

Learning Objectives

By reading and studying these notes you should be able to

1. Describe the difference between a sample and a population
2. Manually create a dot plot for small and modest sample sizes
3. Describe the qualitative properties of a sample mean and a sample standard deviation
4. Compute the mean and standard deviation of a sample

Random Variables

A random variable is a quantity that is not a single value. The variable may have some tendency to hover around a typical value, or it may be so random that there is no one value to describe the physical quantity being observed. The generic random variable is often given the symbol¹ X . If there is a discernable trend to the values of X we might write

$$X = \text{“typical } X\text{”} + \text{noise}$$

where *noise* is the random variation that we observe superimposed on the trend in X .

The temperature of the air in your refrigerator is a random variable. The thermostat in the refrigerator helps to keep the temperature nearly constant by turning the cooling system on and off as needed. If you had a temperature sensor and a means of recording the sensor output, you would find that the air temperature fluctuates around the set point of the thermostat. When you open the door of the refrigerator, room air rushes in, momentarily causing the air temperature to rapidly increase. When the door is closed again, the air temperature is reduced toward the set point.

Populations and Samples

Consider the speed of cars moving on a highway. The cars travel at speeds that are related to the speed limit, but it would be very unlikely to find a single car traveling exactly at the speed limit, say at 55.000 MPH. We would say that the speed of the cars on the highway is a random variable. Suppose we measure the speed of 15 cars passing a landmark like an overpass and obtain the following speeds in MPH

50.6, 53.7, 54.5, 55.6, 51.4, 55.3, 61.0, 52.6, 54.1, 54.4, 58.2, 55.6, 53.3, 57.7, 57.8

The obvious question is, what is the average speed? We will answer that question, but before we do, let's consider the data more carefully.

Is this set of speed measurements an accurate indication of the cars travelling on this highway at this time of day? Could it be that this was an unusually slow or an unusually fast group of cars? Did our speed measuring equipment appear to be a speed trap, which caused some of the cars

¹Statisticians tend to use unnecessary capitalization, and we'll stick that convention. However, a lower case x , or for that matter y or ξ could just as well be used.

to temporarily slow down? To avoid too many complications, we will ignore any uncertainty that might be caused by faulty equipment used to measure the speed, or other conditions that might have interfered with obtaining the true speed of each of the cars.

It would be possible to extend this guessing game until we would not trust any measurement. That is not the point. Rather, by asking these questions about the specific data, we provide a concrete example of important terms of statistical analysis.

Definitions

Population:

A population is the set of all possible values for a random variable of interest. For the car speed example, the population is the speeds of *all* cars that pass by the landmark. It is usually not practical to measure the random variable of interest for all members of a population.

Sample:

A sample is a subset of a population selected to measure the random variable of interest. For the car speed example, the sample is the set of 15 speed values that were measured. This sample was taken in a limited amount of time under conditions of weather and traffic load that may or may not be representative of typical traffic conditions for that landmark on that road. We usually select a sample that is a small subset of the population. The measured values obtained from the sample are the data used as input to a statistical analysis.

The quality of the statistical information obtained with a sample can be strongly affected by the size of the sample and the criteria and procedure by which the sample members are selected. A population is unaffected by the size of the sample or the method of obtaining the sample. A key question in assessing the usefulness of a sample is whether that sample is truly representative of the underlying population. Generally, larger samples are better, but in many situations, larger samples are more costly (or take more effort) than smaller samples. Hence there is a trade-off between sample size and how well the sample truly represents the population.

We cannot emphasize this too much: *the choice of the sample will limit the accuracy and relevance of any subsequent statistical analysis.* A good sample allows the possibility of a good statistical analysis. On the other hand, no amount of sophistication can correct the flaws introduced by a bad or non-representative sample. A well-designed statistical study has a sampling protocol (time of day, number of samples, system operating conditions) that are likely to lead to a subset of measurements that are truly representative of the population.

Dot Plots

Plots help us to grasp overall characteristics of the data. Figure 1 is a dot plot of the car speed data. The horizontal locations of the small circles indicate the values of the speed measurements. Some of the data are stacked one above the other. There are two speeds at 55.6. Two other values at 57.7 and 57.8 are close enough that the circles are considered to be at the same speed, and hence are stacked on top of each other.

The dot plot in Figure 1 shows that measured data are clumped, and that the extreme values are separated by gaps from the bulk of the data.



Figure 1: Dot plot of speed measurements.

Histograms

A histogram is a graph that represents the distribution of data by grouping that data into bins of fixed width. A histogram can be obtained by grouping the data from a dot chart. A histogram can also be obtained directly from a sample.

Consider the car speed data from Figure 1. We could divide the data into two groups. For example there are 11 cars traveling at 56 MPH and slower. There are 4 cars traveling faster than 56 MPH. Separating the speeds into two groups is one arbitrary division. Another choice of grouping would be to consider ranges or bins of speeds in 2 MPH increments. Thus, there are 2 cars traveling between 50 and 52 MPH, 3 cars traveling between 52 and 54 MPH, etc. as summarized in the following table. The first row is the speed range and the second row is the number of cars in that speed range

50 to 52	52 to 54	54 to 56	56 to 58	58 to 60	60 to 62
2	3	6	2	1	1

Each column of the table is a bin. For a given sample, changing the bin width will change the number of values (number of cars in this example) in each bin. There are no bins in a dot chart, but in creating a dot chart the diameter of the dots whether two values should be drawn next to each other or on top of each other.

The tabular representation of the binned data can be represented graphically as in Figure 2, which is called a histogram. A histogram is obtained with a two step process. First, the data is grouped into bins of a pre-determined width. Then a vertical bar is drawn to represent the number of values in each bin. Figure 2 is a graphical representation of how the data was grouped into bins.

The shape of a histogram depends on the bin width. This can lead to misleading interpretations of the data for small data sets. Figure 4 shows three histograms of the car speed data obtained with different definitions of bin width (2 or 3) or two different starting points for the bins (49 or 50) with the same bin width of 2. In the top histogram, the data appears to be skewed slightly to the left. In the bottom two histograms, the data appears to be skewed slightly to the right.

Despite the problem caused by the choice of bin width, histograms help to see how the sample values are distributed along the range of observed values. One of the salient features of all three histograms in Figure 4 is that there is a hump in the middle of the range. In terms of the original variables, there appears to be a range of speeds that is more common, and for this set of data, that popular range of speeds happens to be near the middle of the range. Using the formal language

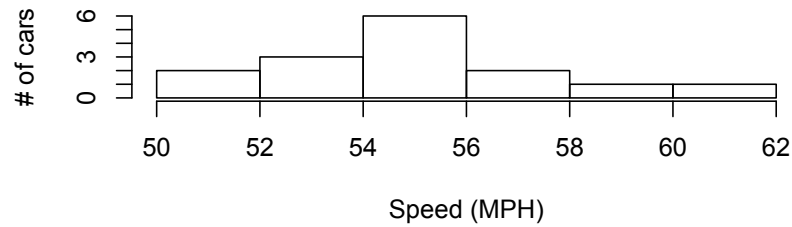


Figure 2: Histogram plot of the speed measurements.

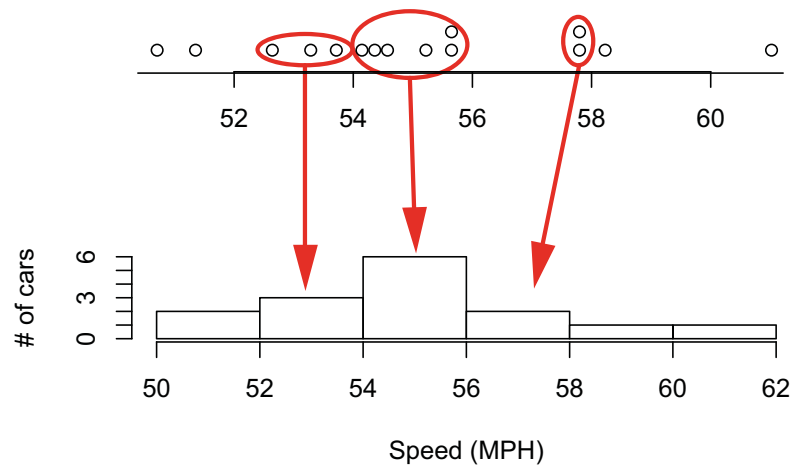


Figure 3: Graphical representation of data binning used to create the histogram in Figure 2.

of statistics, we say that the data has a central tendency (which is not always the case). Using everyday language, we say there appears to be an average speed near the middle of the range.

Review

We can summarize the concepts discussed so far with the following points

- A population is the entire set of values we wish to study
- A sample is a subset of the population
- Dot plots and histograms graphically represent the distribution of values in a sample

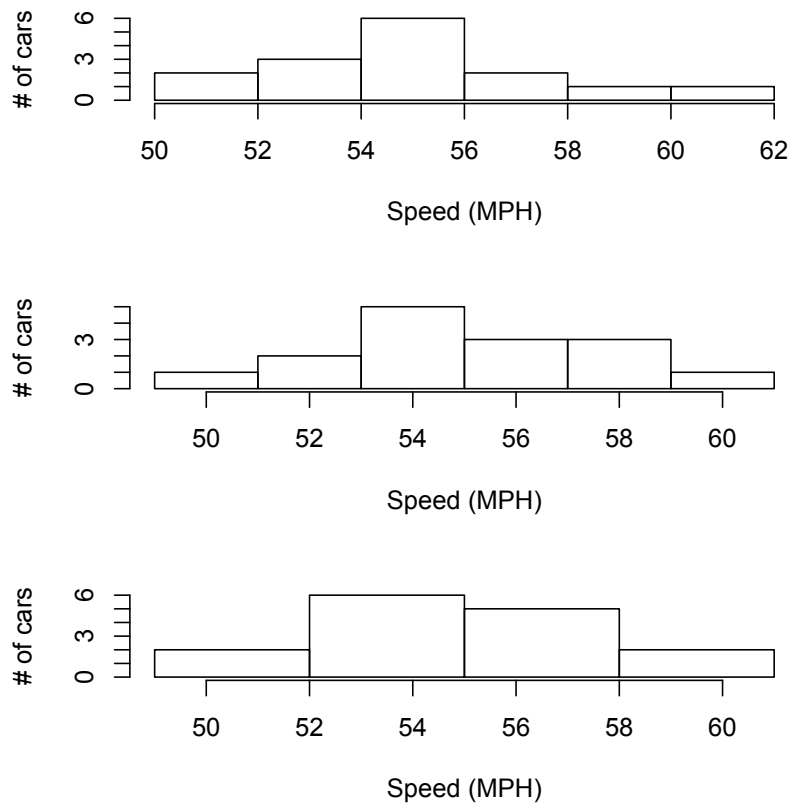


Figure 4: Histogram of car speeds with bin sizes of 2, 2, and 3. The top two histograms differ in the starting point of the first bin: 50 versus 49.

- A dot plot shows the distribution of observations of the random variable. The horizontal axis is the value of the random variable being observed. When multiple values are observed at or near the same value, the dots are stacked.
- Histograms are bar plots obtained when a data set is grouped into bins. The height of each bar is the number of observations in that bin. The bin is defined as a range of the random variable being observed. Histograms can be misleading, especially when the bins are wide.

Quantitative Statistics

Dot plots and histograms help to show how values in a data set are distributed. The shape of the distribution gives useful qualitative information. For routine analysis, it is usually necessary to have quantitative information about the distribution.

Descriptive Statistics

The term *descriptive statistics* refers to the quantitative measures of a population or sample that describe the data as it is. Let's say you have a set of data that is a list of numbers. Often that data is from measurements. From the data you compute values, called *statistics*, that depend solely on the data. The two most basic descriptive statistics are the *mean* (commonly called the *average*) and the *standard deviation*. We will also introduce another important descriptive statistic called the *median*.

Inferential and Predictive Statistics

Statistical analysis is useful for more than just describing data. *Inferential statistics* enables quantitative comparing two or more sets of observations. For example, suppose we wanted to test the strength of different glues, Glue A and Glue B, for joining pieces of wood. We decide to make several measurements for each brand of glue because we know that other variables could affect the strength of the joint, e.g. porosity of the wood and amount of glue applied, and that the testing process itself will introduce some random variation. The table at the right shows the results of 10 measurements of the force necessary to break the glue bond. The last row of the table shows the average of the 10 measurements. As a practical matter, we must decide whether the difference in average breaking force is large enough to make a difference in our application. For example, if the required strength is 50 newtons, then either glue will be fine. The role of inferential statistics is to give us a way of comparing the difference in glue strengths with respect to the variation in the data.

Bond-breaking force (newtons)	
Glue A	Glue B
75.03	73.38
74.91	73.85
74.94	74.87
73.87	74.61
75.61	72.21
74.40	72.39
74.81	74.89
75.22	73.83
74.35	74.85
75.94	72.17
74.9	73.6

Predictive statistics involves developing models of observed behavior that accounts for the variability of the data. Extending the glue example from the preceding paragraph, suppose that the glue strength experiment was repeated several times for a range of temperatures at which the glue was allowed to cure. We could make a curve fit of glue strength as a function of temperature for each glue. A complete statistical model would include uncertainty estimates for the coefficients in the curve fit, as well as some estimate of whether any differences in strength at a given temperature were statistically significant given the variability of the data.

Center of the Data: Mean and Median

Compute the mean (or average)

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

where n is the total number of samples in the data set.

Example: The mean of the speed data on page 1 is

$$\begin{aligned} \bar{x} &= \frac{1}{15} (50.6 + 53.7 + 54.5 + 55.6 + 51.4 + 55.3 + 61.0 \\ &\quad + 52.6 + 54.1 + 54.4 + 58.2 + 55.6 + 53.3 + 57.7 + 57.8) \\ &= 55.05 \end{aligned}$$

The median is not computed with a single formula. Instead the median is found with the following algorithm

1. Sort the data (ascending or descending order)
2. If there is an odd number of values in the data set, the median is in the middle of the sorted list, i.e. at index $(n + 1)/2$
3. If there is an even number of values in the data set, the median is the average of the two values in the middle of the list.

Finding the median by hand is tedious for all but the smallest data sets. Computer software makes it easy to find the median once the data is entered.

Example: Find the median of the speed data from page 1. First, sort the data

50.6, 51.4, 52.6, 53.3, 53.7, 54.1, 54.4, 54.5, 55.3, 55.6, 55.6, 57.7, 57.8, 58.2, 61.0

Since there is an odd number of values, the median is the eight value in the list, or 54.5.

To-do: Give examples of non-symmetric distributions where the mean and median are not equal.

Spread of the Data: Range, IQR, and Standard Deviation

Range: Difference between the maximum and minimum values

Interquartile Range: Difference between the upper and lower quartiles.

- Sort the data
- Divide the values into four groups having the same number of values in the the order in which they appear in the sorted list. The values delineating the boundaries of these groups are the quartiles, usually designated Q_1 , Q_2 , Q_3 and Q_4
- The *interquartile range* or IQR is the difference between Q_3 and Q_1 .

$$\text{IQR} = Q_3 - Q_1$$

Standard Deviation: The standard deviation is a measure of the spread of a data set. The formula for computing the standard deviation, σ , is

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

where \bar{x} is the mean of the sample, and n is the number of values in the sample.

Further Reading

See Barbara Illowsky and Susan Dean, Collaborative Statistics, on-line at <http://cnx.org/content/co110522/1.35/>

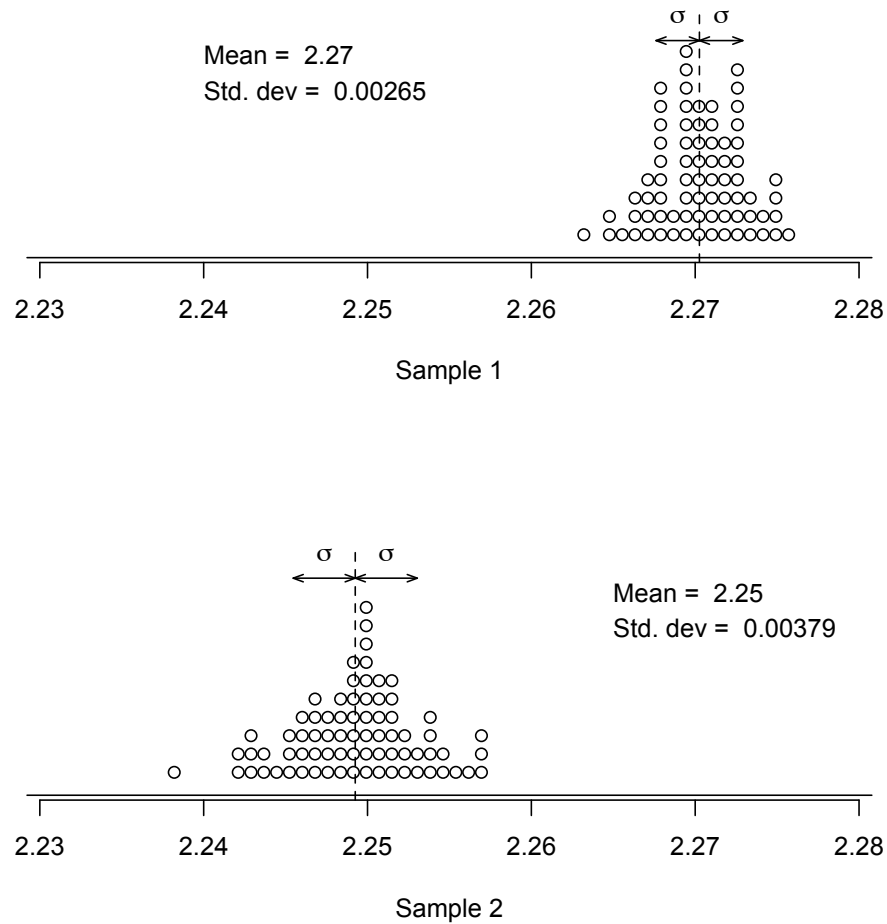


Figure 5: Two samples of measurements from a hypothetical manufacturing process to cut a board to a prescribed length. Sample 1 (top) was taken during a production run when the error in length was discovered. Sample 2 (top) was taken during adjustment of machinery that controls the board length.