Descriptive Statistics for Comprehending the Situation

“...while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.”

Sherlock Holmes in *The Sign of Four*  
(Sir Arthur Conan Doyle, 1890)

The purpose of summary statistics is to take a mass of raw data and condense it in some useful manner. With biosurveillance, the goal is to facilitate an understanding of some aspect of the biosphere. In terms of situational awareness, biosurveillance systems must summarize the data in ways that facilitate an understanding of historical trends, the current state, and how the current state compares to historical trends. That is, as the chapter title says, they must effectively display the situation for the decision maker.

While medical practitioners operate at the individual datum level, meaning they see biosphere one patient at a time, the point of biosurveillance is aggregation of data in order to see events that may not be evident at the individual level. In terms of early event detection, both views are necessary for effective biosurveillance, as there will be some types of events best detected on a case-by-case basis by individual medical practitioners, and there are other types of events best detected in the aggregate.

This chapter focuses on statistical tools and methods useful for perception and comprehension steps of situational awareness. In statistical terms, they would be described as descriptive statistics and exploratory data analysis. Descriptive statistics are first presented in terms of various types of numerical summary statistics followed by types of plots for graphically displaying the data. Exploratory data analysis (EDA) is an interactive approach to exploring and analyzing data, often using the summary statistics just described. The purpose is to explore the data with the purpose of generating insights and hypotheses.
CHAPTER OBJECTIVES

Upon completion of this chapter, you should be able to:

• Define and calculate numerical descriptive statistics for cross-sectional data, including the following.
  – Measures of central tendency: mean, trimmed mean, median, and mode.
  – Measure of variation: variance, standard deviation, standard error, covariance, and correlation.
  – Other summary statistics: percentile, quantile, quartile, interquartile range, range, and mode.

• Define and calculate numerical descriptive statistics for longitudinal data, including repeated cross-sectional statistics, statistics based on moving windows of data and the moving average, and autocorrelation.

• Identify and interpret graphical descriptive statistics for cross-sectional data, including bar charts, histograms, lattice plots, box plots, and scatterplots and scatterplot matrices.

• Identify and interpret graphical descriptive statistics for longitudinal data, including time series plots, repeated cross-sectional plots, and lattice plots conditioned on time.

• Identify and interpret graphical descriptive statistics for spatial and spatio-temporal data, including maps applied to cross-sectional data and repeated with longitudinal data.

• Define exploratory data analysis (EDA), discuss how a biosurveillance system should support EDA, and why effective interface design and data display is important for effective situational awareness.
## MATHEMATICAL NOTATION

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{cov}(x, y))</td>
<td>Sample covariance for variables (x) and (y)</td>
</tr>
<tr>
<td>(i, j)</td>
<td>Indices for either time or the observations</td>
</tr>
<tr>
<td>(k)</td>
<td>Lag</td>
</tr>
<tr>
<td>(\mu)</td>
<td>Population mean</td>
</tr>
<tr>
<td>(n)</td>
<td>Sample size</td>
</tr>
<tr>
<td>(N)</td>
<td>Population size</td>
</tr>
<tr>
<td>(r)</td>
<td>Sample correlation</td>
</tr>
<tr>
<td>(r_k)</td>
<td>Sample autocorrelation for lag (k)</td>
</tr>
<tr>
<td>(R)</td>
<td>Range</td>
</tr>
<tr>
<td>(s)</td>
<td>Sample standard deviation</td>
</tr>
<tr>
<td>(s^2)</td>
<td>Sample variance</td>
</tr>
<tr>
<td>(\text{s.e.}(\bar{y}))</td>
<td>Standard error of the sample mean</td>
</tr>
<tr>
<td>(\hat{\text{s.e.}}(\bar{y}))</td>
<td>Estimated standard error of the sample mean</td>
</tr>
<tr>
<td>(\hat{\text{s.e.}}(\tilde{y}))</td>
<td>Estimated standard error of the median</td>
</tr>
<tr>
<td>(\hat{\text{s.e.}}(s))</td>
<td>Estimated standard error of the sample standard deviation</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>Population standard deviation</td>
</tr>
<tr>
<td>(\sigma^2)</td>
<td>Population variance</td>
</tr>
<tr>
<td>(y)</td>
<td>An observation</td>
</tr>
<tr>
<td>(t)</td>
<td>Time index</td>
</tr>
<tr>
<td>(y(i))</td>
<td>(i)th ordered observation</td>
</tr>
<tr>
<td>(\bar{y})</td>
<td>Sample mean</td>
</tr>
<tr>
<td>(\bar{y}_t)</td>
<td>Moving average at time (t)</td>
</tr>
<tr>
<td>(\bar{y}_{100p%})</td>
<td>(100p%) percent trimmed mean</td>
</tr>
<tr>
<td>(\tilde{y})</td>
<td>Median</td>
</tr>
<tr>
<td>(\tilde{y}_t)</td>
<td>Moving median at time (t)</td>
</tr>
</tbody>
</table>
4.1 Numerical Descriptive Statistics

Numerical descriptive statistics, as should be clear from the name, describe or summarize a set of data in terms of numbers – statistics calculated from the data. Typically the goal is to efficiently summarize a set of data, which is often in terms of where the data is located and how variable it is using statistics such as the average and standard deviation. However, depending on what one is interested in learning about the data, these two statistics may or may not be the most appropriate or effective.

In particular, since biosurveillance data is typically longitudinal, and trends over time are often of interest, simple cross-sectional statistics may be insufficient. Hence, this section is divided into two parts, starting with numerical descriptive statistics that are typically used on cross-sectional data. It then proceeds on to how those statistics can be applied to longitudinal data, as well as numerical descriptive statistics defined explicitly for longitudinal data.

4.1.1 Descriptive Statistics for Cross-sectional Data

As discussed in Chapter 2 (see page 28), cross-sectional data is collected on subjects (such as individuals, hospitals, or regions) at the same point of time. Statistics may then be used to summarize this data, with the intention of reducing the mass of data into a few numbers that usefully provide information about the situation at a particular point in time.

Measures of Central Tendency

*Measures of central tendency* or *measures of location* are typically used to quantify where the “center” of the data is located. The word center is in quotes as there are a number of common measures of central tendency, each of which quantifies the “center” in a different way. The most common measure is the *mean* which is the average of a set of observations in either a sample or a population. That is, the mean is the sum of the observations divided by the total number of items in the sample or population.

Mathematically, for data from an entire population of size $N$, where each observation is denoted $y_i$ and $i = 1, 2, 3, \ldots, N$, the *population mean* $\mu$ is

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i. \quad (4.1)$$

For data from a sample of size $n$, the *sample mean* $\bar{y}$ is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \quad (4.2)$$
Note that the calculation is really the same. The difference is simply a matter of what is being averaged: the entire population or some subset (i.e., sample) of the population. In terms of the notation, a capital letter $N$ is typically used to indicate the number of observations in a population, while the small letter $n$ denotes the number of observations in a sample from the population. In a similar way, the population mean is denoted by the Greek letter $\mu$ while the sample mean is denoted by $\bar{y}$. Often $\mu$ is unobserved and the sample mean $\bar{y}$ is used to estimate the population mean. Also note that the mean does not have to be one of the observed data points and, in fact, it need not even be close to any particular observed value.

A common alternative to the mean as a measure of central tendency is the median. For a sample of data, the median, denoted $\tilde{y}$, is the observation that divides the sample in half. For an odd number of observations, it is literally the value of the middle observation of the ordered data. For an even number of observations, it is the average of the middle two points of the ordered data.

To define the median mathematically, let $y_{(i)}$ denote the $i$th ordered observation (which is also sometimes referred to as the $i$th order statistic). Using this notation, then the sample median is

$$\bar{y} = \begin{cases} y(\frac{n+1}{2}), & \text{if } n \text{ is odd} \\ \frac{y(\frac{n}{2}) + y(\frac{n}{2}+1)}{2}, & \text{if } n \text{ is even}. \end{cases} \quad (4.3)$$

With $n$ replaced by $N$, the calculation for the population median is the same as the sample median.

While both the median and the mean are measures of central tendency, they are distinctly different. In particular, when the data contain one or more outliers, meaning data points that are unusually large or small when compared to the rest of the data, the median is sometimes preferred to the mean. This is because the median is less affected by the outliers.

To illustrate, consider a sample of the following seven ordered observations: \{0, 1, 2, 2, 2, 3, 4\}. These might be, for example, daily ILI counts at a small community hospital for one week. Here the mean equals the median which equals 2. When data is symmetric about the mean, as with the data in this simple example, the mean is equal to the mean. Yet, what if by accident a typo resulted in the first entry being 70, so that the observed ordered observations became \{1, 2, 2, 2, 3, 4, 70\}. Even with the typo, the median remains unchanged at 2, but the mean becomes 12 – clearly a big difference for a change in one data point.

In addition, the use of the mean as a measure of central tendency in the second case with the typo outlier should seem a bit dubious, since the resulting sample average of 12 is larger than all but one of the observations. In contrast, the median is the same regardless of whether the last observation is 4 or 70 or 1,000,000. Whenever data is very skewed, either in the direction of one
or more significantly larger or smaller observations, then the median may be preferred as the measure of central tendency.

The point is that the median is robust with respect to outliers, meaning it is generally not affected much by their presence, while the mean can be very sensitive to one or more outliers. Thus, if the data are likely to contain large or small values atypical of most of the data, then the median may be the preferred measure of central tendency. Similarly, if by the term central tendency one means a typical value in the middle of the data, then it often makes sense to use the median. If, on the other hand, what is of interest is the average value in the data, then the mean is the appropriate measure.

**Example 4.1.** Using the clinic data introduced in Chapter 2 (see page 33), describe the typical age of a patient presenting for the month of January.

**Solution:** In January, 13,223 individuals presented at the clinics. The mean age was \( \bar{y} = 23.8 \) years and the median age was \( \tilde{y} = 23 \) years. Thus, the average patient was almost 24 years old, while roughly half of those presenting were less then 23 years old and half were older than 23 years.

Note that the mean and median for this data are quite close, but the mean is greater than the median. This occurs because there are some ages quite a bit larger than the median, but ages cannot be less than 0, so the data is moderately skewed towards older ages resulting in the mean being larger than the median. However, the difference between the two is modest and thus, in this case, it matters little which is used to describe the location of the data.

The trimmed mean is another measure of central tendency that can be a useful compromise between the mean and the median. For \( 0 \leq p \leq 0.5 \), the 100p percent trimmed mean is calculated by first discarding the 100p percent largest and smallest observations in the data and then averaging the remaining observations.

Mathematically, let \( y(1) \leq y(2) \leq \ldots \leq y(n) \) be the order statistics of \( n \) observations. Then the trimmed mean \( \bar{y}_{tr(p)} \) is defined as

\[
\bar{y}_{tr(p)} = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} y(i),
\]

where \( m = p \times n. \) If \( p \times m \) is not an integer, simply round it down to the nearest integer.

Note that the 100p percent trimmed mean actually deletes \( 2 \times 100p \) percent of the data. Also note that \( \bar{y}_{tr(0)} = \bar{y} \) and \( \bar{y}_{tr(0.5)} = \tilde{y} \), though for any other value of \( 0 < p < 0.5 \), there is no guarantee that the trimmed mean will be between the mean and median.
Example 4.2. Calculate the 20 percent trimmed mean for the previous community hospital example data, both with and without the typo, and then for the clinic data of Example 4.1.

Solution: First, for the community hospital example, $n = 7$. With $p = 0.2$, $0.2 \times 7 = 1.4$, so $m = 1$. Thus, the original set of ordered data \{0, 1, 2, 2, 3, 4\} is trimmed to \{1, 2, 2, 3\}. So

$$\bar{y}_{tr(0.2)} = \frac{1 + 2 + 2 + 2 + 3}{5} = \frac{10}{5} = 2.0 \text{ patients/day}.$$ 

In a similar way, the ordered data with the typo, \{1, 2, 2, 3, 4, 70\}, is trimmed to \{2, 2, 2, 3, 4\}. Thus,

$$\bar{y}_{tr(0.2)} = \frac{2 + 2 + 2 + 3 + 4}{5} = \frac{13}{5} = 2.6 \text{ patients/day},$$

which, while somewhat larger than $\bar{y} = 2$, is nearly as large as $\bar{y} = 12$.

Finally, for the clinic data $0.2 \times 13, 223 = 2, 644.4$, so $m = 2, 644$. From the data, then,

$$\bar{y}_{tr(0.2)} = \frac{1}{7,935} \sum_{i=2,645}^{10,579} y(i) = 20.6 \text{ years},$$

where this is an example in which the trimmed mean turns out to be smaller than both the mean (23.8 years) and the median (23 years).

Note that the mean, median, and trimmed mean are all appropriate to use with either continuous or discrete data. For example, imagine some discrete data comprised of the number of children in a sample of families. While it does not make sense to say that any particular family has 2-1/2 children, it is perfectly fine to say that the average number of children per family is 2.5. However, when data are nominal, neither measure works. For nominal data, the mode, which is the data category with the largest number of observations, is the appropriate measure.

Example 4.3. Using the clinic data, summarize the patients who presented in January in terms of race and ethnicity.

Solution: The most complete way to summarize the data is in a table, as in Table 4.1. As it shows, Hispanic is the modal ethnicity, comprising 82.9 percent (10,963 out of 13,223) of those presenting for whom ethnicity was
recorded. Similarly, the modal race is White, comprising 80.6 percent (10,653 of 13,223) of those presenting, again for those for whom race was recorded. And, White Hispanics are the modal race-ethnic category, comprising 72.3 percent (9,560 of 13,223) of those presenting.

Thus far, the discussion has focused on calculating means and medians for cross-sectional data from one population. However, rate-based information is often important in public health surveillance, particularly when the population of interest is changing, either in terms of size or composition, or when comparing between two or more disparate populations. Thus, rather than simply comparing the number of people with a particular disease or syndrome, the appropriate comparison is the rate: the number of people with the disease or syndrome per some appropriate population unit (e.g., per 100,000 people).

There are various types of rates depending on the medically appropriate way to characterize the prevalence of a disease or event. A crude rate consists of a numerator, which is the total number of observed or estimated events, and a denominator, which is the total population at risk, multiplied by a useful population unit such as 100,000. This then gives the event rate in terms of the average number of events per 100,000 people in the population. There are also various adjusted rates. For example, because many health conditions are related to age, age-adjusted rates are often of interest. See Chapters 2.2 and 2.3 of Waller & Gotway (2004) for a more detailed discussion.

Details of calculating the various types of rates aside, the salient point here is that all of the descriptive statistics discussed thus far (and to follow in the rest of this chapter) are appropriate for use on, and can be useful for summarizing, rate data. The calculations are all the same. The difference is that the rates become the data and the resulting statistics are expressed in the units of the rate. Thus, for example, the mean of some rate data is simply average rate and the median of the rate data is simply the median rate.
Measures of Variation

When describing or summarizing a set of data, providing measures of both location and variation are important. That is, while the mean or median provides information about the average or typical observation in the data, they do not give any information about how the rest of the data are dispersed around that location. But often knowledge about whether the rest of the data tends to be near or far from the center of the data is important. Section 4.2 presents graphical means that allow one to visually assess variation in data. This section presents useful numerical summary measures of variation.

The most common measure is the standard deviation. It is based on the sample variance, $s^2$, which is defined as

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$  \hfill (4.5)

The formula for the sample variance in Equation 4.6 is a bit more complicated than for the sample mean in Equations 4.1 or 4.2. Starting within the parenthesis, the formula says to take the difference between each observation in the sample and the sample mean, square the differences, sum them up, and then divide the sum by $n - 1$.

The sample variance is basically the average squared distance of a sample point from the mean. The larger the sample variance, the more the points are spread out around the mean. The smaller the sample variance, the tighter they are around the sample mean.

The calculation for the population variance, $\sigma^2$, is slightly different from the sample variance in that the population mean replaces the sample mean in the formula and the sum of the squared differences is divided by $N$, not $n - 1$:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2$$  \hfill (4.6)

For both the sample and population calculations, the standard deviation is simply the square root of the variance. So, for a sample it is $s = \sqrt{s^2}$ and for the population it is $\sigma = \sqrt{\sigma^2}$. Typically the standard deviation is used to characterize variation because it is in the same units as the mean. So, returning to the example of summarizing the number of children per family, if the mean is in units of the number of children per family then so is the standard deviation.

The range, $R$, is another measure of variation. The range has the advantage that it is easy to calculate because it is just the difference between the largest and smallest observations. Using the order statistic notation, the range for a sample is $R = y_{(n)} - y_{(1)}$ and for a population it is $R = y_{(N)} - y_{(1)}$. 
Example 4.4. Returning to the clinic age data, summarize the variation in ages of those presenting for the month of January.

Solution: In January, the standard deviation of age for those presenting was $s = 20.7$ years. The standard deviation is so large because there was a wide variety of ages, from the minimum of $y_{(1)} = 0$ years (where 0 years denotes a child less than a year old) to a maximum of $y_{(13,223)} = 95$ years. Thus $R = 95 - 0 = 95$ years.

Another important measure of variation is the standard error, which is the standard deviation of a sample statistic. It is a measure of how variable the statistic itself is (as compared to the standard deviation, for example, which is a measure of the variability of the data). The idea is that different samples from the population will result in different data and thus different sample statistics. It is frequently important to know how much a statistic could vary from sample to sample.

Of course, population statistics have zero standard error since the population mean $\mu$ and standard deviation $\sigma$ are fixed quantities.

Typically the standard error of the sample mean is of most interest. For a sample of size $n$, if the population standard deviation $\sigma$ is known, then the standard error of the sample mean is

$$\text{s.e.}(\bar{y}) = \frac{\sigma}{\sqrt{n}}.$$  

(4.7)

However, it is frequently the case that the population standard deviation is not known. In this case, the standard error is estimated by “plugging in” the sample standard deviation in place of the population standard deviation:

$$\hat{\text{s.e.}}(\bar{y}) = \frac{s}{\sqrt{n}}.$$  

(4.8)

Assuming the data is normally distributed and the sample size is reasonably large ($n \geq 30$), then the estimated standard errors of the median and the sample variance are approximately

$$\hat{\text{s.e.}}(\bar{y}) = 1.25 \frac{s}{\sqrt{n}}$$  

(4.9)

and

$$\hat{\text{s.e.}}(s) = 0.71 \frac{s}{\sqrt{n}}.$$  

(4.10)

Equation 4.10 may seem strange, in the sense that it is calculating a measure of variation (the standard error) of a measure of variation (the standard deviation). However, note that $s$ is a sample statistic and, like all other sample statistics, will vary from sample to sample. Thus, while $s$ is a measure of the
variation in the data, the standard error of $s$ is a measure of the variation of the statistic.

Now, if the assumption that the data is normally distributed does not hold, then it is incorrect to use Equations 4.9 and 4.10. Though beyond the scope of this book, an alternative is a computationally-intensive methodology called the bootstrap which avoids having to make any distributional assumptions.

**Other Numerical Summary Statistics**

There are other useful summary statistics in addition to measures of location and variation. One is the *percentile*. The $p$th percentile is the value of an observation in the data such that $p$ percent of the data is less than or equal to that value. For a given observation, the information communicated by its percentile is where that observation ranks with respect to the rest of the data. That is, it specifies how much of the rest of the data is less than or equal to that observation.

A standardized testing example will make the idea more concrete. If a person scores 720 on the SAT, the 720 score is only somewhat informative because it provides no information about where that places the person among the rest of those who took the test. But, if it turns out that a 720 corresponds to the 96th percentile, then it’s now clear what he score means: 96 percent of the other test takers scored 720 or less or, conversely, that only four percent of the test takers did better.

There are some special percentiles. The 100th percentile is the largest (maximum) observation in the data. The median is the 50th percentile. Finally, the 0th percentile is defined to be the smallest (minimum) observation and thus the range is equal to the 100th percentile minus the 0th percentile.

*Quantiles* are often used in the statistical literature instead of percentiles. Quantiles are simply percentiles divided by 100. That is, while percentiles are on a scale of 0 to 100 (percent), quantiles are on a scale from 0 to 1. Those quantiles that divide the data into fourths are called *quartiles*, where the first quartile is the same as the 25th percentile or the 0.25 quantile. The second quartile is the median, and the third quartile is the same as the 75th percentile or the 0.75 quantile. The *interquartile range* or IQR is defined as the 75th percentile minus the 25th percentile.

---

*Example 4.5.* Returning to the clinic age data, further summarize the distribution of ages of those presenting for the month of January.

---

1 Per Equation 4.3, this is precisely true when $n$ is even, and it is approximately true when $n$ is odd. However, for an odd number of observations, as $n$ gets larger the fraction of observations less than or equal to the median approaches 50 percent.
Solution: The ages varied from the 0th percentile of 0 years to the 100th percentile of 95 years. The 25th percentile (first quartile) of the data is 4 years and the 75th percentile (third quartile) is 34 years, so 50 percent of the patients are between those two ages. A first quartile of 4 years also means that fully one-quarter of the clinic patients are babies and very young children. What fraction are older patients, say 65 years or older? For January, 64 years was the 94.2 percentile, so 5.8 percent of the patients were 65 years or older.

Measures of How Two Variables Co-vary

When looking at a data set with more than one variable, it is often natural to ask whether they seem to be related in some way. For two variables $x$ and $y$, one such measure is the sample covariance, denoted by $\text{cov}(x, y)$ and defined as

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}).$$

(4.11)

Covariance is a measure of how two variables co-vary about their means. Note that the data come in pairs, a pair consisting of one $x$ observation and one $y$ observation. In Equation 4.11, each pair is indexed from 1 to $n$. So, $x_1$ goes with $y_1$, $x_2$ goes with $y_2$, on up to $x_n$ with $y_n$. The equation, then, calculates the average of the product of the differences between each $x$ in the pair from the mean of the $x$s and each $y$ in the pair from the mean of the $y$s.

Covariance is a measure of both the strength and direction of the linear relationship between $x$ and $y$. If the covariance is a large number (either positive or negative) then the strength of linear association is large, if the covariance is near zero, then the strength of linear association is weak or nonexistent. Similarly, if the sign of the covariance is positive, then the association is positive (meaning the $x$s tend to vary in the same direction as the $y$s); if the sign is negative then the association is negative (meaning the $x$s tend to vary in the opposite direction as the $y$s).

The difficulty is determining when to call a covariance “large” because the covariance depends on how the observation is measured. That is, changing the measurement units changes the value of the computed covariance. This is troublesome, since changing the measurement units (inches instead of feet; grams instead of kilograms) does not change the association between the $x$ and $y$. And, if the association is the same, the numerical measure describing that association should be the same.

The sample correlation solves this problem. Sample correlation is the sample covariance divided by the standard deviation of the $x$ values and the standard deviation of the $y$ values. Denoted by $r$, the sample correlation is thus
\[ r = \frac{\text{cov}(x, y)}{s_x \times s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}. \] (4.12)

Dividing by the standard deviations of \( x \) and \( y \) makes the correlation independent of the measurement scale, so that the correlation is always between -1 and 1. That makes interpretation much easier. A correlation near 1 is a strong positive linear association and a correlation near -1 is a strong negative linear association, where a correlation of either +1 or -1 is a perfect linear relationship. And, a correlation of zero means that there is no linear association between \( x \) and \( y \).

The words “association” and “linear” are used purposely in this explanation. It is possible for two variables to be related in a non-linear fashion yet have zero correlation. So, observing a correlation of zero does not mean there is no association between two variables, only that there is no linear relationship. Also, a non-zero correlation does not mean there is a causal relationship between the variables. The correlation between the two variables can occur for reasons not associated with direct causality, so in the absence of other information, the most that can be said if a non-zero correlation is observed is that there is an association between the variables.

Example 4.6. Using the hospital data introduced in Chapter 2 (see page 37), assess the correlation between each pair of hospitals for daily counts of those presenting with GI and respiratory syndromes.

Solution: As shown in Table 4.2, the GI syndrome data are weakly positively correlated across the hospitals.

<table>
<thead>
<tr>
<th>Correlation ((r)) Hospital</th>
<th>Hospital</th>
<th>Hospital</th>
<th>Hospital</th>
<th>Hospital</th>
<th>Hospital</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital #1</td>
<td>1.00</td>
<td>0.14</td>
<td>0.13</td>
<td>0.21</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>Hospital #2</td>
<td>1.00</td>
<td>0.00</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>Hospital #3</td>
<td>1.00</td>
<td>0.11</td>
<td>0.19</td>
<td>0.11</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Hospital #4</td>
<td>1.00</td>
<td>0.09</td>
<td>0.12</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #5</td>
<td>1.00</td>
<td>0.10</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #6</td>
<td>1.00</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #7</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2. Gastrointestinal syndrome daily count correlation matrix for the seven metropolitan hospitals.

Note that, since \( \text{cov}(x, y) = \text{cov}(y, x) \), there is no need to fill in the lower-left part of the table since it would be duplicative of the upper-right. Also,
diagonal contains all ones since each hospital’s syndrome is perfectly correlated with itself. This should also be clear from Equation 4.12 where, if all the $y$s are changed to $x$s then the numerator and denominator are the same and thus the ratio is one.

### Table 4.3. Respiratory syndrome daily count correlation matrix for the seven metropolitan hospitals.

<table>
<thead>
<tr>
<th>Correlation (r)</th>
<th>Hospital #1</th>
<th>Hospital #2</th>
<th>Hospital #3</th>
<th>Hospital #4</th>
<th>Hospital #5</th>
<th>Hospital #6</th>
<th>Hospital #7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital #1</td>
<td>1.00</td>
<td>0.11</td>
<td>0.20</td>
<td>0.27</td>
<td>0.24</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>Hospital #2</td>
<td>1.00</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
<td>0.07</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Hospital #3</td>
<td>1.00</td>
<td>0.16</td>
<td>0.20</td>
<td>0.21</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #4</td>
<td>1.00</td>
<td>0.18</td>
<td>0.10</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #5</td>
<td>1.00</td>
<td>0.15</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #6</td>
<td>1.00</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital #7</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 shows that the correlations are somewhat stronger for the respiratory syndrome compared to GI. However, while both sets of correlations are weak, the syndromes are clearly correlated across the hospitals, and this makes sense since the hospitals are all from the same geographic area.

### 4.1.2 Descriptive Statistics for Longitudinal Data

The previous location and variability measures can be applied to longitudinal data just as well as cross-sectional data. The only real difference is that when applied to longitudinal data one is summarizing the data in terms of one or more statistics per time period. This should seem rather natural for biosurveillance since the goal is to monitor data for changes in disease incidence over time. Indeed, this is precisely what Figures 2.4 and 2.5 in Chapter 2 show, though the statistic plotted in those figures is simply the total number of patients presenting each day who were classified into a particular syndrome category.

**Example 4.7.** Assess the longitudinal trends in GI syndrome by calculating monthly summary statistics for the clinic data.

**Solution:** Table 4.4 gives the minimum, first quartile, median, standard deviation, mean, third quartile, and maximum monthly values for the GI syndrome data.
Table 4.4. Monthly summary statistics for gastrointestinal syndrome using the clinic data.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum ($y(n)$)</td>
<td>9.0</td>
<td>11.0</td>
<td>10.0</td>
<td>10.0</td>
<td>8.0</td>
<td>13.0</td>
<td>15.0</td>
<td>13.0</td>
<td>17.0</td>
<td>11.0</td>
<td>11.0</td>
<td>10.0</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>7.0</td>
<td>8.0</td>
<td>7.0</td>
<td>7.0</td>
<td>5.0</td>
<td>9.0</td>
<td>10.5</td>
<td>9.0</td>
<td>8.0</td>
<td>8.3</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Mean ($\bar{y}$)</td>
<td>5.6</td>
<td>6.7</td>
<td>5.7</td>
<td>5.8</td>
<td>4.2</td>
<td>7.6</td>
<td>8.3</td>
<td>7.5</td>
<td>7.1</td>
<td>7.0</td>
<td>5.2</td>
<td>6.0</td>
</tr>
<tr>
<td>Std Deviation ($s$)</td>
<td>2.4</td>
<td>2.5</td>
<td>2.2</td>
<td>2.1</td>
<td>1.6</td>
<td>3.2</td>
<td>3.4</td>
<td>2.8</td>
<td>4.0</td>
<td>2.5</td>
<td>2.8</td>
<td>2.5</td>
</tr>
<tr>
<td>Median ($\tilde{y}$)</td>
<td>5.0</td>
<td>7.0</td>
<td>6.0</td>
<td>6.0</td>
<td>4.0</td>
<td>8.0</td>
<td>8.0</td>
<td>7.5</td>
<td>6.5</td>
<td>6.5</td>
<td>5.0</td>
<td>5.5</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>4.0</td>
<td>5.0</td>
<td>3.5</td>
<td>4.0</td>
<td>3.0</td>
<td>5.0</td>
<td>6.5</td>
<td>6.3</td>
<td>4.3</td>
<td>5.8</td>
<td>3.0</td>
<td>4.3</td>
</tr>
<tr>
<td>Minimum ($y(0)$)</td>
<td>0.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Thus, one approach to longitudinal data is the repeated calculation of summary statistics for separate subsets of data that correspond to different periods of time. As Example 4.7 shows, comparison of the statistics can usefully help discern trends and patterns in the data.

While doing the calculations in Table 4.4 on a monthly basis is a bit crude, it hints at EED ideas that we will return to in later chapters, namely observing data over time with the objective of trying to determine as early as possible when disease incidence shows an unusual increase. For example, the jump in the mean and median statistics from December or January is an indication that the flu season started sometime in that period.

Now, in addition to calculating statistics for repeated cross-sections of data, there are methods specifically created for longitudinal data. Many of these come in the form of statistical models, and they will be described in the next chapter. In terms of numerical descriptive statistics, the next two subsections focus on statistics calculated from moving “windows” of data and quantifying correlation over time.

Statistics from Moving Windows of Data

One issue with the Example 4.7 approach for summarizing longitudinal data, at least in terms of trying to detect changes, is that the statistic is only updated monthly. As a result, changes are only visible at the end of the month.
so that, for example, if a flu outbreak occurred in the first week of a month, it would not be visible for three or more weeks.

One solution to this is to decrease size of the temporal “window” over which the calculations are conducted. That is, instead of calculating the statistics each month, calculate them every two weeks or perhaps weekly or even daily. This will mitigate the reporting delay problem, but it does not eliminate it since there can still be a delay of up to the size of the temporal window, whatever that may be.

In addition, there is an inherent trade-off made between the size of the temporal window and the variability of the resulting statistics. Specifically, as the window size is decreased, so is the sample size, and as Equations 4.7-4.10 make clear, as \( n \) gets smaller the standard errors get larger. Thus, the smaller the window the more variable (i.e., the noisier) the statistics become.

Calculating the statistics using a moving window of the most recent \( n \) time periods is one way to address this problem. Now, rather than only calculating a statistic once every \( n \) time periods, the statistic is calculated every time period using the \( n \) most recent periods. For example, at time \( t \) the moving average \( \bar{y}_t \) is calculated as

\[
\bar{y}_t = \frac{1}{n} \sum_{i=t-n+1}^{t} y_i,
\]

where in Equation 4.13 it is assumed there is only one observation per time period. For example, \( y_i \) may be the daily ILI syndrome count on day \( i \) and so \( \bar{y}_t \) is the mean ILI syndrome count for the \( n \) days prior to and including day \( t \). At each new time period, the average is recalculated only using the most recent \( n \) periods so that, for example, at time \( t + 1 \) the moving average \( \bar{y}_{t+1} \) is calculated using observations \( y_{t+1}, \ldots, y_{t+n-1} \).

If there is more than one observation per time period, then there are two options. First, the parameter \( n \) in Equation 4.13 can be redefined to be in terms of observations rather than time and thus the moving average will be based on the most recent \( n \) observations. Alternatively, if there are \( m_i \) observations in time period \( i \), calculate

\[
\bar{y}_t = \left( \frac{1}{\sum_{i=t-n+1}^{t} m_i} \sum_{i=t-n+1}^{t} \sum_{j=1}^{m_i} y_{ij} \right) \left/ \sum_{i=t-n+1}^{t} m_i, \right.
\]

where \( y_{ij} \) is the \( j \)th observation in the \( i \)th time period.

The generalization of this idea to other statistics should be fairly obvious. For example, again assuming there is only one observation per time period, the moving median at time \( t \), \( \tilde{y}_t \), is calculated as

\[
\tilde{y}_t = \begin{cases} 
  y_{(n+1)/2}, & \text{if } n \text{ is odd} \\
  \left[ y_{(n+1)/2} + y_{(n+1)/2+1} \right] / 2, & \text{if } n \text{ is even}, 
\end{cases}
\]
which looks exactly like Equation 4.3. The difference is that the order statistics
in Equation 4.15 are calculated from $y_{t-n+1}, \ldots, y_t$.

**Example 4.8.** Show how the moving average and standard deviation compare
to the monthly statistics for March from Example 4.7.

**Solution:** From Table 4.4, the mean and standard deviation for March are
$\bar{y} = 7.5$ and $s = 2.8$. Setting $n = 10$, so the immediate last two weeks of data
are used in the calculation, Table 4.5 shows the GI syndrome moving average
and standard deviation for March 2nd (March 1st was a weekend day) through
the 31st.

|-----|-----|-----|-----|-----|-----|-------|-------|-------|-------|-------|-------|-----
| $\bar{y}$ | 7.1 | 7.6 | 7.8 | 8.2 | 7.6 | 7.6 | 7.8 | 8.2 | 8.5 | 8.3 | 9.0 | 8.8 | ⋮
| $s$ | 3.8 | 3.6 | 3.5 | 3.8 | 3.0 | 3.0 | 3.3 | 2.8 | 2.8 | 3.0 | 2.4 | 2.5 | ⋮
| ⋮ | 8.4 | 7.9 | 7.5 | 7.4 | 7.0 | 6.5 | 6.1 | 6.4 | 6.2 | 6.5 |
| ⋮ | 3.1 | 2.8 | 2.9 | 2.9 | 2.3 | 2.8 | 2.5 | 2.5 | 2.3 | 2.5 |

**Table 4.5.** Moving average and standard deviation for gastrointestinal syndrome
using the March clinic data with $n = 10$.

Note how the moving averages in Table 4.5 show more subtle trends. For
every example, returning to Table 4.4, the monthly mean GI syndrome rate was
down slightly in March from February (from an average of 8.3 per day to 7.5
per day), and the aggregation in that table gives the impression that the GI
syndrome rate peaked in February. However, Table 4.5 shows that, in fact,
the GI syndrome rate was still increasing through mid-March after which it
dropped dramatically.

The choice of $n$ is critical to the behavior of the moving average, or any
other statistic based on a moving window of data. Specifically, the larger $n$
is, the smoother the statistics will be from time period to time period, which
may help make trends in the data more visible by smoothing out the inherent
variation in the data. However, the smoothness comes at the cost of adding
inertia to the statistic since the larger $n$ is the more historical data is included
in the statistic’s calculation. Thus, with the choice $n$, there is an explicit trade-
off to be made between the desired smoothness in the sequence of statistics
and the responsiveness of the statistic to changes in the population.
Another consideration when choosing \( n \) is whether the data contain regular cycles or other periodicities. For example, Figure 2.9 in Chapter 2 shows the hospital respiratory syndrome data have a day-of-the-week effect that occurs on a repeating seven-day cycle. In the presence of such cycles, it is important to choose \( n \) as a multiple of the cycle periodicity in order to mitigate its effects. Otherwise, the statistic will vary according to the cycle. In the case of the hospital respiratory syndrome data, that means choosing a value for \( n \) that is a multiple of seven.

Finally, the choice of a moving window using the past \( n \) time periods or observations is based on the presumption of prospective monitoring or analysis. A prospective analysis is one that only looks forward in time and, in such cases, the only data available is that which is prior to the current time period. However, in a retrospective analysis, where one is looking back in time, the window can be centered around the time period of interest, using data from both before and after it. For example, assuming \( n \) is odd, the moving average calculation in Equation 4.13 could be modified for a retrospective analysis as follows:

\[
\bar{y}_t = \frac{1}{n} \sum_{i=t-n/2}^{t+n/2} y_i. 
\] (4.16)

Retrospective analyses can be useful and appropriate, for example, when studying how a disease progressed through a population or when comparing how various biosurveillance early event detection algorithms would have performed on real, but historical, data. Of course, actual early event monitoring can only be conducted prospectively.

### Autocorrelation

*Autocorrelation* is the correlation of a longitudinal data set with itself. What autocorrelation quantifies is the similarity between longitudinal observations as a function of the time separation (or “lag”) between them. For some lag \( k \), \( k = 1, 2, \ldots \), the sample autocorrelation function \( r_k \) is

\[
r_k = \frac{\sum_{i=k+1}^{n} (y_i - \bar{y}) (y_{i-k} - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{4.17}
\]

where \( \bar{y} \) is the sample mean taken over the entire sample. Perfect positive or negative autocorrelation at lag \( k \) occurs with \( r_k = +1 \) or \( r_k = -1 \); \( r_k \approx 0 \) indicates little to no autocorrelation for lag \( k \).

A plot of \( r_k \) versus sequential values of \( k \), called a *correlogram*, helps to show whether there are dependencies present in the data such as long-term linear or other trends, short- and/or long-term cycles, etc. Figures 4.1 and 4.2 are the correlograms for the clinic GI and ILI syndrome data for \( 1 \leq k \leq 100 \).
Fig. 4.1. Correlogram for the clinic GI syndrome data for various lags, $1 \leq k \leq 100$.

Fig. 4.2. Correlogram for the clinic ILI syndrome data for various lags, $1 \leq k \leq 100$. 
As the plots show, overall the autocorrelation for the GI syndrome is less than that for the ILI syndrome, and both are relatively modest. The roughly linear decrease in $r_k$ is indicative of long-term trends in the data (which are visible in the raw data, as shown in Figures 2.4 and 2.5). There is no evidence of weekly or monthly cycles in the data, which would have manifested as regular spikes at 5- or 20-day increments (since these clinics operate on a 5-day week).

### 4.2 Graphical Descriptive Statistics

In addition to numerical summaries of data (statistics), which can be very useful for condensing the information from a lot of data down into one or a few numbers, graphical summaries are important for visualizing data. Much like the old saying that “a picture is worth a thousand words,” a good graph can be worth a thousand summary statistics.

Furthermore, properly designed, graphics can effectively provide insight into large masses of data and they can help reveal connections and relationships in the data that might otherwise be hard to detect. As John Tukey said, “The greatest value of a picture is when it forces us to notice what we never expected to see” (Tukey, 1977).

This section will examine graphics useful for summarizing categorical data, such as the bar chart, and graphics useful for continuous data, such as the histogram, the box plot, the scatterplot, and the time series plot. As with the numerical summary statistics, the graphics are presented in terms of their application to cross-sectional and longitudinal data, as well as a third category: spatial data.

Many types of statistical graphics, such as bar charts and scatterplots, are now familiar to the average reader. Not only are they fairly commonly used in the popular media, particularly the financial pages of any of the major newspapers, but ubiquitous software such as Microsoft Excel make it easy to generate such plots. On the other hand, more specialized plots, such as box plots, histograms, and bubble charts, may be less familiar.

### 4.2.1 Graphical Methods for Cross-sectional Data

Section 4.1.1 introduced numerical descriptive statistics for cross-sectional data. The graphical methods presented in this section are useful compliments, particularly for further summarizing and displaying cross-sectional data. Appropriately designed, implemented, and presented, these graphical methods give the user intuitive insights into the data that may not be achievable otherwise.
4.2 Graphical Descriptive Statistics

Bar Chart

Bar charts are useful for summarizing categorical data, particularly for visually comparing the relative sizes of the various categories. Bar charts typically display the category titles on one axis and either counts or percentages on the other. For example, Figure 4.3 contains separate bar plots for race and ethnicity for clinic patients presenting in January using the count data from Table 4.1. In these figures, the bars are plotted vertically, though they could just as well have been plotted horizontally. In fact, when the category names are long, it is often preferable to plot the bars horizontally so that the names can also be written out horizontally on the $y$-axis.

![Bar Chart Example]

Fig. 4.3. Examples of bar charts: Plots of race and ethnicity for clinic patients presenting in January. See Table 4.1 for the exact counts.

As shown in Figures 4.4 and 4.5, bar charts can also display sub groupings, either by breaking the bars up to show the constituent subgroups in a stacked bar chart, or by showing each main group as a set of bars in a side-by-side bar chart. Stacked bar charts facilitate comparing between the main groupings (race in the figures) while also allowing for comparison of the relative sizes of the sub groupings within each main group (ethnicity in the figures). In contrast, side-by-side bar charts allow direct comparison of the sizes of the sub groupings between the main groupings, but this comes at the cost of not being able to directly compare the sizes of the main groups.

Histogram

A histogram is akin to a bar chart but for continuous data. As just discussed, bar charts are for discrete data, and hence each bar in the chart corresponds
Fig. 4.4. A stacked bar chart of ethnicity by race for clinic patients presenting in January for the data from Table 4.1.

Fig. 4.5. A side-by-side bar chart of ethnicity by race for clinic patients presenting in January for the data from Table 4.1.
to a distinct category in the data. In contrast, the histogram is applied to continuous data by dividing the real line up into contiguous ranges (typically called “bins”) for which the number of observations that fall into each bin are summed up. Then bars are plotted, where the height of each bar corresponds the observed values that fall in the range of the bar.

Histograms are often used to gain insight into the distribution of the data. What is its shape? Where are most of the observations located? How spread out is the data? Are there any unusual outliers or concentrations of the data?

Figure 4.6 is a histogram of those presenting at the clinics in January, where age is divided into five-year ranges. This data was numerically summarized in Examples 4.1 and 4.4, where it was found that $\bar{y} = 23.8$ and $\tilde{y} = 23$ years with $s = 20.7$ and $R = 95$ years. This is visually evident in the histogram, which shows a bump around the mean, with quite a bit of variation around it, and extreme values that extend from zero to 95 years.

In Figure 4.6, the heights of the bars denote the number of clinic patients in each age range. Visually prominent in the plot is a spike at the 0-5 year age range, clearly showing that a large number of clinic patients are young children. Note that the vertical axis can also be expressed in terms of the percentage or fraction of the total observations that fall within the range of each bar.

If examining the distribution of young patients is of most interest, then it can be useful to plot the data with age rescaled. That is, one way to make the distribution of such data easier to “see” is to transform it and plot the
transformed values. For example, Figure 4.7 shows the square root transformation of age, which compresses the older ages and spreads out the younger ages. What the histogram of the transformed data shows is that the majority of the young patients are, in fact, less than one year. So, a large fraction of the clinic effort is devoted to newborn health care.

![Histogram of square root transformed age for clinic patients presenting in January.](image)

**Fig. 4.7.** A histogram of square root transformed age for clinic patients presenting in January.

Now, the point here is that there can be utility in plotting data in units other than the natural units. For example, with data that contains outliers or that is highly skewed, most of the data can end up being compressed into one or only a few bars. In such cases, transformations can reveal interesting aspects of the data. Another alternative is to restrict the horizontal axis plotting range to focus only on the region of interest.

Given the ubiquity of statistical software, the details of how to manually construct histograms will be ignored here. However, be aware that every software package makes certain choices when drawing a histogram based on default values in the software. Two important choices that need to be made when drawing a histogram: (1) how many “bins” should be used, and (2) where those bins should be located. A common choice is to begin the first bin at the data’s minimum value and it end the last bin at the maximum value. In terms of the right number of bins, when \( n \) is relatively small a good choice is often \( \sqrt{n} \), and when \( n \) is large a good choice is often \( 10 \log_{10} n \).

While the software within the biosurveillance system will likely have defaults, it is important to allow the user to make alternative choices, both in terms of transformations, but also in terms of bin choices, since the informa-
tion communicated with the plot can vary with these choices. For example, Figure 4.8 shows what the same age data looks like with alternative histograms, where the left plot only has two bins, the middle plot has five bins, and the right plot has 96 bins. The plot with only two bins shows almost no details in the data except that the vast majority of clinic patients are less than 50 years old. The histogram with five bins also obscures most of the data and, in fact, is somewhat misleading in the way it shows a fairly constant decrease in counts across the 20-year age bands. The right plot shows the most details and is really just a bar chart since there is a discrete bar for each age year. However, using so many bins is really only effective because the sample size is so large.

**Lattice (or Trellis) Plots**

Lattice plots (also known as trellis plots) are an array of some type of statistical graph of one variable subset according to the values of one or more categorical variables. The categorical variables are also referred to as conditioning variables. The idea is to create a series of plots of one variable, say histograms for the counts of some syndrome, by separate levels of some categorical variable such as zip code or gender or ethnicity. For example, Figure 4.9 is a lattice of histograms of hospital respiratory data for 2003. Here the conditioning variable is hospital number, which results in a lattice of seven histograms, one for each hospital.

What Figure 4.9 shows is that the distributions of respiratory syndrome counts differ by hospital. For example, Hospitals #2 and #3 tend to have much smaller daily counts than the other hospitals. On the other hand, Hospitals #1 and #7 have “right tails” that extend much further than the other hospitals,
Fig. 4.9. Lattice plot of hospital respiratory data for 2003. Here the conditioning variable is the hospital number which results in a lattice of seven histograms, one for each hospital.

which means those hospitals can have much larger daily respiratory counts than the other hospitals.

In some ways, a lattice plot is nothing more than a set of repeated graphs, one for each category of the conditioning variable or set of categories of the conditioning variables. Three things make lattice plots more useful than simply manually repeating some plot by variable levels. The first is that good software facilitates exploring the data by making it easy to generate lattices. Second, and more importantly, the graphs in the lattice are all plotted with the horizontal and vertical axes on the same scale. This makes the plots easier to compare across the various categories. Third, the lattice can be conditioned on more than one categorical variable, which allows for the discovery of more complicated relationships in the data.
Example 4.9. Create and interpret a lattice plot of clinic patient age histograms conditioned on gender and ethnicity.

Solution: Figure 4.10 is the lattice plot of clinic patient age histograms conditioned on gender and ethnicity. The plots clearly show that there is little to no difference in age distributions by ethnicity, but there is by gender.

The difference by gender arises because the clinics provide a significant amount of maternity and well-baby care, where for all ethnic categories most of the patients between their late teens and mid-30s are female. These are the mothers coming to the clinics for maternity care. Similarly, for both gender categories there is a pronounced spike at the younger ages. These are the newborns and young children being brought to the clinic for well-baby care.

Note that the spike for male children is much higher than the spike for female children. This does not mean that there are more male children than female because the bars in each of the plots reflect the percent of those who fall in each bar out of the total of patients in that gender-ethnicity category. Thus, because both the mothers and female babies fall into the same gender category, the female babies are a much smaller fraction of the total than the males.

Box Plot

Box plots are useful for depicting the distributions of continuous data. They do so by displaying summary statistics of the data, including the median and the quartiles. As a result, box plots only require one dimension (unlike histograms that need two dimensions). Because the box plot is based on summary statistics, some information is lost, but box plots can still be very informative, particularly when comparing the distributions of two or more sets of data.

As shown in Figure 4.11, to construct a box plot first calculate the median and the quartiles of a set of data. A box is then plotted that connects the first and third quartiles (the 25th and 75th percentiles), and a line is added inside the box to show the median. At each end of the box “whiskers” are added by extending lines right and left from the box that are 1-1/2 times the interquartile range. These lines are then truncated back to the last point contained within the line. Each whisker thus terminates at an actual data point, which means the whiskers will likely be of different lengths. Finally, observations that fall outside of the whiskers are indicated by dots and are designated outliers.

So, a box plot displays a lot of information: a measure of central tendency, the median; measures of how variable the data is as indicated by the IQR (shown via the width of the box) and the length of the whiskers; and potentially unusual points, the outliers. Note that how box plots are drawn can vary
Fig. 4.10. A lattice plot of clinic patient age histograms conditioned on gender and ethnicity.

by software package. For example, when drawing the whiskers some packages allow the user to choose multiples other than 1-1/2 times the IQR and some extend the whiskers all the way to the minimum and maximum values in the data. In addition, sometimes the box is based on statistics called “hinges” which are similar to, but not precisely the same as, quartiles.
4.2 Graphical Descriptive Statistics

Fig. 4.11. An illustration how a box plot is constructed from summary statistics. Box plots are useful for depicting the distribution of continuous data in one dimension.

Example 4.10. Create and interpret a box plot of clinic patients who presented in January.

Solution: Figure 4.12 shows a box plot for the age of clinic patients who presented in January. As with the histogram in Figure 4.6, the box plot shows the distribution of ages is skewed, with the median age in the low twenties. The right end of the box, which is the third quartile, looks to be somewhere in the mid-30 year range, indicating that 75 percent of the patients are younger than about 35 years or so. On the other hand, Figure 4.12 also shows a long right whisker that extends out to around 80 years, with lots of outliers, indicating that there is a long and not inconsequential “right tail” of significantly older patients.

A very useful way to compare populations is via side-by-side box plots where, similar to lattice plots, a separate box plot is created for each population. Side-by-side box plots require a continuous and a categorical measure on each observation; for example, age and zip code. Also like lattice plots, one of the reasons side-by-side box plots are powerful is that all the box plots are graphed on the same scale which facilitates comparisons between the categories.

Example 4.11. Create and interpret side-by-side box plots of the age of clinic patients who presented in January by ethnicity.

Solution: Figure 4.13 shows side-by-side box plots of the age of clinic patients who presented in January by ethnicity. The plot shows that Hispanic patients tended to be younger than non-Hispanic patients. Note, for example, that the median age of the non-Hispanic patients is roughly equal to the 75th percentile of the Hispanic patients and the box for Hispanic patients is shifted towards
Fig. 4.12. A box plot of age for clinic patients presenting in January. Note that box plots cannot show multiple modes in data which means, for example, that the spike from 0-5 years evident in the Figure 4.6 histogram is invisible when the data is depicted with a box plot.

younger ages compared to the non-Hispanic patients. On the other hand, note that the extreme ages are very similar, with the left (lower) whiskers of both groups extending to zero and the right (upper) whisker of the non-Hispanic group extends to about the same age as the greatest outlier for the Hispanic group.

Scatterplot

A scatterplot is a graph of one continuous variable versus another. Scatterplots are useful for summarizing large sets of data, and they are very effective at showing whether there is an association between two variables, where such an association would show up as a pattern in the plot. For example, if a scatterplot shows a distribution of points that fall roughly in an ellipse starting at the lower left corner of the plot and extending to the upper right corner, indicating that larger values of the variable on the horizontal axis are associated with larger values of the variable on the vertical axis, then the two variables will also be positively correlated.

Figure 4.14 is a scatterplot of Hospital #1 versus Hospital #2 respiratory syndrome daily count data. The figure shows there is very little association between the two hospitals in terms of daily respiratory syndrome counts,
Fig. 4.13. Side-by-side box plots of the ages of clinic patients who presented in January by ethnicity. The plot shows that Hispanic patients tended to be younger than non-Hispanic patients.

Where large and small counts at Hospital #1 occur both with large and small counts at Hospital #2. This lack of association most likely exists because the hospitals serve very different populations and respiratory symptoms are not frequently associated with contagious diseases. The plot is consistent with Table 4.3 which shows the correlation is only weakly positive \( r = 0.11 \). Note that the observations in the plot have been “jittered,” meaning a bit of random noise was added to each observation so that days with exactly the same counts for both hospitals did not overplot.

Compare Figure 4.14 to Figure 4.15, which is scatterplot of clinic GI versus ILI daily count data. Here the correlation moderately positive \( r = 0.48 \), which is visually evident in the plot showing that larger GI values tend to be associated with larger ILI values. This makes clinical sense since flu and gastrointestinal symptoms often occur simultaneously in patients. Again note that the observations have been “jittered,” meaning a bit of random noise to mitigate over plotting so that all the data is visible.

A correlation of +1 would correspond to all the points in a scatterplot falling perfectly in a straight line, where every “\( x \)” value would be precisely equal to its associated “\( y \)” value. In contrast, a correlation of -1 would also correspond to all the points falling in a straight line, but the line would go from the upper left corner of the plot to the lower right. Note that scatterplots
can show other relationships between two variables, perhaps non-linear, that may not be evident in the correlation alone. For example, two variables with a perfect sinusoidal relationship between them would have a correlation of zero.

When there are more than two continuous variables, scatterplot matrices can be used. They display in a lattice-like format scatterplots for all pairs of the data. For example, in the hospital data there are seven hospitals. Figure 4.14 only showed the scatterplot for Hospital #1 versus #2, but for $k$ variables there are a total of $k(k - 1)/2$ possible pairwise comparisons.

**Example 4.12.** Create a scatterplot matrix for the hospital respiratory data and interpret it.

**Solution:** Figure 4.16 is the scatterplot matrix for the hospital respiratory data. There are $7 \times 6/2 = 21$ unique scatterplots, where every plot above the diagonal has an equivalent plot below it, just with the axes reversed. Note how the plot in the first row, second column, is the same as the scatterplot in Figure 4.14.

What Figure 4.16 shows is there are no strong associations between any of the hospitals. This is consistent with the correlations in Table 4.3, where the largest correlation is $r = 0.27$ between Hospitals #1 and #4. As previously described, this lack of strong correlation in daily respiratory counts is clinically plausible. The scatterplot matrix also shows there are no non-linear relationships that are not evident in the correlation matrix of Table 4.3.
Fig. 4.16. A scatterplot matrix of the hospital respiratory data. See Table 4.3 for the associated correlation values for each scatterplot. Note that the plot in the first row, second column, is the same as the scatterplot in Figure 4.14.

4.2.2 Graphical Methods for Longitudinal Data

Time Series Plots

Time series data are variables or measures taken repeatedly over time, so in that sense the term “time series” is simply a synonym for longitudinal data. See Chapter 2 (page 28) for additional discussion about types of data. Daily syndrome counts, for example, are time series data and biosurveillance by its very nature focuses on such data.

Time series plots display longitudinal data plotted with respect to time. By convention, the data is plotted with the magnitude of the observation on the vertical axis and time on the horizontal axis in the appropriate units corresponding to how the data was collected. Time series plots are useful for displaying whether there are trends in the data, such as a regular increase in the number of people presenting at a facility with a particular syndrome (perhaps as a result of regular population growth in the region), or whether there are cycles in the data (perhaps as a result of seasonal or other influences).
Referring back to Chapter 2, Figure 2.5 is a time series plot of influenza-like illness syndrome data from the clinics, where each point on the plot is the daily count of the number of people presenting who were classified into the ILI syndrome. In this plot, time is represented in terms of days, where the clinics are not open on weekends or holidays, so there are 252 sequential daily observations in the data.

Figure 2.5 displayed the time series by plotting the raw data and then overlaying a line that represented the trend in the average daily ILI. This is just one way the data can be displayed and there are a variety of alternatives depending on what is intended to be emphasized. For example, in Figure 4.17 the upper left plot just shows the data. In the lower left plot, a smoothed line is added to help direct attention to the trends and cycles in the data (much like Figure 2.5). In contrast, the upper right plot connects all the data with a line and that tends to emphasize the variation in the data from day to day. And the lower right plot adds the smoothing line to try to show both how variable the data is and also the trends and cycles. Which of these is the best, or most appropriate display, depends on the purpose of the plot and what one is trying to communicate with it.

Another alternative for displaying times series data is shown in Figures 2.7 and 2.8, where only the smoothed lines are plotted and the data itself is suppressed. In those figures this was done for the purposes of clarity, where graphing the raw data for all seven hospitals made the plots too busy and virtually unintelligible. However, displaying the time series this way also completely obscures the variation in the data since the data itself is not even shown.

If adding a smoothing line is desired, that naturally leads to the question of how to calculate it. In Figures 2.4-2.5 and 2.7-2.8 lowess was used (see the footnote on page 33 for a brief description). Alternatives also include the moving average, the exponentially weighted moving average, and time series models. These will all be discussed in more detail in Chapter 5. For now, simply note that adding such smoothers to a plot can help show trends that may be more or less discernible to the unaided eye.

Repeated Cross-sectional Plots

Another way to present time series data is via a series of repeated plots, where each plot represents a different (but sequential) time period. This is simply the graphical equivalent to the approach in Section 4.1.2 for summary statistics. For example, box plots repeated over time can be be very informative. Figure 4.18 shows this approach for the clinic ILI data, where each box plot is based on a month of data. Compare Figure 4.18 to Figure 2.5 and the plots in Figure 4.17.

Figure 4.18 does show time trends, but the repeated box plots tend to emphasize the variation within each month. An abstraction of the repeated
Fig. 4.17. Time series plotting variants for the clinic ILI data first presented in Figure 2.5.

box plots is shown in Figure 4.19, where each of the box plot statistics are joined with a line and the region between the lines corresponding to the first and third quartiles is darkly shaded while the region between the lines corresponding to the whiskers is lightly shaded. This plot, then, gives some idea of both the overall trends in the data as well as the variation.

Of course, lattice plots can also be used to display repeated cross-sectional plots if the conditioning variable corresponds to time. For example, Figure 4.20 is a lattice plot of the clinic ILI monthly histograms.

4.2.3 Graphical Methods for Spatial and Spatio-temporal Data

Informative displays of spatial data almost by definition require them to be viewed in the context of a map. However, unlike a lot of data that is displayed on geographic information systems (GIS) as precise points on a map, biosurveillance data are frequently areal data. Areal data is data that is associated with, or can only be linked to an area or region, not a precise location.
Fig. 4.18. Repeated box plots for the clinic ILI data by month

Fig. 4.19. An alternate way to display the box plot statistics from the clinic ILI data shown with repeated box plots in Figure 4.18.
Fig. 4.20. A lattice plot applied to the clinic ILI data, where the conditioning variable is the month.

For example, patients in the clinic data can only be located to residential zip codes, not to precise residential addresses.

Of course, in some data sets the precise residential address of the patients is recorded. However, even if they are known, some aggregation across patients and regions is necessary to discern trends and patterns. Furthermore, there is an inherent ambiguity in location since whatever location is precisely recorded for a given patient only corresponds to one location of many that they likely visit over time. That is, employed people spend a significant portion of their time in at least two locations: their residence and their location of employment. Children in school similarly spend much of their time both at home and at school. Thus, whatever location is recorded is at best a proxy for the general location where an individual spends some or more of their time.

Maps

Figures 4.21 and 4.22 are examples of maps displaying areal data. Figure 4.21 is a map of the clinic population density by zip code. Clinic locations (which are point data) are denoted by the white diamonds, where it is visually evident that the clinics are located in the more densely populated areas. Figure 4.22 is a map of annual ILI count from the clinics by ZIP code. Note how, not surprisingly, the distribution of ILI is directly associated with population size and with clinic location shown in Figure 4.21.
Fig. 4.21. Map of clinic population density by ZIP code. Clinic locations are denoted by the white diamonds, where it is evident that the clinics are located in the more densely populated areas.

Fig. 4.22. Map of annual ILI count from the clinics by ZIP code. Note distribution is directly associated with population size and with clinic location shown in Figure 4.21.
These two maps illustrate that care must be taken with how one interprets results (using maps as well as the numerical statistics and other descriptive described in this chapter) from this type of biosurveillance data. At issue is that this data comes from what statisticians call *convenience sampling*, which means that there is no control by the public health system of whether or how an individual ends up in the data. For the clinic data, for example, it’s likely a complex interaction of socio-economics and living and working locations that affect whether an individual goes to a clinic. As a result, it’s difficult at best, and often impossible, to use results such as that displayed in Figure 4.22 to make inference about ILI incidence in the general population.

**Fig. 4.23.** Weekly ILI rates for a non-outbreak period, seasonal ILI, and the initial wave of H1N1 in 2009. The rate here is simply the total ILI syndrome count for the period of time divided by seven. Note the subtle increase in the weekly ILI for the two outbreak periods compared to the non-outbreak period.
Nonetheless, the data can be useful for detecting changes in disease incidence, at least among the populations who use the facilities being monitored. The key is monitoring and displaying data from the same source over time. Figure 4.23 is an example of such a display, where the weekly ILI rates for the clinic are mapped both for a non-outbreak period (8/1/08-12/11/08), a seasonal ILI period (12/12/08-2/13/09), and the initial wave of H1N1 in 2009 (4/6/09-5/8/09). The plots show a subtle, yet visible, increase in the weekly ILI for the two outbreak periods compared to the non-outbreak period.

One thing that tends to obscure trends in Figure 4.23 is that different zip codes have different non-outbreak weekly ILI rates because those rates are dependent on the size and composition of each ZIP code’s population as well as proximity to the clinics. Particularly for low density regions and those regions distant from the clinics, the non-outbreak rate is low and thus increases from that base rate are difficult to discern on a scale that must span the range of rates across all the regions.

Figure 4.24 solves this problem by plotting the percent change from the base rate, where the base rate was calculated over the non-outbreak period (8/1/08-12/11/08) shown in the top plot of Figure 4.23 and then the percent change is calculated for each week for the flu outbreak from 12/22/08 to 2/16/09 (corresponding to the bottom left plot in Figure 4.23). Here the increase in ILI is much more visible, as is the fact that there was an anomalous drop-off in ILI across the region for the week of 1/12/09.

Looking for increases from a base rate, which ostensibly reflects the “normal” state of a population in terms of disease or syndrome incidence, is the fundamental idea of early event detection. In Figure 4.24 the progression of the flu, as manifested via the proxy measure of ILI syndrome incidence, is quite clear. However, these plots were made retrospectively and with the knowledge of the flu period. The challenge in EED, and any type of prospective monitoring, is that one does not know whether the flu period has started or not.

Indeed, the whole point of monitoring is to sequentially observe the data as it unfolds over time and determine as early as possible then there is enough evidence to conclude that the flu season (or some other outbreak or attack) has begin. However, such a determination can always turn out to be a false positive and a major challenge with EED is to appropriately balance how aggressively one makes an outbreak determination with an acceptable rate of false positives. These issues will be discussed in more detail in Part III.

Note that Figure 4.24 is essentially just a lattice plot of maps using time (weeks) as the conditioning variable. This often is an effective way to look at spatio-temporal data. However, while these types of plots are useful when viewing the data in a static format, such as on a page in a book, animation is another powerful option when the data can be viewed dynamically. For example, on a computer display one might have a slider bar under a map that
controls what point in time is displayed. By sliding the bar to the right, the user can observe how the incidence rate dynamically changes over time.

**Bubble Charts and Other Spatial Displays**

Maps, in the strict cartographic sense, are often not necessary to discern important spatial changes in disease incidence. In fact, maps can be limiting in various ways, such as the fact that the map itself uses two dimensions and thus it is challenging to display multidimensional data on a map. Thus it is difficult to impossible to clearly convey more than one additional dimension on a map. With the addition of animation, that can be increased to two dimensions, where the second is time.
In addition, areal data maps can fool the eye when, as is frequently the case, region size is inversely related to population density. The problem is the human brain tends to equate size with importance. Thus changes in color-coded areal maps may be misleading when the least populated areas are the most noticeable because those areas are the largest.

One solution is a bubble chart which centers bubbles in each area and the size of the bubble represents the magnitude. For example, Figure 4.25 uses bubbles to display the population size for the clinic data. Compare this plot to Figure 4.21, where the ZIP codes at the bottom of the plot are large but have comparatively small populations, while some very small zip codes near the top are small but have much larger populations. This is somewhat more difficult to discern in Figure 4.21, because the size of the ZIP codes confounds the interpretation, while the bubbles in Figure 4.25 more clearly convey the information.

Other options include cartograms which distort a map by making the area of each region proportional to the measure of interest. The result is a recognizable but distorted shape but where the area of each region is resized to match the measure of interest. A cartogram can then convey two dimensional data, where area represents one dimension and color a second. With animation, time can also be added for a total of three dimensions.

4.3 Discussion & Summary

This chapter described numerous numerical and graphical statistical methods for summarizing and displaying data. The goal is to use these methods for enhancing and promoting situational awareness, particularly Level 2 situational awareness – comprehension – which depends on appropriately summarizing the raw data as it is received by a biosurveillance system.

Now, while this chapter has presented a suite of useful numerical and graphical statistical methods, how these methods are implemented in a computer interface is also critically important for facilitating situational awareness. That is, while it is beyond the scope of this text, how the biosurveillance system interface is designed both in terms of usability and graphical display can either enhance or inhibit SA.

The general question of effective statistical graphics design has been addressed by Tufte (2001) and others in the statistical literature. For example, Tufte says,

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

• show the data"
4.3 Discussion & Summary

Bubble Chart of County Population Distribution

Fig. 4.25. Bubble Chart of clinic population density by ZIP code. This is an alternative to Figure 4.21 for displaying the data.

- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from the broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
be closely integrated with the statistical and verbal descriptions of
the data set” (Tufte, 2001, p. 13).

These are critical and relevant guidelines. However, effective interface and
graphical design is situation and context dependent where, for biosurveillance,
as of this writing no studies or research have been done that would provide
appropriate guidance specific to this particular application. As discussed in
the previous chapter, as systems such as BioSense mature, understanding
and improving how biosurveillance system interfaces enhance (or degrade)
situational awareness – particularly when under stress, such as with a large
outbreak – is an important area for future research.

Now, in addition to a system supporting situational awareness through
effective display of information, it must also support the effective exploration
of the data, for example, to help localize possible sources of an outbreak. As
with the numerical and graphical methods, effective exploratory data analy-
sis or EDA is as much a function of system and interface design as it is of
individual statistical methods.

In the statistics literature, EDA is an approach to data analysis that fo-
cuses summarizing the data so that it is easy to understand, often with visual
graphs, without the use of formal statistical models hypotheses. As described
by the National Institute of Standards and Technology (NIST, 2012):

“Exploratory Data Analysis (EDA) is an approach/philosophy for
data analysis that employs a variety of techniques (mostly graphical)
to
• maximize insight into a data set;
• uncover underlying structure;
• extract important variables;
• detect outliers and anomalies;
• test underlying assumptions;
• develop parsimonious models; and
• determine optimal factor settings.”

NIST goes on to say,

“The particular graphical techniques employed in EDA are often quite
simple, consisting of various techniques of:
• Plotting the raw data (such as data traces, histograms, bihis-
tograms, probability plots, lag plots, block plots, and Youden plots.
• Plotting simple statistics such as mean plots, standard deviation
plots, box plots, and main effects plots of the raw data.
• Positioning such plots so as to maximize our natural pattern-
recognition abilities, such as using multiple plots per page.”
Promoting and enhancing EDA in biosurveillance requires more than good static summaries of data, it also requires an interface design that allows the user to easily and appropriately explore the data. Important software features include the ability to “drill down” into the data for details, in order to, for example, facilitate easy identification of individual records associated with an EED signal, and dynamic and interactive graphics. Good examples of statistical software that emphasize data and graphical interactivity include JMP and the GGobi, Mondrian, and Manet packages in R. Important features include linking plots and “brushing” where, for example, highlighting one or more data points in one plot also highlights them in the other plots and in the original database, and the ability to “tour” through the data, particularly higher dimensional data.

Finally, it is important to recognize that this chapter is focused on descriptive and not inferential statistics. Descriptive statistics are used to quantitatively describe a set of data, while inferential statistics are used to learn about some larger unobserved population from an observed sample of data. There are two reasons for the focus on descriptive statistics. The first reason is that the data used in many biosurveillance systems such as BioSense are convenience samples from some larger population. Because they are convenience samples, there is not enough information available to make statistically defensible and quantitatively rigorous inferences from sample to population. The second reason is that often inference from sample to population is not the immediate goal of prospective biosurveillance. Rather, the goal is early detection of and situational awareness about an outbreak as it is occurring.

Now, in many traditional health surveillance systems the goal is inference to some larger population using a sample of data. For example, the Behavioral Risk Factor Surveillance System (BRFSS) is designed to track health conditions and risk behaviors and make inference about these conditions and behaviors to all adults in the United States. This is done via a monthly surveys that collect information on health risk behaviors, preventive health practices, and health care access primarily as related to chronic disease and injury. An inferential approach is necessary, reasonable, and appropriate in this situation. It is necessary since it is impossible to survey the entire population; it is reasonable since trends in chronic conditions change slowly; and, it is appropriate since the health system is interested in population trends, not just those of a particular sample.

For prospective biosurveillance of a rapidly evolving situation, such as the spread of a virulent contagion, this type of formal sample data collection and analysis is impractical at best. That said, even from a prospective biosurveillance perspective, clearly the more that is known about or that can be inferred to an entire population the better. Information about populations can be obtained either by sampling and then doing statistical inference, or by collecting data on the entire population. At least currently, the emphasis seems to be more on acquiring large convenience samples of data that can be transmitted
electronically than on ensuring that the data either supports formal statistical inference or that covers the entire population. Essentially it is a trade-off of timeliness of data for data coverage, though as electronic sources of data continue to expand it may well be that eventually electronic data on most of the population is available to biosurveillance. At that point inference may become irrelevant. Until then, as biosurveillance systems mature, attention should be paid to managing data sources. In so doing, the goal should be clearly articulated, which is either obtaining complete data coverage for entire populations (which, realistically, means obtaining data on as high a fraction of the population as possible) or the careful and appropriate selection of data sources and statistical methods to allow formal inference (including the calculation of margins of error) from sample to population.

**Additional Reading**

For those who would like to delve more deeply into the material, consider the following.

- For additional information on, examples of, and further discussion about numerical descriptive statistics, see most standard undergraduate-level statistics textbooks. However, note that such texts vary in the amount of detail devoted to discussing descriptive statistics. Suggested texts include *Statistics* by Voelker *et al.* (2001, chpts. 3-9), *Statistics* by Freedman *et al.* (1998, chpts. 3-9), *Mathematical Statistics and Data Analysis* by Rice (2006, chpt. 10), and *Using R for Introductory Statistics* by Verzani (2005, chpts. 2-4).


- For more information about exploratory data analysis, the seminal text is *Exploratory Data Analysis* by Tukey (1977).

- To read more about interactive graphics, see *Visual Statistics: Seeing Data with Dynamic Interactive Graphics* by Young *et al.* (2006), *Dynamic Graphics for Statistics* by Cleveland & McGill (1988), and *Interactive and Dynamic Graphics for Data Analysis with R and GGobi* by Cook & Swayne (2007, chpts. 1 & 2). While most of these discuss dynamic graphics in the context of a specific software system, the principles and ideas generalize to any system.
Finally, though only briefly mentioned in this chapter, the rigorous definition of and appropriate estimation approaches for the risk of contracting a disease and the rate of disease in a population are critical in public health surveillance. Refer to Sections 2.2 and 2.3 of *Applied Spatial Statistics for Public Health Data* by Waller & Gotway (2004) for additional discussion in a traditional health surveillance context.