3.2 Numerical Ways to Describe Data

When several observations of the same type are obtained (e.g., gene expression of many genes in a microarray experiment), then it is often desirable to report the findings in a summarized way rather than list all individual measurements. This can be done in different ways. We will next describe the most common numerical measures that summarize categorical and quantitative data.

3.2.1 Categorical Data

Categorical data are most commonly summarized in tables. The tables contain the labels or possible categories and COUNTS of how often these labels were observed in the experiment. If two categorical variables are observed on the same individuals, then the data can be summarized in the form of a two-dimensional CONTINGENCY TABLE.

Example 3.3

To test the effectiveness of two potential yellow fever vaccines A and B, laboratory mice are vaccinated with vaccine type A, or vaccine type B. Some mice are left unvaccinated to function as a control. All mice are infected with the yellow fever virus, and after an appropriate incubation period, live and dead mice are counted. Thus, data are collected on two categorical variables per mouse. One variable describes the type of vaccine the mouse received (A, B, or none) and the other variable states whether the mouse is alive or dead. The experimental results in form of a contingency table look like this:

	A	B	none
Live	7	5	3
Dead	3	7	12

For example, ten mice were vaccinated with the type A vaccine, and of those, seven survived.

3.2.2 Quantitative Data

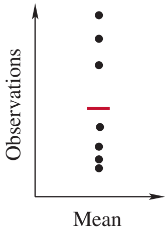
The mean and median are commonly used to describe a typical or center value of a quantitative variable, respectively. Percentiles can be used to convey information on both the center and the spread.

Variance, standard deviation, range, and interquartile range are common measures for the spread or variability in the data.

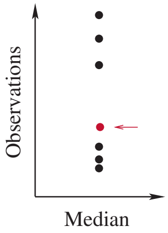
Mean: The average of all observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

where the observations are denoted $x_1, x_2, \ldots, x_n$ and n is the number of observations (sample size). Other commonly used terms for the mean are expected value and average.



Median: The middle observation, if the number of observations n is odd. If n is even, then the median is the average of the two middle observations. Half of the observations are always larger than the median and the other half are smaller.

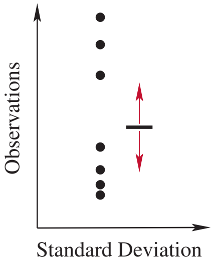


Percentiles: Similar to the median, the p^{th} percentile of the observations is the observation value such that $p\%$ of the observations are smaller than it. Consequently, the median can also be thought of as the 50th percentile. The 25th and 75th percentiles are sometimes referred to as the first and third quartiles, respectively.

Variance: The variance is the average squared distance (deviation) of observations from the mean.

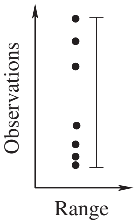
$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

where $x_1, \ldots, x_n$ denote the observations and n is the sample size. Variance is used as a measure of variation in the observations and describes the spread of the distribution.



Standard deviation: This is simply the square root of the variance. Unlike the variance, which has nonsensical units, the units of the standard deviation are the same as the original observation measurements. Commonly used symbols for standard deviation (variance) are s, σ (s^2, σ^2). Standard deviation should not be confused with standard error (Section 3.6).

Range: This is the distance between the largest and smallest observation. The range can be used as a crude measure of spread in the data, but it is more susceptible than the variance (standard deviation) to misrepresent the true variation in the data if there are uncharacteristically large or small observations among the data points.



Interquartile Range: Another measure for variability in the data which is less dependent on extreme (big or small) data values is the interquartile range (IQR). It is the distance between the third and first quartile of the data. This unit is sometimes used to determine whether or not a data point may be considered an outlier.

In R Commander: To compute summary statistics on a quantitative variable, click on "Statistics," "Summaries," and select "Numerical Summaries." Choose the variable(s) for which to compute summary statistics. Click "OK." R will report the mean, the standard deviation (sd), the interquartile range (IQR), and the 0^{th}, 25^{th}, 50^{th} (this is also the median), 75^{th}, and 100^{th} percentile, respectively. The 0^{th} percentile is the smallest and the 100^{th} percentile is the largest observation. Therefore, the range can be obtained by subtracting the smallest from the largest observation. R also reports the sample size (n).

3.2.3 Determining Outliers

An outlier is a data point whose value is very different from that of the majority of the data. Outliers can be caused by errors during the measurement process or during the recording of data. Outliers may also be measurements that differ from the majority of the data points for legitimate reasons (e.g., one patient with a rare extreme reaction to a treatment). The decision of whether or not an observation is an outlier is a subjective one.

A statistical rule of thumb to decide whether or not an observation may be considered an outlier uses the IQR of the data. Using all data points (including possibly suspected outliers), compute the first and third quartile Q_1 and Q_3, as well as the interquartile range IQR $= Q_3 - Q_1$, for the data. An observation can be considered to be an outlier if it is either larger than $Q_3 + 1.5\times$ IQR or smaller than $Q_1 - 1.5\times$ IQR.

If an observation is suspected of being an outlier, double check the recording of the observation to rule out typographical errors. If the measurement cannot be repeated, a statistical analysis should be performed with and without the outlier to determine if this one data point influences the results of the analysis. Because a single data point should not change the outcome of any statistical analysis, removal of the data point may be required. If the data point is omitted in the subsequent analysis, then the decision to remove the data point should be accompanied with an explanation as to why this value is considered an outlier, and what may have caused it to be so different from the majority of the observations.

Example 3.4

The iron content of various foods was measured by G. von Bunge (Bunge 1902). In this experiment, spinach was determined to contain 35 mg of iron per 100 g of spinach. When this value was used by other scientists later on, an error was made. The scientists used the value measured by von Bunge, but failed to notice that it was attributed to dried spinach rather than raw leaves. This error gave rise to a campaign for the health benefits of spinach, as well as a popular comic figure.

The table below lists iron content (in mg) in 100 g of particular foods as reported by the USDA in the national nutrient database for standard reference; USDA (Release 18).

Food	Iron per 100 g (in mg)
Beef, cooked	6.16
Sunflower seeds, roasted, salted	3.81
Chocolate, semisweet	3.13
Tomato paste, canned	2.98
Kidney beans, boiled	2.94
Spinach, raw	2.70
Brussel sprout, cooked	1.20
Soy milk	1.10
Lettuce, raw	1.00
Broccoli, raw	0.91
Cabbage, red, raw	0.80
Raspberries, raw	0.69
Strawberry, raw	0.42
Potato, baked	0.35

Food (von Bunge)	Iron per 100 g (in mg)
Spinach, dried	35.00

If these data values were combined, would the high iron content of dried spinach be considered a statistical outlier? What about the value for beef? Computing the quartiles for the above 15 data points yields $Q_1 = 0.855$ and $Q_3 = 3.055$ and hence an interquartile range of $3.055 - 0.855 = 2.2$. Because the value 35 is larger than $Q_3 + 1.5 \times$ IQR $= 3.055 + 1.5 \times 2.2 = 6.355$, the dried spinach value would be considered an outlier in these data. This makes sense, as all other values are taken from fresh or cooked foods, and the dried spinach is fundamentally different from all other foods listed. The iron value for beef (6.16 mg) is smaller than $Q_3 + 1.5 \times IQR$ and hence would not be considered an outlier.

3.2.4 How to Choose a Descriptive Measure

The mean and median are numerical measures that are used to describe the center of a distribution, or to find a “typical” value, for a given set of observations. If there are some atypical values (outliers) among the observations, then the median is a more reliable measure of center than the mean. On the other hand, if there are no outliers and especially if the number of observations is large, then the mean is the preferred measure of center. If the data are to be used to answer a specific question, then working with means rather than medians will make subsequent statistical analysis much more straightforward.

Variance, standard deviation, range, and interquartile range are all measures that can be used to describe the spread or variability in the observations. If the variability in the data is small, then this means that the observations are all grouped closely together. If, on the other hand, the observations cover a wide range of values, then the variation measure will be large. As is the case for the measure of center, the variance and standard deviation are better suited for situations where there are no extreme outliers among the observations. The range is very susceptible to outliers because it is based entirely on the largest and smallest observation values.

Data characteristic	Statistical measure	When to use
Center	mean median	no outliers, large sample possible outliers
Variability	standard deviation interquartile range range	no outliers, large sample possible outliers use with caution

3.3 Graphical Methods to Display Data

Bar plot: For categorical data, a common graphical display method is a bar plot. The number of observations that fall into each category are counted and displayed as bars. The length of the bars represents the frequency for each category. Because the values of a categorical variable may not be ordinal, the order of the bars (each labeled by the category it represents) can be altered without changing the meaning of the plot.

Example 3.5

Suppose the variable recorded is flower color of a plant. In an experiment, 206 progeny of a certain parental cross of a flowering