1 Statistics

The course thus far has concerned itself primarily with PROBABILITY THEORY. In studying probability, we have made the *assumption* that we fully knew or understood the parameters of the underlying distribution.

For example,
- $X \sim N(\mu = 15, \sigma^2 = 4)$
- $Y \sim \text{Poisson}(\lambda = 2)$
- $W \sim \text{Uniform}(0, 2)$

In a sense, we have left nothing to the imagination, that is, we have *fully specified* the distribution parameters for our Random Variables. Now, we may (and shall) directly compute P(X<13), P(Y=1), or P(.25≤ W ≤1.75).

$P(X < 13) = P\left(\frac{x-\mu}{\sigma} < \frac{13-15}{2}\right) = P(z < -1) = \Phi(-1) \approx 0.1587$

$P(Y = 1) = \frac{2^1}{1!}e^{-2} \approx 0.271$

$P(.25 < W < 1.75) = \int_{.25}^{1.75} \frac{1}{2}\, dx = .75$

.750

Is this the case in general? No! How could we possibly expect to see this every time, in every situation? Consider the situation in which you know nothing about the underlying population!

In STATISTICS, we will use *observed* data to compute probabilities, related quantities of interest, or to make decisions and predictions.

1.1 Terminology

To get us started, here are some key terms essential to understanding the "big picture."

POPULATION The total set of observations or individuals that we may describe or attempt to draw conclusions about. This set is what you describe logically.

SAMPLE A subset which is collected or observed from the population. i.e. what you see yourself.

EXAMPLE: Suppose we are interested in the damage sustained by model year 2003 BMW M5's when crashed at 85 miles per hour.

- The population consists all 2003 BMW M5. Of which there are 1,710.
- Collecting information for the entire population would require collecting each car, and crashing it at 85mph.

$\Longrightarrow$ Constraints on money, time, practicality, or scarce resources make obtaining information on the entire population. It is always important to make note of what is the population, and what is the sample.

VARIABLE A characteristic whose value may chance from one member (or observation) to another in the population.

EXAMPLE: Suppose we are interested in making generalizations about about the redwood trees in California.
- The population consists all redwood trees that exist now, have ever existed, or ever will exist.
- A possible *sample* may consist of a random selection of 10 redwood trees from each of the national parks (Yosemite, Sequoia, and Kings Canyon) in the Serra Nevada mountains.

UNIVARIATE DATA consists of observations made on a single variable. Think list of observations for which you may (and shall) describe with descriptive statistics. This class focuses on univariate probability and statistics.

MULTIVARIATE DATA consists of observations made on multiple variables simultaneously. Think giant spread sheets, numbers become vectors and matrices. Ask yourself: what kind of data do you think Facebook collects? What kind of data do you think your insurance company uses when they charge you for insurance?

2 Numerical Summaries of Univariate Data

When we collect data we are often interested in concise ways to communicate features of the data. As we have already seen, we typically discuss the theoretical mean and variance. Now, in dealing with sample data, we wish to convey sample statistics as well.

Suppose we have a data set consisting of n-many observations on one variable. The individual observations themselves are denoted by $x_1, x_2, \ldots, x_n$ where each x_i is a number.

2.1 Measures of Location or Central Tendency

As with population models, we are concerned with ways to describe *where* most of the data may occur.

1. **The Sample Mean** of the observations $x_1, x_2, \ldots, x_n$ is denoted with $\bar{x}$ as is computed as the arithmetic average of the observations

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

2. **The Sample Median** of the observations, denoted $\tilde{x}$ is simply the middle value of the observations after ordering from smallest to largest.

$$\tilde{x} = \begin{cases} (\frac{n+1}{2})\text{th ordered value} & \text{if } n \text{ is odd} \\ \text{average of } \frac{n}{2}\text{th and the } (\frac{n}{2}+1)\text{th ordered values} & \text{if } n \text{ is even} \end{cases}$$

EXAMPLE: It is worth noting that the sample mean $\bar{x}$ is sensitive to extreme values. You have likely seen this where one poor midterm destroys your otherwise strong grade in the class.

- Suppose your first 4 test scores are 90, 90, 96, and 96. The sample average: $\bar{x} = \frac{90+90+96+96}{4} = 93$
- And now, you took the final and earned a score of 10; what would that do to your average? $\bar{x} = \frac{90+90+96+96+10}{5} = 76.4$
- the sample median on the other hand, moves from 93 to 90.

3. **The Sample Quartiles** divide the the data set into 4 equal parts.
   - **The First Sample Quartile** denoted Q_1 is the value *of the data* such that at most $\frac{1}{4}$ of the observations are smaller than Q_1 and at most $\frac{3}{4}$ are larger than Q_1.
   - **The Second Sample Quartile** denoted Q_2 is the sample median.
   - **The Third Sample Quartile** denoted Q_3 is the value *of the data* such that at most $\frac{3}{4}$ of the observations are smaller than Q_3 and at most $\frac{1}{4}$ are larger than Q_3.

EXAMPLE Suppose you have the collected the following data on gas prices in Santa Clara County (as of November 1, 2015).

$$2.41,\ 2.59,\ 2.65,\ 2.71,\ 2.79,\ 2.85,\ 2.89,\ 3.09,\ 3.29,\ 3.79$$

Describe the Sample Quartiles:
at most $10/4 = 2.5$ observations are less than $Q_1 \implies Q_1 = 2.65$
the sample median is middle value, but n=10, so its the average of the two middle values... $\frac{2.79+2.85}{2} = 2.82$
at most $3\text{x}10/4 = 7.5$ observations are less than $Q_3 \implies Q_3 = 3.09$

2.2 Measures of Spread or Variability

Again, as in population models, we are concerned with describing *how much* spread or variability there is in the sample.

1. **The Sample Range** is perhaps the most simple measure of variability, given as the difference between the largest and smallest sample values.
2. **The Interquartile Range** usually denoted as the IQR, or the Fourth Spread is given by

$$IQR = Q_3 - Q_1$$

3. **The Sample Variance** is commonly denoted as s^2 is given by

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

Proof.

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

PROPOSITION let $x_1, x_2, \ldots, x_n$ be a sample with sample variance s_x^2

1. Let $y_1 = x_1 + a, y_2 = x_2 + a, \ldots, y_n = x_n + a$, where $a \in \mathbf{R}$, and let s_y^2 denote the sample variance of the y_i. Then $s_x^2 = s_y^2$.
2. let $z_1 = bx_1, z_2 = bx_2, \ldots, z_n = bx_n$ where $b \neq 0 \in \mathbf{R}$. Then $s_z^2 = b^2 s_x^2$

Proof.
1.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i + a) - (\bar{x} + a))^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$$

2.

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(bx_i - b\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} b^2(x_i - \bar{x})^2 = b^2 s_x^2$$

EXAMPLE Describe the variability in the data from the gasoline prices in Santa Clara County.
summary:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.410   2.665   2.820   2.906   3.040   3.790
```

sd(gas) = 0.3985306
mean(gas) = 2.906
range = max(gas) - min(gas) = 3.79-2.41=1.38
IQR(gas) = 0.375

2.3 Boxplots

A boxplot is a powerful, yet simple visual summary of data. The most simple boxplot conveys the following 5-number summary: (1) the minimum, (2) first sample quartile, (3) the sample median, (4) the third sample quartile, and (5) the maximum. The position of the median conveys information regarding skewness in the middle 50% of the data.

To draw a boxplot:

1. Step 1. Draw a measurement scale horizontally and place a rectangle above this axis where the left-edge is at Q_1 and the right edge is at Q_3. $\implies$ The width of the rectangle is given by the IQR= $Q_3 - Q_1$.
2. Step 2. Place a verticle line segment inside the rectangle at the location of the sample median.
3. Step 3. Draw the whiskers out from both ends of the rectangle to the respective minimum and maximum observations.

EXAMPLE For our gasoline data the 5 number summary is as follows:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.410   2.665   2.820   2.906   3.040   3.790
```

Concluding our discussion on simple descriptive statistics, we look at OUTLIERS, observations that are found unusually far from the bulk of the data. So now the question becomes: *How far is far?*

- We say that a univariate observation x_i is a **mild** outlier if $x_i < Q_1 - 1.5IQR$ or $x_i > Q_3 + 1.5IQR$
- If x_i is an outlier if $x_i < Q_1 - 3IQR$ or $x_i > Q_3 + 3IQR$ then that observation is considered an **extreme** outlier.

Referring to the boxplot above and determine the cutoff values for mild and extreme outliers. The sample IQR is 0.375
$Q_1 = 2.665$, $Q_3 = 3.040$
Mild Lower: $Q_1 - 1.5 \times IQR = 2.665 - 1.5 \times .375 = 2.0875$
Mild Upper: $Q_3 + 1.5 \times IQR = 3.040 + 1.5 \times .375 = 3.6025$
Extreme Lower: $Q_1 - 3 \times IQR = 2.665 - 3 \times .375 = 1.525$
Extreme Upper: $Q_3 + 3 \times IQR = 3.040 + 3 \times .375 = 4.165$