1$. However, we shall also encounter circumstances in which division by n is preferable. We will drop the subscript x if there is no ambiguity.

We can use the distributive property to simplify the calculation of the sample variance.

Example 5. For the data set on *Bacillus subtilis* data, we have $\bar{x} = 498/200 = 2.49$

length	frequency	length - $\bar{\text{length}}$	(length - $\bar{\text{length}}$) ²	product
1.5	18	-0.99	0.9801	17.6418
2.0	71	-0.49	0.2401	17.0471
2.5	48	0.01	0.0001	0.0048
3.0	37	0.51	0.2601	9.6237
3.5	16	1.01	1.0201	16.3216
4.0	6	1.51	2.2801	13.6806
4.5	4	2.01	4.0401	16.1604
sum	200			90.4800

So the sample variance $s_x^2 = 90.48/199 = 0.4546734$ and standard deviation $s_x = 0.6742947$.

To accomplish this in R

```
> bacteria<-c(rep(1.5, 18), rep(2.0, 71), rep(2.5, 48), rep(3, 37), rep(3.5, 16), rep(4, 6),
+ rep(4.5, 4))
> length(bacteria)
[1] 200
> mean(bacteria)
[1] 2.49
> var(bacteria)
[1] 0.4546734
> sd(bacteria)
[1] 0.6742947
```

Exercise 6. For the data set x_1, x_2, \dots, x_n , let

$$y_i = ax_i + b.$$

Give the summary statistics for the y data set given the corresponding values of the x data set. (Consider carefully the consequences of the fact that a might be less than 0.)

Among these, the **quadratic identity**

$$\text{var}(ax + b) = a^2 \text{var}(x)$$

is one of the most frequently used and useful in all of statistics.

Exercise 7. Show that $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

We next perform a little algebra. We record this for use in later sections. This is similar to the computation of the parallel axis theorem in physics. This is used to determine the moment of inertia of a rigid body about any axis, given the moment of inertia of the object about the parallel axis through the object's center of mass (\bar{x}) and the perpendicular distance between the axes. In this case, we are looking at the rigid motion of a finite number of equal point masses.

Define the sum of squares about the value α ,

$$SS(\alpha) = \sum_{i=1}^n (x_i - \alpha)^2.$$

Then,

$$\begin{aligned} SS(\alpha) &= \sum_{i=1}^n ((x_i - \bar{x}) + (\bar{x} - \alpha))^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \alpha) + \sum_{i=1}^n (\bar{x} - \alpha)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - \alpha)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \alpha)^2. \end{aligned}$$

By Exercise 6, the cross term above $2 \sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \alpha)$ equals to zero. Thus, we have *partitioned the sums of squares* into two levels. The first term gives the sums of squares about the sample mean \bar{x} . The second gives square of the difference between \bar{x} and the chosen value α . We shall see this idea of partitioning in other contexts.

Note that the minimum value of $S(\alpha)$ takes place by minimizing the second term. This takes place at $\alpha = \bar{x}$. Thus,

$$\min_{\alpha} SS(\alpha) = SS(\bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Our second use for this identity provides an alternative method to compute the variance. Take $\alpha = 0$ to see that

$$SS(0) = \sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2. \quad \text{Thus, } \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$$

Divide by $n - 1$ to see that

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

Exercise 8. Show that $SS'(\bar{x}) = 0$ and $SS''(\bar{x}) > 0$ to see that SS has a minimum at the value $\alpha = \bar{x}$.

3 Quantiles and Standardized Variables

A single observation, say 87 on an exam, gives little information about the performance on the exam. One way to include more about this observation would be to give the value of the empirical cumulative distribution function. Thus,

$$F_n(87) = 0.7223$$

tells us that about 72% of the exam scores were below 87. This is sometimes reported by saying that 87 is the 0.7223 **quantile** for the exam scores.

This is computed using the R command `quantile`. For the ages of presidents at inauguration, we have that the 72% quantile is 57 year old.

```
> quantile(age, 0.72)
72%
57
```

Thus, for example, for the ages of the president, we have that `IQR(age)` can also be computed using the command `quantile(age, 3/4) - quantile(age, 1/4)`. R returns the value 7.

Another, and perhaps more common use of the term quantiles is a general term for partitioning ranked data into equal parts. For example, quartiles partition the data into 4 equal parts. Percentiles partition the data into 100 equal parts. Thus, the k -th q -tile is the value in the data for which k/q of the values are below the given value. This naturally leads to some rounding issues which leads to a large variety of small differences in the definition.

Exercise 9. For the example above, describe the quintile, decile, and percentile of the observation 87.

A second way to evaluate a score of 87 is to related it to the mean. Thus, if the mean $\bar{x} = 76$. Then, we might say that the exam score is 11 points above the mean. If the scores are quite spread out, then 11 points above the mean is just a little above average. If the scores are quite tightly spread, then 11 points is quite a bit above average. Thus, for comparisons, we will sometimes use the **standardized version** of x_i ,

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

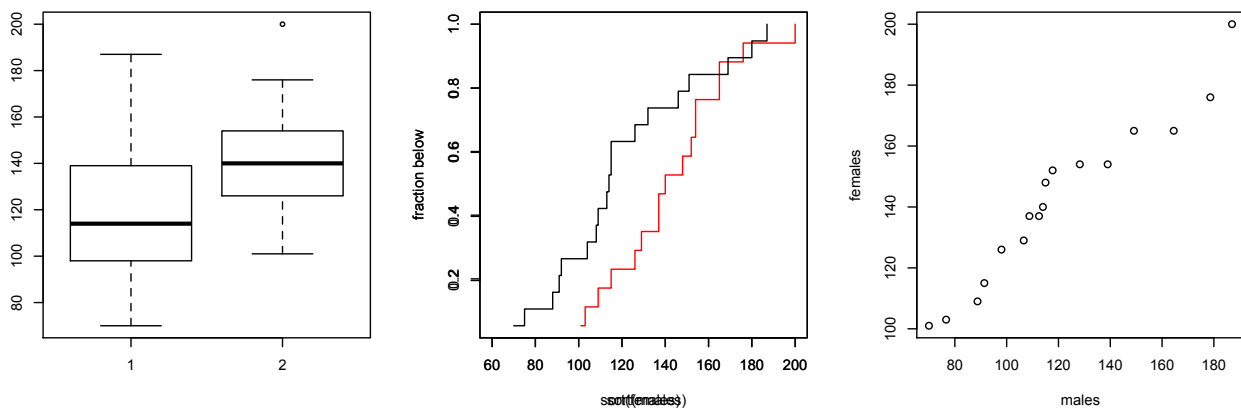
The observations z_i have mean 0 and standard deviation 1. The value z_i is also called the **standard score**, the **z-value**, the **z-score**, and the **normal score**. An individual z-score, z_i , gives the number of standard deviations an observation x_i is above (or below) the mean.

Exercise 10. What are the units of the standard score? What is the relationship of the standard score of an observation x_i and $y_i = ax_i + b$?

4 Quantile-Quantile Plots

In addition to side by side boxplots or histograms, we can also compare two cumulative distribution function directly with the **quantile-quantile** or **Q-Q plot**. If the quantitative data sets x and y have the same number of observations, then this is simply `plot(sort(x), sort(y))`. In the case the Q-Q plot matches each of the quantiles for the two data sets. If the data sets have an unequal number of observations, then observations from the larger data are reduced by interpolation to create data sets of equal length and the Q-Q plot is `plot(sort(x), sort(y))` for the reduced data set.

Example 11. The Survey of Study Habits and Attitudes is a psychological test that measures motivation, attitude toward school, and study habits. Scores range from 0 to 200. Below are side by side boxplots, empirical cumulative distribution function, and Q-Q plots. The data sets and the R code are given below.



```
> females<-c(154,109,137,115,152,140,154,176,101,103,126,137,165,165,129,200,148)
> males<-c(108.140,114,91,180,115,126,92,169,146,109,132,75,88,113,151,70,
115,187,104)
> par(mfrow=c(1,3))
> plot(sort(females),1:length(females)/length(females),type="s",
ylab=c("fraction below"),xlim=c(60,200),col="red")
> par(new=TRUE)
> plot(sort(males),1:length(males)/length(males),type="s",ylab=c("fraction below"),
xlim=c(60,200))
> qqplot(males,females)
```

5 Answers to Selected Exercises

3. Check the formula

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}.$$

For k additional observations, write

$$\bar{x}_{n,k+n} = \frac{1}{k}(x_{n+1} + \cdots + x_{n+k}).$$

Then the mean of the $n + k$ observations is

$$\bar{x}_{n+k} = \frac{n}{n+k}\bar{x}_n + \frac{k}{n+k}\bar{x}_{n,k+n}.$$

6.

statistic	
median	If m_x is the median for the x observations, then $am_x + b$ is the median of the y observations.
mean	$\bar{y} = a\bar{x} + b$
variance	$\text{var}(y) = a^2\text{var}(x)$
standard deviation	$s_y = a s_x$
first quartile	If Q_1 is the first quartile of the x observations and if $a > 0$, then $aQ_1 + b$ is the first quartile of the y observations. If $a < 0$, then $aQ_3 + b$ is the first quartile of the y observations.
third quartile	If Q_3 is the third quartile of the x observations and if $a > 0$, then $aQ_3 + b$ is the third quartile of the y observations. If $a < 0$, then $aQ_1 + b$ is the third quartile of the y observations.
interquartile range	$IQR(y) = a IQR(x)$.

7. Divide the sum into 2 terms.

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n \left(\frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \right) = 0.$$

8.

$$S(\alpha) = \sum_{i=1}^n (x_i - \alpha)^2. \quad \text{Thus,} \quad S'(\alpha) = -2 \sum_{i=1}^n (x_i - \alpha)$$

and $S'(\bar{x}) = 0$. Next, $S''(\alpha) = 2n$ for all α and thus $S''(\bar{x}) = 2n > 0$. Consequently, \bar{x} is a minimum.

9. 87 is between the 3-rd and the 4-th quintile, between the 7-th and the 8-th decile and the 72-nd and 73-rd percentile.

10. Both the numerator and the denominator of the z -score have the same units. Their ratio is thus unitless. The standard score for y ,

$$z_i^y = \frac{y_i - \bar{y}}{s_y} = \frac{(ax_i + b) - (a\bar{x} + b)}{|a|s_x} = \frac{a(x_i - \bar{x})}{|a|s_x} = \frac{a}{|a|} z_i^x.$$

Thus, if $a > 0$, the two standard scores are the same. If $a < 0$, the two standard scores are the negative of one another.