1

DESCRIPTIVE STATISTICS

1.1 MEASURES OF CENTRAL TENDENCY

One wishes to establish some basic understanding of statistical terms before we deal in detail with the laboratory applications. We want to be sure to understand the meaning of these concepts, since one often describes the data with which we are dealing in summary statistics. We discuss what is commonly known as measures of central tendency such as the mean, median, and mode plus other descriptive measures from data. We also want to understand the difference between samples and populations.

Data come from the samples we take from a population. To be specific, a population is a collection of data whose properties are analyzed. The population is the *complete* collection to be studied; it contains all possible data points of interest. A sample is a part of the population of interest, a subcollection selected from a population. For example, if one wanted to determine the preference of voters in the United States for a political candidate, then all registered voters in the United States would be the population. One would sample a subset, say, 5000, from that population and then determine from the sample the preference for that candidate, perhaps noting the percent of the sample that prefer that candidate over another. It would be impossible logistically and costwise in statistics to canvass the entire population, so we take what we believe to be a representative sample from the population. If the sampling is done appropriately, then we can generalize our results to the whole population. Thus, in statistics, we deal with the sample that we collect and make our decisions. Again, if

Introduction to Statistical Analysis of Laboratory Data, First Edition.
Alfred A. Bartolucci, Karan P. Singh, and Sejong Bae.

we want to test a certain vegetable or fruit for food allergens or contaminants, we take a batch from the whole collection, send it to the laboratory and it is, thus, subjected to chemical testing for the presence or degree of the allergen or contaminants. There are certain safeguards taken when one samples. For example, we want the sample to appropriately represent the whole population. Factors relevant in considering the representativeness of a sample include the homogeneity of the food and the relative sizes of the samples to be taken, among other considerations. Therefore, keep in mind that when we do statistics, we always deal with the sample in the expectation that what we conclude generalizes to the whole population.

Now let's talk about what we mean when we say we have a distribution of the data. The following is a sample of size 16 of white blood cell (WBC) counts $\times 1000$ from a diseased sample of laboratory animals:

$$5.13, 5.4, 5.4, 5.7, 5.7, 5.7, 6.0, \underline{6.0, 6.0}, 6.0, 6.13, 6.13, 6.13, 6.4, 6.4, 6.8.$$

Note that this data is purposely presented in ascending order. That may not necessarily be the order in which the data was collected. However, in order to get an idea of the range of the observations and have it presented in some meaningful way, it is presented as such. When we rank the data from the smallest to the largest, we call this a distribution.

One can see the distribution of the WBC counts by examining Figure 1.1. We'll use this figure as well as the data points presented to demonstrate some of the statistics that will be commonplace throughout the text. The height of the bars represents the frequency of counts for each of the values 5.13–6.8, and the actual counts are placed on top of the bars. Let us note some properties of this distribution. The mean is easy. It is obviously the average of the counts from 5.13 to 6.8 or $(5.13 + 5.4 + \cdots + 6.8)/16 = 5.939$. Algebraically, if we denote the elements of a sample of size

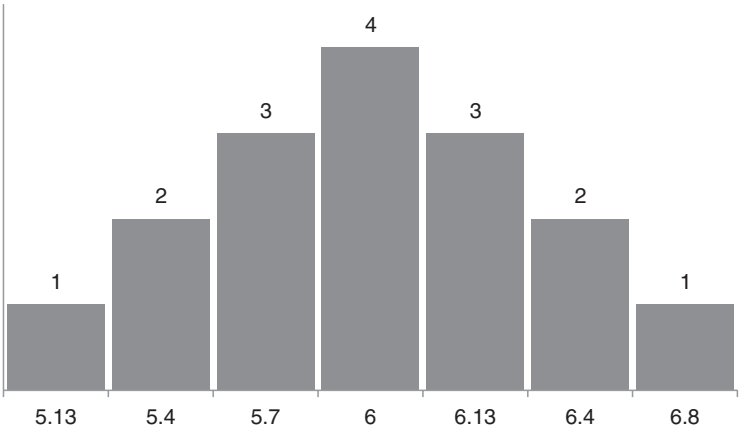

Figure 1.1 Frequency Distribution of White Cell Counts

n as $X_1, X_2, \ldots, X_n$, then the sample mean in statistical notation is equal to

$$\overline{X} = \frac{(X_1 + X_2 + \cdots + X_n)}{n}. \tag{1.1}$$

For example, in our aforementioned WBC data, $n = 16$, $X_1 = 5.13, X_2 = 5.4$, and so on, where $X_{16} = 6.8$.

Then the mean is noted as earlier, $(5.13 + 5.4 + \cdots + 6.8)/16 = 5.939$.

The median is the middle data point of the distribution when there is an odd number of values and the average of the two middle values when there is an even number of values in the distribution. We demonstrate it as follows.

Note our data is:

$$5.13, 5.4, 5.4, 5.7, 5.7, 5.7, 6.0, 6.0, \underline{6.0, 6.0}, 6.13, 6.13, 6.13, 6.4, 6.4, 6.8.$$

The number of data points is an even number, or 16. Thus, the two middle values are in positions 8 and 9 underlined above. So the median is the average of 6.0 and 6.0 or $(6.0 + 6.0)/2 = 12.0/2 = 6.0$ or median = 6.0.

Suppose we had a distribution of seven data points, which is an odd number, then the median is just the middle value or the value in position number 4. Note the following: 5.13, 5.4, 5.6, 5.7, 5.8, 5.8, 6.0. Thus, the median value is 5.7. The median is also referred to as the 50th percentile. Approximately 50% of the values are above it and 50% of the values are below it. It is truly the middle value of the distribution.

The mode is the most frequently occurring value in the distribution. If we examine our full data set of 16 points, one will note that the value 6.0 occurs four times. Also see Figure 1.1. Thus, the mode is 6.0. One can have a distribution with more than one mode. For example, if the values of 5.4 and 6.0 were each counted four times, then this would be a bimodal distribution or a distribution with two modes.

We have just discussed what is referred to as measures of central tendency. It is easy to see that the measures of central tendency from this data (mean, median, and mode) are all in the center of the distribution, and all other values are centered around them. In cases where the mean = median = mode as in our example, the distribution is seen to be symmetric. Such is not always the case.

Figure 1.2 deals with data that is skewed and not symmetric. Note the mode to the left indicating a high frequency of low values. These are potassium values from a laboratory sample. This data is said to be skewed to the right or positively skewed. We'll revisit this concept of skewness in Chapter 2 and later chapters as well. There are 23 values (not listed here) ranging from 30 to 250. One usually computes the geometric mean (GM) of the data of this form. Sometimes, GM is preferred to the arithmetic mean (ARM) since it is less sensitive to outliers or extreme values. Sometimes, it is called a "spread preserving" statistic. The GM is always less than or equal to the ARM and is commonly used with data that may be skewed and not normal or not symmetric, such as much laboratory data is not symmetric.

Suppose we have n observations $X_1, X_2, \ldots, X_n$, then the GM is defined as

$$\text{GM} = \prod_{i=1,n} X_i^{1/n} = {X_1}^{1/n} {X_2}^{1/n} {X_n}^{1/n}, \tag{1.2}$$

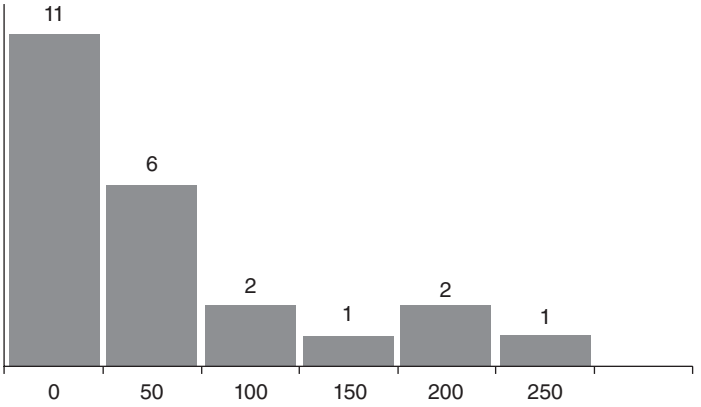

Figure 1.2 Frequency Distribution of Potassium Values

or equivalently

$$\text{GM} = \exp\left[\frac{\left(\log X_1 + \log X_2 + \cdots + \log X_n\right)}{n}\right]. \tag{1.3}$$

In our potassium example GM $= 30^{1/23}35^{1/23}, \ldots, 250^{1/23} = 65.177$. Note that the ARM $= 75.217$.

1.2 MEASURES OF VARIATION

We've learned some important measures of statistics. The mean, median, and mode describe some sample characteristics. However, they don't tell the whole story. We want to know more characteristics of the data with which we are dealing. One such measure is the dispersion or the variance. This particular measure has several forms in laboratory science and is essential to determining something about the precision of an experiment. We will discuss several forms of variance and relate them to data accordingly.

The range is the difference between the maximum and minimum value of the distribution. Referring to the WBC data:

$$\text{Range} = \text{maximum value–minimum value} = 6.8 - 5.13 = 1.67.$$

Obviously, the range is easy to compute, but it only depends on the two most extreme values of the data. We want a value or measure of dispersion that utilizes all of the observations. Note the data in Table 1.1. For the sake of demonstration, we have three observations: 2, 4, and 9. These data are seen in the data column.

TABLE 1.1 Demonstration of Variance

Observation	Data	Deviation	(Deviation)^2
1	2	$(2-5)=-3$	$(2-5)^2=9$
2	4	$(4-5)=-1$	$(4-5)^2=1$
3	9	$(9-5)=4$	$(9-5)^2=16$
Sum	15	0	26
Average	5	0	$26/(3-1)=13$

Note their sum or total is 15. Their mean or average is 5. Note their deviation from the mean, $2-5=-3$, $4-5=-1$ and $9-5=4$. The sum of their deviations is 0. This property is true for any size data set, that is, the sum of the deviations will be close to 0. This doesn't make much sense as a measure of dispersion or we would have a perfect world of no variation or dispersion of the data. The last column denoted as (Deviation)^2 is the deviation column squared. And the sum of the squared deviations is 26.

The variance of a sample is the average squared deviation from the sample mean. Specifically, from the previous sample of three values, $[(2-5)^2 + (4-5)^2 + (9-5)^2]/(3-1) = [9 + 1 + 16]/2 = 26/2 = 13$. Thus, the variance is 13. Dividing by $(3-1)=2$ instead of 3 gives us an *unbiased* estimator of the variance because it tends to closely estimate the true population variance. Note that if our sample size were 100, then dividing by 99 or 100 would not make much of a difference in the value of the variance. The adjustment of dividing the sum of squares of the deviation by the sample size minus 1, $(n-1)$, can be thought of as a small sample size adjustment. It allows us not to underestimate the variance but to conservatively overestimate it.

Recall our WBC data:

$$5.13, 5.4, 5.4, 5.7, 5.7, 5.7, 6.0, 6.0, 6.0, 6.0, 6.13, 6.13, 6.13, 6.4, 6.4, 6.8.$$

The mean or average is: $5.939 = 5.94$.

So the variance is

$$\text{Var} = \frac{[(5.13-5.94)^2 + (5.4-5.94)^2 + \cdots + (6.8-5.94)^2]}{15} = 0.1798$$

Algebraically, one may note the variance formula in statistical notation for the data in Table 1.1, where the mean is $\overline{X} = 5$.

One defines the sample variance as S^2_{n-1} or

$$S^2_{n-1} = \frac{\sum (X_i - \overline{X})^2}{n-1} \tag{1.4}$$

So for the data in Table 1.1 we have

$$S^2_{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2}{3-1}$$

$$= \frac{(2-5)^3 + (4-5)^2 + (9-5)^2}{2} = \frac{(-3)^2 + (-1)^2 + (4)^2}{2}$$

$$= \frac{9+1+16}{2} = \frac{26}{2} = 13$$

The sample standard deviation (SD), S_{n-1}, is the square root of sample variance $= \sqrt{S^2_{n-1}}$, or in our case $\sqrt{13} = 3.606$.

$$\sqrt{S^2_{m-1}} = \sqrt{13} = 3.606 \tag{1.5}$$

The variance is a measure of variation. The square root of the variance, or SD, is a measure of variation in terms of the original scale.

Thus, referring back to the aforementioned WBC data, the SD of our WBC counts is the square root of the variance, that is, $\sqrt{0.1798} = 0.4241$.

Just as we discussed the GM earlier for data that may be possibly skewed, we also have a geometric standard deviation (GSD). One uses the log of the data as we did for the GM. The GSD is defined as

$$\text{GSD} = \exp\left[\sqrt{\frac{\sum_{i=1,n}\left\{\log X_i - \log \text{GM}\right\}^2}{n-1}}\right]. \tag{1.6}$$

As an example, suppose we have $n = 10$ data points 100, 99, 100, 90, 90, 70, 89, 70, 64, 56.

Then from (1.6), the GSD = 1.233. Unlike the GM, the GSD is not necessarily a close neighbor of the arithmetic SD, which in this case is 16.315.

Another measure of variation is the standard error of the mean (SE or SEM), which is the SD divided by the square root of the sample size or

$$\text{SE} = \frac{\text{SD}}{\sqrt{n}}. \tag{1.7}$$

For our aforementioned WBC data, we have $\text{SE} = 0.4241/\sqrt{16} = 0.4241/4 = 0.1060$.

The standard error (SE) of the mean is the variation one would expect in the sample means after repeated sampling from the same population. It is the SD of the sample

means. Thus, the sample SD deals with the variability of your data while the SE of the mean deals with the variability of your sample mean.

Naturally, we have only one sample and one sample mean. Theoretically, the SE is the SD of many sample means after sampling repeatedly from the same population. It can be thought of as a SD of the sample means from replicated sampling or experimentation. Thus, a good approximation of the SE of the mean from one sample is the SD divided by the square root of the sample size as seen earlier. It is naturally smaller than the SD. This is because from repeated sampling from the population one would not expect the mean to vary much, certainly not as much as the sample data. Rosner (2010, Chapter 6, Estimation) and Daniel (2008, Chapter 6, Estimation) give an excellent demonstration and explanation of the SD and SE of the mean comparisons.

Another common measure of variation used in laboratory data exploration is the coefficient of variation (CV), sometimes referred to as the relative standard deviation (RSD). This is defined as the ratio of the SD to the mean expressed as a percent.

It is also called a measure of reliability – sometimes referred to as precision and is defined as

$$\text{CV} = \left(\frac{\text{SD}}{\text{mean}}\right) \times 100. \tag{1.8}$$

Our Sample CV of the WBC measurements is $\text{CV} = \left(\frac{0.4241}{5.94}\right) \times 100 = 7.14$.

The multiplication by 100 allows it to be referred to as the percent CV, %CV, or CV%.

The %CV normalizes the variability of the data set by calculating the SD as a percent of the mean. The %CV or CV helps one to compare the precision differences that may exist among assays and assay methods. We'll see an example of this in the following section. Clearly, an assay with CV = 7.1% is more precise than one with CV = 10.3%.

1.3 LABORATORY EXAMPLE

The following example is based on the article by Steele et al. (2005) from the *Archives of Pathology and Laboratory Medicine*. The objective of the study was to determine the long-term within- and between-laboratory variation of cortisol, ferritin, thyroxine, free thyroxine, and Thyroid-Stimulating Hormone (TSH) measurements by using commonly available methods and to determine if these variations are within accepted medical standards, that is to say within the specified CV.

The design – Two vials of pooled frozen serum were mailed 6 months apart to laboratories participating in two separate College of American Pathologists' surveys. The data from those laboratories that analyzed an analyte in both surveys were used to determine for each method the total variance and the within- and between-laboratory variance components. For our purposes, we focus on the CV for one of the analytes, namely, the TSH. There were more than 10 analytic methods studied in this survey. The three methods we report here are as follows: A – Abbott AxSYM, B – Bayer

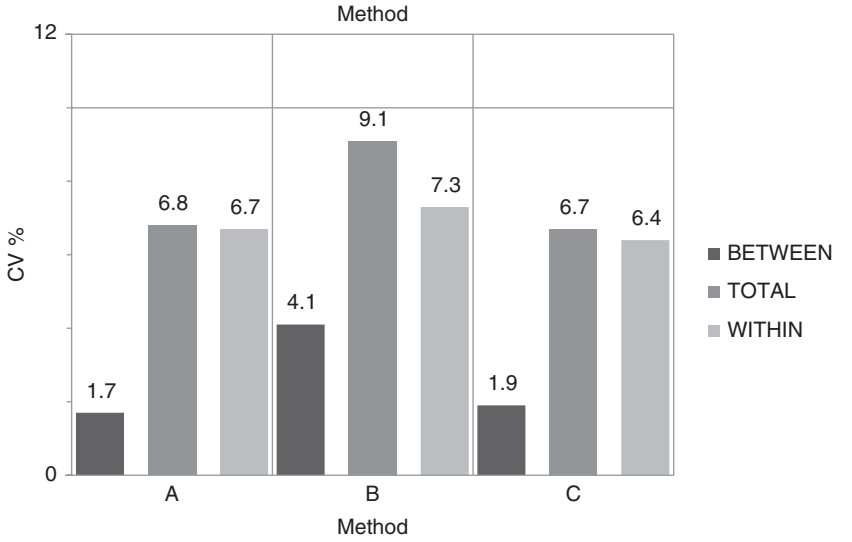

Figure 1.3 CV% for TSH. Reproduced in part from Steele et al. (2005) with permission from Archives of Pathology and Laboratory Medicine. Copyright 2005 College of American Pathologists

Advia Centaur, and C – Bayer Advia Centaur 3G. The study examined many endpoints directed to measuring laboratory precision with a focus on total precision overall and within- and between-laboratory precision. The within-laboratory goals as per the %CV based on biological criteria were cortisol – 10.43%, ferritin – 6.40%, thyroxine – 3.00%, free thyroxine – 3.80%, and TSH – 10.00%. Figure 1.3 shows the graph for analytic methods A, B, and C, for TSH. The horizontal reference line across the top of the figure at 10% indicates that all of the bars for the total, within- and between-laboratory %CV met the criteria for the three methods shown here. Also, note in examining Figure 1.3 that the major source of variation was within-laboratory as opposed to the between- or among-laboratory variation or %CV.

When examining the full article, the authors point out that the number of methods that met within-laboratory imprecision goals based on biological criteria were 5 of 5 for cortisol; 5 of 7 for ferritin; 0 of 7 for thyroxine and free thyroxine; and 8 of 8 for TSH. Their overall conclusion was that for all analytes tested, the total within-laboratory component of variance was the major source of variation. In addition, note that there are several methods, such as thyroxine and free thyroxine that may not meet analytic goals in terms of their imprecision.

1.4 PUTTING IT ALL TOGETHER

Let's consider a small data set of potassium values and demonstrate summary statistics in one display. Table 1.2 gives the potassium values denoted by the X_i, where

TABLE 1.2 Potassium Values and Descriptive Statistics

i	X_i	$Y_i = \ln(X_i)$
1	3.2	1.163
2	4.2	1.435
3	3.8	1.335
4	2.3	0.833
5	1.5	0.405
6	15.5	2.741
7	8.5	2.140
8	7.9	2.067
9	3.1	1.131
10	4.4	1.482
Mean	$\overline{X} = 5.44$	$\overline{Y} = 1.47$
Variance	$S^2_X = 17.53$	$S^2_Y = 0.46$
Standard deviation	$\mathrm{SD}_X = 4.19$	$\mathrm{SD}_Y = 0.68$
Standard error of the mean	$\mathrm{SE}_X = 1.32$	$\mathrm{SE}_Y = 0.22$

$i = 1, 2, \ldots, 10$. The natural log of the values are seen in the third column denoted by $Y_i = \ln(X_i)$. The normal range of values for adult laboratory potassium (K) levels are from 3.5–5.2 milliequivalents per liter (mEq/L) or 3.5–5.2 millimoles per liter (mmol/L). Obviously, a number of the values are outside the range. The summary statistics are provided for both raw and transformed values, respectively. The Y values are actually from what we call a log-normal distribution, which we will discuss in the following chapter. Focusing on the untransformed potassium values of Table 1.2, Table 1.3 gives a complete set of summary statistics that one often encounters. We've discussed most of them and will explain the others. The minimum and maximum values are obvious, being the minimum and maximum potassium values from Table 1.2. The other two added values in Table 1.3 are 25th percentile (first quartile) and 75th percentile (third quartile). They are percentiles just like the median. Just as the median is the 50th percentile (second quartile) in which approximately 50% of the values may lie above it as well as below it, the 25th percentile is the value of 2.9, meaning that approximately 25% of the values in the distribution lie below it, which implies about 75% of the values in the distribution lie above the value 2.9. Thus, the 75th percentile is the value of 8.05, meaning that 75% of the values in the distribution are less than or equal to 8.05, implying that about 25% of the values lie above it. Note that the median is in the middle of the 25th and 75th percentile. These values between the 25th and 75th quartile are called the interquartile range (IQR). Note that approximately 50% of the data points are in the IQR.

Let's revisit the GM and GSD. From Table 1.2, we note that

$$\mathrm{GM} = \exp\left[\frac{\left(\log X_1 + \log X_2 + \cdots + \log X_n\right)}{n}\right] = \exp[\overline{Y}] = \exp[1.47] = 4.349.$$

TABLE 1.3 Descriptive Statistics of 10 Potassium (X) Values

Mean	5.44
Standard deviation	4.19
Standard error mean	1.32
N	10
Minimum	1.5
25th percentile	2.9
Median	4.0
75th percentile	8.05
Maximum	15.5

Also, the relation between the arithmetic standard and GSD is such that ln (GSD) = arithmetic SD of the Y_i;s in Table 1.2. Thus, ln(GSD) = 0.68 or GSD = exp(0.68) = 1.974.

1.5 SUMMARY

We have briefly summarized a number of basic descriptive statistics in this chapter such as the measures of central tendency and measures of variation. We also put them in the context of data that has a symmetric distribution as well as data that is not symmetrically distributed or may be skewed. It is important to note that these statistics just describe some property of the sample with which we are dealing in laboratory experimentation. Our goal in the use of these statistics is to describe what is expected to be true in the population from which the sample was drawn. In the next chapter, we discuss inferential statistics, which leads us to draw scientific conclusions from the data.

REFERENCES

Daniel WM. (2008). Biostatistics: A Foundation for Analysis in the Health Sciences, 9th ed., John Wiley & Sons, New York.

Rosner B. (2010). Fundamentals of Biostatistics, 7th ed., Cengage Learning.

Steele BW, Wang E, Palmer-Toy DE, Killeen AA, Elin RJ and Klee GG. (2005). Total long-term within-laboratory precision of cortisol, ferritin, thyroxine, free thyroxine, and Thyroid-Stimulating Hormone (TSH) assays based on a College of American Pathologists fresh frozen serum study: do available methods meet medical needs for precision? Archives of Pathology and Laboratory Medicine 129(3): 318–322.