

CHAPTER THREE

COMMON DESCRIPTIVE STATISTICS

The analysis of data begins with **descriptive statistics** such as the **mean, median, mode, range, standard deviation, variance, standard error of the mean, and confidence intervals**. These statistics are used to summarize data and provide information about the sample from which the data were drawn and the accuracy with which the sample represents the population of interest. The mean, median, and mode are measurements of the “**central tendency**” of the data. The range, standard deviation, variance, standard error of the mean, and confidence intervals provide information about the “**dispersion**” or **variability** of the data about the measurements of central tendency.

MEASUREMENTS OF CENTRAL TENDENCY

The appropriateness of using the mean, median, or mode in data analysis is dependent upon the nature of the data set and its distribution (normal vs non-normal). The **mean** (denoted by \bar{x}) is calculated by dividing the sum of the individual data points (where Σ equals “sum of”) by the number of observations (denoted by n). It is the arithmetic average of the observations and is used to describe the center of a data set.

$$\text{mean} = \bar{x} = \frac{\Sigma x}{n}$$

The mean is commonly used to describe numerical data that is normally distributed. It is very sensitive to extreme values in the data set. For example, the mean of the data set {1,2,3,4,5} is 15/5 or 3. If the number 20 is substituted for the 5, the data set becomes {1,2,3,4,20} and the mean is 30/5 or 6. Whereas a mean of 3 accurately describes the “center” of the first data set, a mean of 6 does not accurately describe the distribution of the second data set. Thus, the mean is subject to extreme values or “outliers” and may not accurately represent the true center of the data if such outlying values are present. The mean is only appropriate if the data are normally distributed as is the case in the first data set.

The **median** is another method of describing the center of a data set. It is the middle value of the data if the number of observations, n , is odd, or the average of the two middle values if n is even. By definition, half of the data points reside above the median and half reside below the median. For example, the median for each of the above data sets is 3 despite the outlying value of 20 in the second data set. The median is therefore useful for describing the center of a data set that is non-normally distributed as is the case in the second data set. It is also commonly used with ordinal data that is non-numerical.

The **mode** is the value which occurs most frequently in the data. There may be one or more modes for each data set and this makes the mode a useful method for describing a population that is bimodal. In the data set {2,3,4,4,5,6,7,7,8,9}, for example, there are two modes, 4 and 7. Whereas a data set may have one or more modes, there can only be one mean and one median.

The mean, median, and mode can be used to evaluate the symmetry of a data distribution. This is essential to choosing the appropriate statistical test. If the mean and median are equal, the data is usually symmetrical or normally distributed and a test for normally distributed data should be used. If the mean and median differ markedly, the data are likely skewed and a test for non-normally distributed data is appropriate.

MEASUREMENTS OF VARIABILITY

As we have seen, the mean, median, and mode are used to describe the central tendency of the data. Used alone, however, they do not adequately describe a data set. We also need a way of accurately describing the “dispersion” or variability of the data about the measurements of central tendency. Remember that it is this variability that mandates the use of statistical analysis. One method of describing this variability is the **range** which is defined as the difference between the smallest and largest values in the data set. It is frequently given as the minimum and maximum values and is used to demonstrate the presence of extreme values or “outliers” which would tend to skew the mean and median in one direction or another.

Perhaps the most commonly used method to describe the variability of a data set is the **standard deviation**. It is the cornerstone behind most of the commonly used statistical tests. The standard deviation is an estimate of the average distance of the values from their mean. Assuming the data is normally distributed (i.e., assumes a bell-shaped curve with half of the data lying on either side of the mean), approximately 68% of the data will lie within 1 standard deviation, approximately 95% within 2 standard deviations, and approximately 99% within 3 standard deviations.

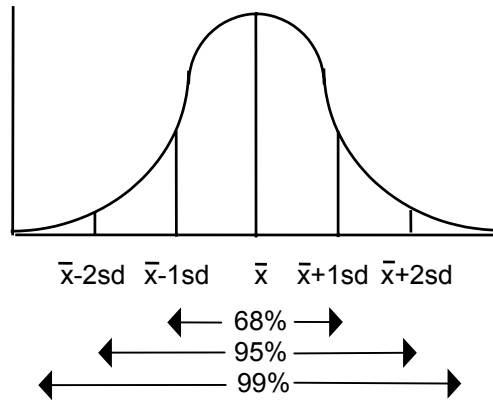


Figure 3-1: The Normal Distribution

The standard deviation is easily calculated by virtually any computer, but the equation is included below as we will use it repeatedly in describing the commonly used statistical tests for data analysis. The **variance** is the square of the standard deviation.

$$\text{standard deviation} = \text{sd} = \sqrt{\frac{\sum(x - \bar{x})^2}{(n - 1)}}$$

where x = each data point, \bar{x} = mean,
and n = number of observations

Assume we wish to calculate the mean and standard deviation for a series of serum sodium measurements to determine their central tendency and variability. Such calculations are the basis for the “normal laboratory ranges” which we use everyday. The normal range for most laboratory tests is defined as the mean \pm 2 standard deviations, which encompasses the test results for 95% of the population. Although easily calculated with a computer, creation of a table such as the one below illustrates the steps involved in calculating the standard deviation.

Serum Na (x)	(x- \bar{x})	(x- \bar{x}) ²
138	-2	4
137	-3	9
143	3	9
141	1	1
138	-2	4
142	2	4
136	-4	16
140	0	0
145	5	25
143	3	9
mean = \bar{x} = 140		sum = 81

$$\text{standard deviation} = \text{sd} = \sqrt{\frac{81}{9}} = \sqrt{9} = 3$$

In describing this data, we would state that the mean is 140 mEq/L with a standard deviation of 3 and variance of 9, and that 95% of the values are within 6 mEq of the mean (2 standard deviations). This is similar to the normal clinical range for serum sodium (135-145 mEq/L).

The mean, standard deviation, and variance are inappropriate if the data contain significant outliers or are skewed (i.e., non-normally distributed). In such a situation, the median and range may more accurately describe the central tendency and dispersion of the data set. Charts and other graphics can similarly be used to describe data where the variability is notable.

STANDARD ERROR OF THE MEAN

The **standard error of the mean** (se) is frequently confused with the standard deviation. Whereas the standard deviation describes the variability of the data about the mean, the standard error of the mean estimates how closely the sample mean approximates the population mean. It is essentially the standard deviation of all possible sample means if we were to repeatedly draw samples from our population and calculate the mean for a particular variable of interest in each sample. It is calculated by dividing the standard deviation by the square root of n , the number of observations, and will always be smaller than the standard deviation.

$$sd = se \times \sqrt{n}$$

As the sample size (n) increases and approaches the population size, the standard error of the mean approaches zero as the sample mean will more closely approximate the true population mean. The standard deviation, however, remains fairly constant with changes in sample size as the sample standard deviation is an estimate of the population standard deviation whereas the standard error of the mean is not.

The standard error of the mean is useful in comparing the means of two samples to determine whether they are representative of the same population. It does not describe the variability of the data, as does the standard deviation, and should not be used in place of the standard deviation. The mean and standard deviation allow the clinician to apply the data to individual patients. The standard error of the mean describes only the sample as a group and cannot be applied to patients clinically.

The following example illustrates the use of descriptive statistics in the reporting of data. The Preoperative Evaluation study compared surgical patients who had pulmonary artery catheters placed preoperatively in an intensive care unit ("study patients") with those who had catheters placed in the operating room prior to their operation ("control patients"). The data presented are the mean economic data on intensive care unit (ICU) and total hospital length of stay (LOS) and ICU and total hospital charges.

Preoperative Pulmonary Artery Catheterization

	<u>ICU LOS</u>	<u>Hospital LOS</u>	<u>ICU Charges</u>	<u>Hospital Charges</u>
Study	8.7 days	33.2 days	\$38,557	\$78,887
Control	5.7 days	21.4 days	\$21,140	\$56,517

Study - patients with pulmonary artery catheters placed preoperatively in the ICU

Control - patients with pulmonary artery catheters placed in the operating room

Data presented as the mean of each group

From this data, we might conclude that patients who had pulmonary artery catheters placed preoperatively in the ICU had increased ICU and total hospital LOS and increased ICU and total hospital charges compared to patients whose pulmonary artery catheter was placed in the operating room. Such conclusions would be correct if the data are normally distributed, but without some measure of variability we do not know that to be true. If we review the raw data, in fact, we find that the "study" data is heavily skewed by a single patient (whom we will call patient "E") who had an ICU LOS of 179 days, a hospital LOS of 254 days, and total hospital charges of \$838,282. Since the data are skewed and therefore non-normally distributed, we cannot use the mean to accurately describe the central tendency of the data and any conclusions we might make using the currently available data would be invalid. Consider the effect on the data when patient E is excluded from the data analysis.

Without Patient E:

	<u>ICU LOS</u>	<u>Hospital LOS</u>	<u>ICU Charges</u>	<u>Hospital Charges</u>
Study	4.6 days	20.3 days	\$15,031	\$49,289
Control	5.7 days	21.4 days	\$21,140	\$56,517

After excluding patient E from the analysis, our conclusions would be very different. We would now conclude that preoperative evaluation decreases ICU and total hospital LOS as well as ICU and total hospital charges. From a statistical standpoint, however, patient E cannot be excluded from the data analysis as this introduces **selection bias** into the data and clearly alters the results of the study. A better method for presenting this non-normally distributed data would be to report not only the mean, but also the median and range of each group to demonstrate the presence and impact of outlying data such as that of patient E.

Preoperative Pulmonary Artery Catheterization				
Study	ICU LOS	Hospital LOS	ICU Charges	Hospital Charges
mean	8.7 days	33.2 days	\$38,557	\$78,887
median	3 days	15 days	\$10,496	\$39,758
range	1-178 days	2-254 days	\$1,123-\$304,152	\$7,144-\$838,282
Control				
mean	5.7 days	21.4 days	\$21,140	\$56,517
median	3 days	14 days	\$9,836	\$40,314
range	1-43 days	4-128 days	\$1,714-\$188,245	\$9,644-\$289,045

When appropriate statistics for non-normally distributed data are used, it becomes clear that the two groups are equivalent and that our original conclusions, which were dependent upon the data being normally distributed, were in error. Presenting the data in this way makes clear the presence and effect of outliers to the reader. It also emphasizes the importance of reviewing the raw data to establish the presence of significant trends or errors in data entry that might be missed if the data were analyzed using only the mean. This can be done most effectively by tabulating the occurrences of each value (a **frequency distribution**) or graphing the data (a **histogram** or **scatter plot**) such that outliers, skewed distributions, and data entry errors become readily apparent. Had we not identified the effect of patient E on the data, we might easily have concluded that a significant difference existed between the two groups (committing a Type I error) and have altered our clinical practices based on these erroneous conclusions.

CONFIDENCE INTERVALS

Probability limits establish a range and specify the probability that an observed value occurs within that range. They can be used to determine whether a particular value is likely to have come from a specified population. For example, the 95% probability interval (assuming a normally distributed data set) is approximately given by the observed value ± 2 standard deviations. **Confidence intervals** are similar to probability limits. They establish a range and specify the probability of the true population mean being within that range. They are most commonly reported as the **95% confidence interval** which is defined as the mean ± 2 standard errors of the mean.

$$95\% \text{ confidence interval} = \bar{x} \pm 2 \text{ se}$$

If we were to look at repeated samples from our population of interest and calculate the 95% confidence interval for each, 95% of the confidence intervals so obtained would contain the true population mean. A 99% confidence interval represents the mean ± 3 standard errors of the mean. The width of the confidence interval is determined by the standard error of the mean and by the sample size; as the sample size increases, the confidence interval becomes narrower as the sample and true population means approach one another. Similarly, a narrow confidence interval suggests that the sample is very representative of the population from which it was drawn. In the situation in which the data are skewed or non-normally distributed, using the median to calculate the confidence interval, rather than the mean, may be preferable. Another option is to “transform” the data to a normal distribution so that it can be analyzed using the mean and standard deviation (see Chapter Six).

“**Confidence limits**” and “confidence intervals” are occasionally used interchangeably in the medical literature. Technically, confidence limits refer to the outer boundaries of the interval while the confidence interval refers to the range of values between the limits. It is helpful when reporting a sample mean as an estimate of a population mean to report the confidence limits as they describe how certain you are that the sample mean represents the true population mean. Confidence intervals also provide information on the size of the difference between groups whereas the more commonly used “p-values” only indicate that a significant difference exists. Some researchers thus prefer to report confidence intervals instead of, or in addition to, p-values as the confidence interval provides more information on the data and its validity than does a p-value.

alone. Further, confidence intervals provide information on important effects that, although not statistically significant, may be useful clinically. They are also helpful in determining whether the sample size was large enough to detect a significant difference to begin with (i.e., whether the study had sufficient statistical power).

COEFFICIENT OF VARIATION, BIAS, AND PRECISION

Several descriptive statistics exist for the special situation in which we wish to compare two laboratory tests or measurements to determine whether they accurately measure the same physiologic parameter. The **coefficient of variation** is one method that is frequently used to describe the reliability of a laboratory test or measurement. It is calculated by dividing the standard deviation by the mean and multiplying by 100%.

$$\text{coefficient of variation} = \text{cov} = \frac{\text{standard deviation}}{\text{mean}} \times 100\%$$

As it has no units, the coefficient of variation can be used to compare two tests which are measured on different scales. It is commonly used to describe the reliability and measurement error associated with electronic instruments and monitors. For example, a cardiac output computer might be described as having a coefficient of variation of 5% which means that the measurement error associated with sequential cardiac output measurements will be less than or equal to 5%.

Two other statistics which are used to gauge measurement accuracy are **bias** and **precision**. Note that this bias is different from **statistical bias** which was discussed in the previous chapter. The bias of two methods is simply the difference between their measurements. For example, if a cardiac output by thermodilution is 3.4 L/min and that by radionuclide ventriculography is 3.2 L/min, the bias involved in these two measurements is 3.4-3.2 or 0.2 L/min. Precision is defined as the standard deviation of the bias (the difference between measurements) and is a measure of the variability of the difference. Bias and precision are both necessary to evaluate the agreement between two methods of measurement.

DESCRIBING NOMINAL DATA

Proportions, percentages, ratios, and rates are used to describe nominal data. A **proportion** is defined as the number of observations possessing a given characteristic of interest divided by the total number of observations. If we are evaluating successful extubation of patients from mechanical ventilation, the proportion of patients successfully extubated is:

$$\frac{\text{successfully extubated patients}}{\text{successfully extubated} + \text{reintubated patients}}$$

A **percentage** is defined as a proportion multiplied by 100%. A **ratio** compares the incidence of an event or disease in one group with that in another group. The ratio of successfully extubated to reintubated patients is therefore:

$$\frac{\text{successfully extubated patients}}{\text{reintubated patients}}$$

The **odds** for an event (see Chapter One) is an example of a ratio. A **rate** is defined as a proportion multiplied by a particular "base" and is expressed per unit time. If, for example, it is known that the proportion of male patients with angina who have an acute myocardial infarction is 0.05 (i.e., 5 out of every 100 patients) per year, the **incidence rate** for myocardial infarction in male patients with angina will be 500 per 100,000 patients per year. Rates will be discussed further in the next chapter.

SUGGESTED READING

1. Emerson JD, Colditz GA. Use of statistical analysis in the *New England Journal of Medicine*. In: Bailar JC, Mosteller F (Eds.). *Medical uses of statistics* (2nd Ed). Boston: NEJM Books, 1992:45-57.
2. Wassertheil-Smoller S. *Biostatistics and epidemiology: a primer for health professionals*. New York: Springer-Verlag, 1990:41-63.
3. Dawson-Saunders B, Trapp RG. *Basic and clinical biostatistics* (2nd Ed). Norwalk: Appleton and Lange, 1994:41-63.
4. Moses LE. Statistical concepts fundamental to investigations. In: Bailar JC, Mosteller F (Eds.). *Medical uses of statistics* (2nd Ed). Boston: NEJM Books, 1992:5-26.
5. O'Brien PC, Shampo MA. Statistics for clinicians: introduction. *Mayo Clin Proc* 56:45-46, 1981.
6. O'Brien PC, Shampo MA. Statistics for clinicians: 1. descriptive statistics. *Mayo Clin Proc* 56:47-49, 1981.
7. O'Brien PC, Shampo MA. Statistics for clinicians: 4. estimation from samples. *Mayo Clin Proc* 56:274-276, 1981.
8. O'Brien PC, Shampo MA. Statistics for clinicians: 9. evaluating a new diagnostic procedure. *Mayo Clin Proc* 56:573-575, 1981.
9. Fletcher RH, Fletcher SW, Wagner EH. *Clinical epidemiology - the essentials*. Baltimore: Williams and Wilkins, 1982:27-31.
10. Bartko JJ. Rationale for reporting standard deviations rather than standard errors of the mean [Editorial]. *Am J Psychiatry* 1985: 1060.
11. Altman DG. Statistics and ethics in medical research: V - analyzing data. *Brit Med J* 1980:1473-1475.
12. Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. *Brit Med J* 1986: 746-750.