Chapter 1

Descriptive Statistics

1.1 Descriptive vs. Inferential

There are two main branches of statistics: descriptive and inferential. Descriptive statistics is used to say something about a set of information that has been collected only. Inferential statistics is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population. Thus, inferential statistics involves generalizing beyond the data, something that descriptive statistics does not do.

Other distinctions are sometimes made between data types.

- Discrete data are whole numbers, and are usually a count of objects. (For instance, one study might count how many pets different families own; it wouldn't make sense to have half a goldfish, would it?)
- Measured data, in contrast to discrete data, are continuous, and thus may take on any real value. (For example, the amount of time a group of children spent watching TV would be measured data, since they could watch any number of hours, even though their watching habits will probably be some multiple of 30 minutes.)
- Numerical data are numbers.
- Categorical data have labels (i.e. words). (For example, a list of the products bought by different families at a grocery store would be categorical data, since it would go something like {milk, eggs, toilet paper, ...}.)

1.2 Means, Medians, and Modes

In everyday life, the word "average" is used in a variety of ways - batting averages, average life expectancies, etc. - but the meaning is similar, usually

the center of a distribution. In the mathematical world, where everything must be precise, we define several ways of finding the center of a set of data:

> **Definition 1: median**.
> The median is the middle number of a set of numbers arranged in numerical order. If the number of values in a set is even, then the median is the sum of the two middle values, divided by 2.

The median is not affected by the magnitude of the extreme (smallest or largest) values. Thus, it is useful because it is not affected by one or two abnormally small or large values, and because it is very simple to calculate. (For example, to obtain a relatively accurate average life of a particular type of lightbulb, you could measure the median life by installing several bulbs and measuring how much time passed before half of them died. Alternatives would probably involve measuring the life of each bulb.)

> **Definition 2: mode**.
> The mode is the most frequent value in a set. A set can have more than one mode; if it has two, it is said to be bimodal.

Example 1:
The mode of {1, 1, 2, 3, 5, 8} is 1.
The modes of {1, 3, 5, 7, 9, 9, 21, 25, 25, 31} are 9 and 25. Thus, the set is bimodal.

The mode is useful when the members of a set are very different - take, for example, the statement "there were more Ds on that test than any other letter grade" (that is, in the set {A, B, C, D, E}, D is the mode). On the other hand, the fact that the mode is absolute (for example, 2.9999 and 3 are considered just as different as 3 and 100 are) can make the mode a poor choice for determing a "center". For example, the mode of the set {1, 2.3, 2.3, 5.14, 5.15, 5.16, 5.17, 5.18, 10.2} is 2.3, even though there are many values that are close to, but not exactly equal to, 5.16.

> **Definition 3: mean**.
> The mean is the sum of all the values in a set, divided by the number of values. The mean of a whole population is usually denoted by μ, while the mean of a sample is usually denoted by $\overline{x}$. (Note that this is the arithmetic mean; there are other means, which will be discussed later.)

Thus, the mean of the set $\{a_1, a_2, \cdots, a_n\}$ is given by

$$\mu = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{1.1}$$

The mean is sensitive to *any* change in value, unlike the median and mode, where a change to an extreme (in the case of a median) or uncommon (in the case of a mode) value usually has no effect.

One disadvantage of the mean is that a small number of extreme values can distort its value. For example, the mean of the set {1, 1, 1, 2, 2, 3, 3, 3, 200} is 24, even though almost all of the members were very small. A variation called the **trimmed mean**, where the smallest and largest quarters of the values are removed before the mean is taken, can solve this problem.

1.3 Variability

Definition 4: range.
The range is the difference between the largest and smallest values of a set.

The range of a set is simple to calculate, but is not very useful because it depends on the extreme values, which may be distorted. An alternative form, similar to the trimmed mean, is the interquartile range, or IQR, which is the range of the set with the smallest and largest quarters removed. If $Q1$ and $Q3$ are the medians of the lower and upper halves of a data set (the values that split the data into quarters, if you will), then the IQR is simply $Q3 - Q1$.

The IQR is useful for determining outliers, or extreme values, such as the element {200} of the set at the end of section 1.2. An outlier is said to be a number more than 1.5 IQRs below $Q1$ or above $Q3$.

Definition 5: variance.
The variance is a measure of how items are dispersed about their mean. The variance σ^2 of a whole population is given by the equation

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma x^2}{n} - \mu^2 \tag{1.2}$$

The variance s^2 of a sample is calculated differently:

$$s^2 = \frac{\Sigma(x-\overline{x})^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)} \tag{1.3}$$

Definition 6: standard deviation.
The standard deviation σ (or s for a sample) is the square root of the variance. (Thus, for a population, the standard deviation is the

square root of the average of the squared deviations from the mean. For a sample, the standard deviation is the square root of the sum of the squared deviations from the mean, divided by the number of samples minus 1. Try saying that five times fast.)

Definition 7: relative variability.
The relative variability of a set is its standard deviation divided by its mean. The relative variability is useful for comparing several variances.

1.4 Linear Transformations

A linear transformation of a data set is one where each element is increased by or multiplied by a constant. This affects the mean, the standard deviation, the IQR, and other important numbers in different ways.

Addition. If a constant c is added to each member of a set, the mean will be c more than it was before the constant was added; the standard deviation and variance will not be affected; and the IQR will not be affected. We will prove these facts below, letting μ and σ be the mean and standard deviation, respectively, before adding c, and μ_t and σ_t be the mean and standard deviation, respectively, after the transformation. Finally, we let the original set be $\{a_1, a_2, \ldots, a_n\}$, so that the transformed set is $\{a_1 + c, a_2 + c, \ldots, a_n + c\}$.

$$\begin{aligned}
\mu_t &= \frac{(a_1 + c) + (a_2 + c) + \cdots + (a_n + c)}{n} = \frac{a_1 + a_2 + \cdots + a_n + n \cdot c}{n} \\
&= \frac{a_1 + a_2 + \cdots + a_n}{n} + \frac{cn}{n} = \mu + c \\
\sigma_t &= \sqrt{\frac{\sum_{i=1}^{n} ((a_i + c) - (\mu + c))^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (a_i - \mu)^2}{n}} = \sigma \\
IQR_t &= Q3_t - Q1_t = (Q3 + c) - (Q1 + c) = Q3 - Q1 = IQR
\end{aligned}$$

where we use the result of the first equation to replace μ_t with $\mu + c$ in the second equation. Since the variance is just the square of the standard deviation, the fact that the standard deviation is not affected means that the variance won't be, either.

Multiplication.

Another type of transformation is multiplication. If each member of a set is multiplied by a constant c, then the mean will be c times its value before the constant was multiplied; the standard deviation will be $|c|$ times its value before

the constant was multiplied; and the IQR will be $|c|$ times its value. Using the same notation as before, we have

$$\begin{aligned}
\mu_t &= \frac{(a_1c)+(a_2c)+\cdots+(a_nc)}{n} = \frac{(a_1+a_2+\cdots+a_n)\cdot c}{n} \\
&= \frac{a_1+a_2+\cdots+a_n}{n}\cdot c = \mu\cdot c \\
\sigma_t &= \sqrt{\frac{\sum_{i=1}^{n}((a_ic)-(\mu c))^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}c^2(a_i-\mu)^2}{n}} = \sqrt{\frac{c^2\sum_{i=1}^{n}(a_i-\mu)^2}{n}} \\
&= \sqrt{c^2\frac{\sum_{i=1}^{n}(a_i-\mu)^2}{n}} = \sqrt{c^2}\cdot\sigma = |c|\sigma \\
IQR_t &= |Q3_t - Q1_t| = |Q3\cdot c - Q1\cdot c| = |c|(Q3-Q1) = |c|IQR
\end{aligned}$$

1.5 Position

There are several ways of measuring the relative position of a specific member of a set. Three are defined below:

Definition 8: simple ranking.
As the name suggests, the simplest form of ranking, where objects are arranged in some order and the rank of an object is its position in the order.

Definition 9: percentile ranking.
The percentile ranking of a specific value is the percent of scores/values that are below it.

Definition 10: z-score.
The z-score of a specific value is the number of standard deviations it is from the mean. Thus, the z-score of a value x is given by the equation

$$z = \frac{x-\mu}{\sigma} \tag{1.4}$$

where μ is the mean and σ is the standard deviation, as usual.

Example 2:
In the set of grade point averages {1.1, 2.34, 2.9, 3.14, 3.29, 3.57, 4.0}, the value 3.57 has the simple ranking of 2 out of 7 and the percentile ranking of $\frac{5}{7} \approx 71.43\%$. The mean is $\frac{20.34}{7} \approx 2.91$ and the standard deviation is 0.88, so the z-score is $\frac{3.57-2.91}{0.88} = 0.75$.

Conversely, if given a z-score, we can find a corresponding value, using the equation

$$x = \mu + z\sigma \tag{1.5}$$

Example 3:
The citizens of Utopia work an average (mean) of 251 days per year, with a standard deviation of 20 days. How many days correspond to a z-score of 2.3?

Since each z corresponds to one standard deviation, a z-score of 2.3 means that the desired value is 2.3 standard deviations more than the mean, or $251 + 2.3 \cdot 20 = 297$.

1.6 Dispersion Percentages

> **Theorem 1: empirical rule**
> For data with a "bell-shaped" graph, about 68% of the values lie within one standard deviation of the mean, about 95% lie withing two standard deviations, and over 99% lie within three standard deviations of the mean.

Note that since 99% of the data fall within a span of six standard deviations (z-scores of -3 to +3), the standard deviation of a set of values that are somewhat bell-shaped should be about $\frac{1}{6}$ of the range. This can be useful in checking for arithmetic errors.

> **Theorem 2: Chebyshev's Theorem**
> For any set of data, at least $1 - \frac{1}{k^2}$ of the values lie within k standard deviations of the mean (that is, have z-scores between $-k$ and $+k$).

Example 4:
Matt reads at an average (mean) rate of 20.6 pages per hour, with a standard deviation of 3.2. What percent of the time will he read between 15 and 26.2 pages per hour?

15 pages per hour corresponds to a z-score of $\frac{15-20.6}{3.2} = -1.75$, and 26.2 pages per hour corresponds to a z-score of $\frac{26.2-20.6}{3.2} = 1.75$. Chebyshev's Theorem says that $1 - \frac{1}{1.75^2} = .673$ of the values will be within 1.75 standard deviations, so 67.3% of the time, Matt's reading speed will be between 15 and 26.2 pages per hour.
