1

CHAPTER

Basic Descriptive Statistics

1.1 Types of Biological Data

Any observation or experiment in biology involves the collection of information, and this may be of several general types:

Data on a Ratio Scale

Consider measuring heights of plants. The difference in height between a 20-cm-tall plant and a 24-cm-tall plant is the same as that between a 26-cm-tall plant and a 30-cm-tall plant. These data have a "constant interval size." They also have a true zero point on the measurement scale, so that ratios of measurements make sense (e.g., it makes sense to state that one plant is three times as tall as another). A measurement scale that has constant interval size and a true zero point is called a "ratio scale." For example, this applies to measurements of weights (mg, kg), lengths (cm, m), volumes (cc, cu m), and lengths of time (s, min).

Data on an Interval Scale

Measurements with an interval scale but having no true zero point are of this type. Examples are temperatures measured in Celsius or Fahrenheit: it makes no sense to say that 40 degrees is twice as hot as 20 degrees. Absolute temperatures, however, are measured on a ratio scale.

Data on an Ordinal Scale

Data that can be ordered according to some measurements are on an ordinal scale. Examples would be rankings based on size of objects, the speed of an individual relative to another individual, the depth of the orange hue of a shirt, and so on. In some cases (e.g., size), there may be an underlying ratio scale, but if all that is provided is a ranking of individuals (e.g., you are told only that tomato genotype A is larger than tomato genotype B, not how much larger), there is a loss of information if we are given only the ranking on an ordinal scale. Quantitative comparisons are not possible on an ordinal scale (how can one say that one shirt is half as orange as another?).

Data on a Nominal Scale

When a measurement is classified by an attribute rather than by a quantitative, numerical measurement, then it is on a nominal scale (male or female; genotype AA, Aa or aa; in the taxa *Pinus* or in the taxa *Abies*; etc.). Often, these are called categorical data because you classify the data elements according to their category.

Continuous vs. Discrete Data

When a measurement can take on any conceivable value along a continuum, it is called continuous. Weight and height are continuous variables. When a measurement can take on only one of a discrete list of values, it is discrete. The number of arms on a starfish, the number of leaves on a plant, and the number of eggs in a nest are all discrete measurements.

1.2 Summary of Descriptive Statistics of DataSets

Any time a data set is summarized by its statistical information, there is a loss of information. That is, given the summary statistics, there is no way to recover the original data. Basic summary statistics may be grouped as

(i) measures of central tendency (giving in some sense the central value of a data set) and
(ii) measures of dispersion (giving a measure of how spread out that data set is).

Measures of Central Tendency

Arithmetic Mean (the average)

If the data collected as a sample from some set of observations have values $x_1, x_2, \ldots, x_n$, then the mean of this sample (denoted by $\bar{x}$) is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Note the use of the $\sum$ notation in the above expression, that is,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Median

The median is the middle value: half the data fall above this and half below. In some sense, this supplies less information than the mean since it considers only the ranking of the data, not how much larger or smaller the data values are. But the median is less affected than the mean by "outlier" points (e.g., a really large measurement or data value that skews the sample). The LD 50 is an example of a median: the median lethal dose of a substance (half the individuals die after being given this dose, and half survive). For a list of data $x_1, x_2, \ldots, x_n$, to find the median, list these in order from smallest to largest. This is known as "ranking" the data. If n is odd, the median is the number in the $1 + \frac{n-1}{2}$ place on this list. If n is even, the median is the average of the numbers in the $\frac{n}{2}$ and $1 + \frac{n}{2}$ positions on this list.

Quartiles arise when the sample is broken into four equal parts (the right end point of the 2nd quartile is the median), quintiles when five equal parts are used, and so on.

Mode

The mode is the most frequently occurring value (or values; there may be more than one) in a data set.

Midrange

The midrange is the value halfway between the largest and smallest values in the data set. So, if x_{min} and x_{max} are the smallest and largest values in the data set, then the midrange is

$$\bar{x}_{\text{mid}} = \frac{x_{\text{min}} + x_{\text{max}}}{2}.$$

Geometric Mean

The geometric mean of a set of n data is the nth root of the product of the n data values,

$$\bar{x}_{\text{geom}} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}.$$

The geometric mean arises as an appropriate estimate of growth rates of a population when the growth rates vary through time or space. It is always less than the arithmetic mean. (The arithmetic mean and the geometric mean are equal if all the data have the same value.)

Harmonic Mean

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the data,

$$\bar{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}.$$

It also arises in some circumstances as the appropriate overall growth rate when rates vary.

Example 1.1 (Describing a Data Set Using Measures of Central Tendency)

After developing some heart troubles, John was told to monitor his heart rate. He was advised to measure his heart rate six times a day for 3 days. His heart rate was measured in beats per minute (bpm).

65	70	90	95	82	84
61	83	120	83	72	70
72	71	92	85	102	69

(*Continued*)

(a) What was John’s mean heart rate over the 3 days? Calculate the three different means (arithmetic, geometric, and harmonic).
(b) What was John’s median heart rate?
(c) What were the modes of John’s heart rate?
(d) What was the midrange of John’s heart rate?

Solution:

(a) Arithmetic mean:

$$\bar{x} = \frac{65 + 70 + 90 + \cdots + 85 + 102 + 69}{18} = 81.4$$

Geometric mean:

$$\bar{x}_{\text{geom}} = (65 \times 70 \times 90 \times \cdots \times 85 \times 102 \times 69)^{1=18} = 80.3$$

Harmonic mean:

$$\bar{x}_{\text{harm}} = \frac{18}{\frac{1}{65} + \frac{1}{70} + \frac{1}{90} + \cdots + \frac{1}{85} + \frac{1}{102} + \frac{1}{69}} = 79.2$$

Notice that the three means do not yield equal values.

(b) Arranging the numbers from smallest to largest, we get

61 65 69 70 70 71 72 72 82
83 83 84 85 90 92 95 102 120

Since there are 18 data points, we take the average of the middle two numbers: 82 and 83. Thus, the median is 82.5.

(c) There are three modes in this data set: 70, 72, and 83.

(d) Midrange: $\bar{x}_{\text{mid}} = \frac{61 + 120}{2} = 90.5$. Notice that this is different from the median.

Measures of Dispersion

Range

The range is the largest minus the smallest value in the data set: $x_{\max} - x_{\min}$. This does not account in any way for the manner in which data are distributed across the range.

Variance

The variance is the mean sum of the squares of the deviations of the data from the arithmetic mean of the data. The *best* estimate of this (take a good statistics class to find out how *best* is defined) is the sample variance, obtained by taking the sum of the squares of the differences of the data values from the sample mean and dividing this by the number of data points minus one,

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

where n is the number of data points in the data set, x_i is the ith data point in the data set x, and $\bar{x}$ is the arithmetic mean of the data set x.

Standard Deviation

The variance has square units, so it is usual to take its square root to obtain the standard deviation,

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

which has the same units as the original measurements. The higher the standard deviation s, the more dispersed the data are around the mean.

Both the variance and the standard deviation have values that depend on the measurement scale used. So measuring body weights of newborns in grams will produce much higher variances than if the same newborns were measured in kilograms. To account for the measurement scale, it is typical to use the coefficient of variability (sometimes called the coefficient of variance): the standard deviation divided by the arithmetic mean, which is dimensionless and has no units. This coefficient of variability is thus independent of the measurement scale used.

Example 1.2 (Describing a Data Set Using Measure of Dispersion)

In a summer ecology research program, Jane is asked to count the number of trees per hectare in five different sampling locations in King's Canyon National Park in California. Each sampling location is referred to as a plot, and each plot is a different size. Here are the data she collected:

Plot Size (hectares)	No. of Trees in Plot
1.50	20
2.30	31
1.75	43
3.10	58
2.65	29

Given the data Jane collected, (a) construct the data set that represents the number of trees per hectare for each of the five plots and then calculate the (b) range, (c) variance, and (d) standard deviation of the data set you constructed.

(*Continued*)

Solution:

(a) For each plot, the number of trees per hectare is

$$\frac{\text{\# trees in plot}}{\text{plot size}}.$$

For example, the first plot has $20/1.5 = 13.3$ trees/hectare. Thus, the data set that represents the number of trees per hectare for each of the five plots is

$$x = \{13.3, 13.5, 24.6, 18.7, 10.9\}.$$

(b) To calculate the range, we need to know x_{max} and x_{min} (the maximum and minimum values of the data set x). Looking at the data set constructed in (a), $x_{\text{min}} = 10.9$ and $x_{\text{max}} = 24.6$. Thus,

$$\text{range} = 24.6 - 10.9 = 13.7.$$

(c) Recall that to calculate the variance of a data set, you must first know the arithmetic mean of that data set. For the data set constructed in (a),

$$\bar{x} = \frac{13.3 + 13.5 + 24.6 + 18.7 + 10.9}{5} = 16.2.$$

Then, the variance is

$$\begin{aligned} s^2 &= \frac{1}{5-1}\Big[(13.3 - 16.2)^2 + (13.5 - 16.2)^2 + (24.6 - 16.2)^2 \\ &\qquad + (18.7 - 16.2)^2 + (10.9 - 16.2)^2\Big] \\ &= \frac{1}{4}\Big[(-2.9)^2 + (-2.7)^2 + (8.4)^2 + (2.5)^2 + (-5.3)^2\Big] \\ &= \frac{1}{4}[8.41 + 7.29 + 70.56 + 6.25 + 28.09] \\ &= \frac{1}{4}[120.6] \\ &= 30.15. \end{aligned}$$

(d) Recall that the standard deviation of a data set is the square root of the variance of that data set. Thus, the standard deviation is

$$s = \sqrt{30.15} = 5.491.$$

Dispersion over Nominal Scale Data and the Simpson Index

All the above measures of dispersion apply to ratio scale data. For nominal scale data, there is no mean or variance that makes sense, but there certainly can be a measure of how spread out the data are among the various categories, a concept called diversity. In ecology, the two main factors taken into account when measuring diversity are richness and evenness. Species richness is the number of different species present, while evenness is a measure of the relative abundance of the different species making up the richness of an area. The area has uneven diversity if virtually all the individuals found are of one species with only rare individuals of the other species. The area has even diversity if all species have the same abundances. Simpson's index of diversity (SID) is one of several diversity indices. The SID represents the probability that two individuals randomly selected from a sample will belong to different species. In a certain area or sample, let

$$D = \sum_{i=1}^{S} \frac{n_i(n_i - 1)}{N(N - 1)},$$

where n_i is the number of individuals in species i, N is the total number of individuals, and S is the number of species. Then, the SID is

$$\text{SID} = 1 - D.$$

When SID is close to 1, the sample is considered to be highly diverse.

1.3 Matlab Skills

If you are not familiar with the software Matlab, review "Getting Started with Matlab" in Appendix A.

Entering Data Sets in Matlab

In Matlab, data sets are entered as arrays, and arrays are denoted with square brackets: [ ]. If we wanted to enter the trees per hectare data from Example 1.2, we would type

```
[13.3 13.5 24.6 18.7 10.9]
```

into Matlab. Notice that the data points in the set are separated by spaces. If we want to refer back to this data set using Matlab, we need to name the data set. In Example 1.2, we called the data set x. To call the data set x in Matlab, we type

```
x = [13.3 13.5 24.6 18.7 10.9]
```

into Matlab. Now, whenever we want to refer back to our data set, we can just use x instead of typing the entire data set again.

Table 1.1. Matlab commands for a variety of descriptive statistics. In each case, *x* refers to the data set.

Command	Description
`mean(x)`	Returns arithmetic mean of data set *x*
`prod(x)ˆ(1/length(x))`	Returns geometric mean of data set *x*
`geomean(x)`	Returns geometric mean of data set *x* (using the Statistics Toolbox is available)
`length(x)/sum(1./x)`	Returns harmonic mean of data set *x*
`harmmean(x)`	Returns harmonic mean of data set *x* (using the Statistics Toolbox is available)
`median(x)`	Returns median of data set *x*
`mode(x)`	Returns mode of data set *x*
	(when there are multiple values occurring equally frequently,
	`mode(x)` Returns the smallest of those values)
`min(x)`	Returns minimum value of data set *x*
`max(x)`	Returns maximum value of data set *x*
`var(x)`	Returns the variance of data set *x*
`std(x)`	Returns the standard deviation of data set *x*

Calculating Descriptive Statistics in Matlab

Now that we know how to enter our data sets into Matlab, we can use Matlab to quickly compute basic descriptive statistics. Table 1.1 shows the commands for the descriptive statistics described earlier in this chapter.

Each of the commands in Table 1.1 returns its corresponding answer and names the answer `ans`. If we wish to save the answer for future use, we must name the output of the command. For example, if we wish to save the arithmetic mean, we can type

```
xbar = mean(x)
```

into Matlab. If you are typing this into the command window, you will see that the value that is returned is named `xbar`.

Notice there are no commands for calculating the range or the midrange. We can calculate these, however, by using the min and max commands. To calculate the midrange, we use

```
(min(x)+max(x))/2
```

and to calculate the range, we use

```
max(x)-min(x)
```

As an example, suppose we wanted to calculate the mean, median, mode, midrange, geometric mean, harmonic mean, range, variance, and standard deviation for the data set in Example 1.1.

The following shows the input typed into the command window (always proceeded by ») and its corresponding output:

Command Window

```
>> y = [65 70 90 95 82 84 61 83 120 83 72 70 72 71 92 85 102 69]
y =
  Columns 1 through 11
    65    70    90    95    82    84    61    83   120    83    72

  Columns 12 through 18
    70    72    71    92    85   102    69

>> ybar = mean(y)
ybar =
   81.4444

>> ymed = median(y)
ymed =
   82.5000

>> ymode = mode(y)
ymode =
    70

>> ymidrange = (min(y)+max(y))/2
ymidrange =
   90.5000

>> ygeo = geomean(y)
ygeo =
   80.2747

>> yharm = harmmean(y)
yharm =
   79.1871

>> yrange = max(y)-min(y)
yrange =
    59

>> yvar = var(y)
yvar =
  217.3203

>> ystd = std(y)
ystd =
   14.7418
```

1.4 Exercises

1.1 The capacity for physical exercise (in seconds) was determined for each of 11 patients who were being treated for chronic heart failure.

906 1320 711 1170 684 1200 837 1056 897 882 1008

(a) Determine the mean and the median of the data.
(b) Determine the geometric and harmonic means of the data.
(c) How do the three different measures of the mean differ?

1.2 Daily crude oil output (in millions of barrels) for the U.S. is shown below for the years 1971 to 1990.

9.45	9.40	9.25	8.75	8.30	8.10	8.25	8.70	8.55	8.60
8.55	8.65	8.70	8.70	8.91	8.60	8.20	7.70	7.20	6.75

Compute the mean, median, and mode for the data.

1.3 Suppose the scale of a data set is changed by multiplying each measurement by a positive constant. How would this affect the mean, median, mode, and range?

1.4 Ten hospital employees on a standard American diet agreed to adopt a vegetarian diet for 1 month. Below is the change in the serum cholesterol level (before − after).

49	−10	27	13	36
19	48	21	8	16

(a) Compute the median and mean change in cholesterol.
(b) Compute the range, variance, and standard deviation of the data. Are the data fairly spread out or close together?

1.5 Twelve sheep were fed pingue (a toxin-producing weed of the southwestern United States) as a part of an experiment and died as a result. The time of death in hours after the ingestion of pingue for each sheep follows:

44	27	24	24	36	36
44	120	29	36	36	36

Compute the range, variance, and standard deviation of the sample.

1.6 The National Weather Service reports data on the number of hurricanes to strike the United States in decades in the last century (using the Saffir-Simpson category). Calculate the mean of the number of hurricanes per decade.

Decade	No. of Hurricanes
1901–1910	18
1911–1920	21
1921–1930	13
1931–1940	19
1941–1950	24
1951–1960	17
1961–1970	14
1971–1980	12
1981–1990	15
1990–2000	14

1.7 Consider these two sets of data [71]:

$$A = \{0, 5, 10, 15, 25, 30, 35, 40, 45, 50, 71, 72, 73, 74, 75, 76, 77, 78, 100\}$$

$$B = \{0, 22, 23, 24, 25, 26, 27, 28, 29, 50, 55, 60, 65, 70, 75, 85, 90, 95, 100\}$$

For both sets of data, calculate the range, median, the first quartile, and the third quartile. Do these values adequately represent the distribution in each data set?

1.8 Suppose the mean score on a national test is 400 with a standard deviation of 50. If each score is increased by 25, what are the new mean and standard deviation?

1.9 Suppose the mean score on a national test is 400 with a standard deviation of 50. If each score is increased by 25%, what are the new mean and standard deviation?

1.10 Use the following simple data set to calculate the SID for these trees in a particular plot [21]. Interpret your results as a probability.

Species of Trees	No. of Trees in Plot
Eastern rosebud	3
Black oak	4
Post oak	5
White pine	3
Honey locust	1

1.11 Below are some data from the Citizen Science program in the Great Smoky Mountains National Park that record the species of salamanders observed in a particular area in 2000 [21]. Calculate the SID for salamanders in this area using these data.

Species	No. of Salamanders
Desmog	3
Spotted dusky	7
Black bellied	22
Seal	16
Blue ridged Two lined	8
Imitator	2
Southern redback	1
Black chinned	1