4. DESCRIPTIVE STATISTICS

Descriptive Statistics is a body of techniques for summarizing and presenting the essential information in a data set.

Eg: Here are daily high temperatures for Jan 16, in 30 U.S. cities:


Clearly, a long list of numbers is not very informative.
A better presentation is the **Frequency Distribution**: Group the data into intervals called **classes**, and record the frequency (i.e., the number of observations) for each class.

**Eg:** Daily High Temperatures:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>1</td>
</tr>
<tr>
<td>30-39</td>
<td>7</td>
</tr>
<tr>
<td>40-49</td>
<td>8</td>
</tr>
<tr>
<td>50-59</td>
<td>7</td>
</tr>
<tr>
<td>60-69</td>
<td>2</td>
</tr>
<tr>
<td>70-79</td>
<td>3</td>
</tr>
<tr>
<td>80-89</td>
<td>2</td>
</tr>
</tbody>
</table>

**Histograms** provide a graphical representation of a frequency distribution.

![Histogram of Daily High Temperatures](image)
Measures of Central Tendency (Location)

Needed: An objective, concise summary of a data set.

For many purposes, just two numbers will suffice:

(1) A measure of central tendency (i.e., the typical value, or location),
(2) A measure of dispersion (fluctuation).

Here, we discuss measures of central tendency. The two most popular measures are:

The **Mean**; The **Median**.

Of these, the **mean** is the most important.
Sample Mean

The mean of a sample of \( n \) measurements \( x_1, \ldots, x_n \) is

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + \cdots + x_n)
\]

**Eg:** The first five high temperatures in our data set are 35, 57, 68, 43, 40. The sample size is \( n=5 \), and the sample mean is

\[
\bar{x} = \frac{1}{5} (35 + 57 + 68 + 43 + 40) = \frac{1}{5} (243) = 48.6
\]

- \( \bar{x} \) is the average value. It is also the center of gravity (balance point) of the data set.
Population Mean

The mean of a population of \( N \) measurements \( x_1, \ldots, x_N \) is

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} (x_1 + \cdots + x_N)
\]

**Eg:** If we view our data set of 30 high temperatures as a population, the population mean is

\[
\mu = \frac{1}{30} \sum_{i=1}^{30} x_i = \frac{1}{30} (35+57+\cdots+55) = 50.87
\]

- \( \bar{x} \) is a statistic, \( \mu \) is a parameter.
• Often, we would like to know the population mean $\mu$, but our information is limited to a small random sample. Then we can use $\bar{x}$ as an estimate of $\mu$. Using principles of statistical inference, we can even assess the accuracy of this procedure, and thereby draw conclusions ("make inferences") about $\mu$.

• **The Problem With $\bar{x}$**: It is extremely sensitive to outliers ("extreme observations", or "wild values").

These outliers may be due to errors in recording the data, or they may be real (but exceptional) observations. In either case, it is usually best to set aside the outliers (to be described separately) before computing $\bar{x}$. Alternatively, use the median.
Median

Given \( n \) measurements arranged in order of magnitude,

Median = The Middle Value if \( n \) is odd
Median = The average of the two middle values if \( n \) is even.

Eg: Top 5 CEO Compensations for US Companies ($Millions)

<table>
<thead>
<tr>
<th>Company</th>
<th>Compensation ($Millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla</td>
<td>595</td>
</tr>
<tr>
<td>Apple</td>
<td>134</td>
</tr>
<tr>
<td>Charter Comm.</td>
<td>117</td>
</tr>
<tr>
<td>ViacomCBS</td>
<td>117</td>
</tr>
<tr>
<td>Chewy</td>
<td>108</td>
</tr>
</tbody>
</table>
Arranging the data in order gives: 108, 117, 117, 134, 595.

The median compensation is $117,000,000.
The mean compensation is $214,200,000.
The mean is substantially larger than the median due to the outlier, Tesla.

If we remove the outlier, then $\bar{x}$ becomes $119,000,000$, while the sample median remains at $117,000,000$. 
• The median divides a data set into two equal parts. Half of the data lie below the median, and half lie above it.

• The median is resistant to outliers. Thus it can be safely used on a raw, unexamined data set. (Of course, it is always best to look at the data; you will usually learn something.)

• Although the median is very useful as a descriptive statistic, it is rarely used for statistical inference. Reason: No simple mathematical theory for the median.
Simple Measures of Dispersion

The mean (or median) cannot completely summarize a data set. Once we know the typical value, the next question is:

To what extent do the data fluctuate from their typical value?

Eg: Consider the lifetimes of “GE” and “Philips” light bulbs. Both brands are rated for 750 hours (average lifetime).
"GE" and "Philips" Lightbulb Lifetimes (in hours)

The GE bulbs exhibit better **quality control**:
performance is consistent, since there is not much variation.

Philips’ performance is more erratic: There’s more fluctuation, although the **average** is the same as for GE.
The **range** is the difference between the largest and smallest values.

**Eg:** For the baseball salaries, the highest and lowest values were $42,142,857 (for Max Scherzer) and $555,000 (for the 41 lowest-paid players), so the range was

$$42,142,857 - 555,000 = 41,587,857$$

The range is a very crude measure, containing no information about the dispersion of the values between the extremes.

It has absolutely no resistance to outliers. (Why?)
A resistant measure of dispersion is provided by the **interquartile range**.

\[ \text{IQR} = Q_U - Q_L = 75^{\text{th}} \text{ Percentile} - 25^{\text{th}} \text{ Percentile} \]

**Definition of Quartiles:**

The first (or lower) quartile \( Q_L = 25^{\text{th}} \text{ percentile} = \) 
Value such that 25% of the distribution is below it.

The second quartile \( = 50^{\text{th}} \text{ percentile} = \text{Median}. \)

The third (or upper) quartile \( Q_U = 75^{\text{th}} \text{ percentile} = \) 
Value such that 75% of the distribution is below it.
• IQR is the width of the “middle half” of the data set. The IQR is resistant to outliers.

Eg: For the baseball data, the 25th percentile is $567,500, the 75th percentile is $6,000,000, so the IQR is

$$\$6,000,000 - \$567,500 = \$5,432,500$$

• Using the extremes, quartiles and median, we can draw a boxplot, a graphical summary which reveals basic distributional properties (Center, spread, skewness, outliers), and which is especially useful for comparing several data sets, side-by-side.
Five Year Performance of Mutual Funds
Based on a $2000/yr investment
For each data set, we draw a box extending from $Q_L$ (bottom) to $Q_U$ (top). We draw a horizontal line at the median. Then, we draw two vertical “whiskers” from the box to the most extreme non-outlying observations. Any data value beyond the whiskers is declared to be an outlier and flagged with an asterisk or circle.

The height of the box is the IQR.

For symmetric distributions, median will be halfway between $Q_L$ and $Q_U$. Otherwise, the distribution is skewed.

The width of the box doesn’t mean anything!
How outliers are labeled:

“Suspect outliers”, labeled with an asterisk, are those more than $1.5 \times \text{IQR}$ above $Q_U$ or below $Q_L$.

“Highly suspect outliers”, labeled with a circle, are those more than $3 \times \text{IQR}$ above $Q_U$ or below $Q_L$.
Note: Boxplots can hide bimodality.

Eg: Old Faithful eruptions.
Distribution Shape

A distribution may be symmetric or skewed, it may be unimodal, bimodal or multi-modal, it may be long-tailed (lots of outliers) or short-tailed (almost no outliers). Histograms and boxplots help us to see the distribution shape.

1) Symmetrical: Roughly equal tails.
   **Eg:** Bell-Shaped Distribution.

2) Positively Skewed (skewed to the right): Longer tail on right.
   **Eg:** Income Distributions.

3) Negatively Skewed (skewed to the left): Longer tail on left.
   **Eg:** Scores on an easy exam.
(a) Bell-shaped distribution

(b) Left-skewed distribution

(c) Right-skewed distribution
• For nearly symmetrical distributions, mean \approx \text{median}, (Also, Median is about halfway between 25\textsuperscript{th} and 75\textsuperscript{th} percentiles)

\textbf{Eg:} A sample of student heights.

\textbf{Descriptive Statistics}

\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
\textbf{Variable: height} & \\
\hline
\textbf{Andeerson-Darling Normality Test} & \\
A-Squared: & 0.201 & \\
P-Value: & 0.874 & \\
\hline
\textbf{Mean} & 65.2900 & \\
\textbf{StDev} & 1.8994 & \\
\textbf{Variance} & 3.60786 & \\
\textbf{Skewness} & -1.3E-02 & \\
\textbf{Kurtosis} & -5.4E-01 & \\
\textbf{N} & 50 & \\
\hline
\textbf{Minimum} & 61.5000 & \\
\textbf{1st Quartile} & 63.9750 & \\
\textbf{Median} & 65.3000 & \\
\textbf{3rd Quartile} & 66.6000 & \\
\textbf{Maximum} & 69.5000 & \\
\textbf{68\% Confidence Interval for Mu} & 65.0201 & 65.5599 & \\
\textbf{68\% Confidence Interval for Sigma} & 1.7344 & 2.1232 & \\
\textbf{68\% Confidence Interval for Median} & 65.0000 & 65.5014 & \\
\hline
\end{tabular}
For positively skewed distributions, mean > median (Also, Median is closer to the 25th percentile than to 75th)

Eg: Salaries of the 2019 Baseball players.
For negatively skewed distributions, mean < median (Also, Median is closer to the 75th percentile than to 25th).

Eg: Class scores on an individual project

Descriptive Statistics

Anderson-Darling Normality Test
A-Squared: 1.610
P-Value: 0.000

Mean 92.6071
StDev 6.7404
Variance 45.4325
Skewness -1.59939
Kurtosis 2.41526
N 28

Minimum 72.000
1st Quartile 91.000
Median 95.000
3rd Quartile 96.750
Maximum 100.000

95% Confidence Interval for Mu
89.994 95.221

95% Confidence Interval for Sigma
5.329 9.175

95% Confidence Interval for Median
92.448 96.000