Chapter 6

Def'n: Statistics:

1) are commonly known as numerical facts.
2) is a field of discipline or study.

In this class, statistics is the science of collecting, analyzing, and drawing conclusions from data. The methods help describe and understand variability.

Types of statistics:

- Descriptive: methods to view a given dataset.
  →
- Inferential: methods using sample results to infer conclusions about a larger pop'n.
  →

Def'n: A variable is any characteristic that is recorded for subjects in a study.

- Categorical (qualitative): cannot assume a numerical value but classifiable into 2 or more non-numeric categories. →
- Numerical (quantitative): measured numerically.
  - Discrete: only certain values with no intermediate values. →
  - Continuous: any numerical value over a certain interval or intervals.
    →

Population vs. Sample:

Def'n: A population consists of all elements whose characteristics are being studied.
Ex6.1)
A sample is a portion of the population selected for study.
Ex6.2)

6.2 Data Presentation

Graphical Summaries for Categorical Data

Def'n: A bar chart is a graph of bars whose heights represent the (relative) frequencies of respective categories. Ex6.3) (created in class)

Look for: frequently and infrequently occurring categories.

A Pareto chart arranges the categories in order of decreasing frequency
Ex6.4) (created in class)

Look for: frequently and infrequently occurring categories.

A pie chart is a circle divided into portions that represent (relative) frequency belonging to different categories. Ex6.5) (created in class)

Look for: categories that form large and small proportions of the data set.

Graphical Summaries for Numerical Data

Def'n: A frequency distribution (for numerical data) is a listing of non-overlapping intervals, together with the # of observations for each interval (a.k.a. class or bin).

$$Relative\ frequency = \frac{f}{\sum f}$$ (Cumulative relative frequency also exists)

The data divide into intervals (normally of equal width).

Worldwide Box Office (in millions)	Number of movies *f*	Relative Frequency	Cumulative rel. freq.
200 to 599			
600 to 999			
1000 to 1399			
1400 to 1799			
1800 to 2199			
2200 to 2599			
2600 to 3000			

Def'n: A histogram graphically shows a frequency distribution for *numerical* data.

Look for: - central or typical value and corresponding spread
- gaps in the data or outliers
- presence of symmetry in the dist'n
- number and location of peaks

An outlier is an obs'n that falls well above or below the overall bulk of the data.

Ex6.6) (preceding table used for example in class)

Histogram traits: (corresponding curves drawn in class)

1. Modes (unimodal, bimodal, multimodal)
2. Skewness (symmetric, left-skewed & right-skewed) → term refers to "TAIL"
3. Tail weight (normal, heavy-tailed, light-tailed)

6.3 Sample Statistics (and more)

Def'n: A parameter is a summary measure calculated for pop'n data.

A statistic is a summary measure calculated for sample data.

Measures of Center:

Def'n: If *n* observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

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

Also, if there are *N* obs'ns in an entire population, then the population mean is

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The median is the value of the midpoint of a data set that has been ranked in order, increasing or decreasing. If dataset has an even # of observations, use the average of the middle 2 values.

The mode is the most frequent value in a data set.

NOTE: median (and mode) resistant to outliers, mean uses all observations.

Table 6X0 – Estimated provincial populations circa Jul. 2011 (in millions)

ON	PQ	BC	AB	MB	SK	NS	NB	NL	PEI
13.373	7.980	4.573	3.779	1.250	1.058	0.945	0.756	0.511	0.146

Ex6.7) Avg. pop'n of all provinces:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} =$$

median =

mode = no mode here

Avg. pop'n from sample of 3 provinces:

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

Outlier effect? (remove Ontario & Quebec)

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

Comparing Mean, Median & Mode in Histograms: (graphs drawn in class)

1. Symmetric curve & histogram
   - all 3 identical and lie at the center of the distribution
2. Right-skewed: Mode < Median < Mean

3. Left-skewed: Mean < Median < Mode

Measures of Spread:

Def'n: The <u>sample range</u> is *range* = max(x_i) – min(x_i)

Ex6.8) (from Table 6X0) *range* =

Deviations from the Mean:

Ex6.9) 1, 2, 4, 3

x_i	$x_i - \bar{x}$
1	1 – 2.5 = -1.5
2	2 – 2.5 = -0.5
4	4 – 2.5 = 1.5
3	3 – 2.5 = 0.5
	$\sum (x_i - \bar{x}) = 0$

Note that $\sum_{i=1}^{N}(x_i - \mu)$ and $\sum_{i=1}^{n}(x_i - \bar{x})$, aka deviation of *x* from the mean, both equal zero.

Variance and Standard Deviation:

The most common measure of spread is standard deviation. Variance, however, must be calculated first. The basic formulas for variance are:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where σ^2 is the population variance and s^2 the sample variance.

Since $\sum (x_i - \bar{x})^2 = \sum x_i^{\,2} - \frac{(\sum x_i)^2}{n}$, the variance formulas become

$$\sigma^2 = \frac{1}{N}\left[\sum x_i^{\,2} - \frac{(\sum x_i)^2}{N}\right] \quad \text{and} \quad s^2 = \frac{1}{n-1}\left[\sum x_i^{\,2} - \frac{(\sum x_i)^2}{n}\right]$$

Finding the standard deviation only requires taking the square root of the variance.
Population: $\sigma = \sqrt{\sigma^2}$ Sample: $s = \sqrt{s^2}$

Ex6.10) 1, 2, 4, 3

$$\sum x_i = 10,\ \sum x_i^2 = 30,\ \sigma^2 = \frac{1}{N}\left[\sum x_i^{\,2} - \frac{(\sum x_i)^2}{N}\right] =$$

$\sigma =$

If the data was a sample, $s^2 =$, $s =$

Important notes:

1. Standard deviation measures spread *only* about the mean (i.e. not the median).
2. Values of variance and std. dev. are never negative. (Equals zero only if *no spread.*)
3. The measurement units of variance are always the square of the units of the original data.
4. Standard deviation, like the mean, is not resistant to outliers.
5. Consider the sample variance s^2 to have n – 1 degrees of freedom. There are n observations, and n deviations from the mean. Since the total always sums to zero, n – 1 of these quantities determines the remaining one. Thus, only n – 1 of the n deviations, $x_i - \bar{x}$, are freely determined. (Degrees of freedom apply only to samples.)

Boxplots:

Def'n: The p^{th} quantile is a value such that p percent of the observations fall below or at that value. Three useful quantiles are quartiles. The *lower* (or *first*) *quartile* has $p = 25$, the median has $p = 50$, and the *upper* (or *third*) *quartile* has $p = 75$.

The five-number summary consists of the min, Q_1, median, Q_3, and the max.

(visual representation of above drawn in class)

Def'n: The interquartile range (IQR) is the difference between the first and third quartiles. IQR = $Q_3 - Q_1$ (IQR is *also* a measure of spread)

Ex6.11) 7.9 9.1 9.2 9.3 9.4 9.4 9.5 9.6 9.6 9.7

(examples regarding finding quartiles with other #'s of observations discussed in class)

Note: For exams, INCLUDE the median in calculating Q_1 and Q_3.

Def'n: A boxplot shows the center, spread, and skewness of a data set.
To construct it:
Step 1: Rank the data in increasing order and find the median, Q_1, Q_3, and IQR.
Use data from Ex6.11) to construct a boxplot.
Step 2: Find the points beyond the boundaries: 1.5*IQR below Q_1 and 1.5*IQR above Q_3, known as the lower & upper inner fences, respectively. These points are outliers.
1.5*IQR =
Lower i.f. = Upper i.f. =
Step 3: Determine smallest & largest values within the respective inner fences.
small = large =
Step 4: Draw linear scale containing entire range of data.
Step 5: Draw perpendicular lines to the scale to indicate Q_1 and Q_3. Connect ends of both lines. Box width = IQR
Step 6: Draw another line perpendicular to the scale to indicate the median inside the box.
Step 7: Draw two smaller lines perpendicular to the scale for the values from Step 3. Join their centers to the box to make whiskers.

(boxplot drawn in class)

What to do with outliers?
Consider lower & upper outer fences at 3.0*IQR below Q_1 and 3.0*IQR above Q_3.
Ex6.12) 3.0*IQR =
Lower o.f. = Upper o.f. =

A *mild outlier* is outside an inner fence but inside the outer fence.
An *extreme outlier* is outside either outer fence.
Some textbooks represent mild outliers by shaded circles, extreme outliers by open circles. Our textbook uses asterisks, '*', as well. Overall, classifying outliers is important whereas drawing them a certain way is subjective.

Whiskers extend on each end to the most extreme obs'ns that are *NOT* outliers.
Ex6.13)

Analyzing center, spread, and skewness:
- What is the approximate value of the center?
- What is the width of the IQR?
- Are the data symmetric or skewed?

Boxplot vs. Histogram:
Each graph highlights different features of a data set (layers of skewness and skewness/modality, respectively), so it's always better to construct both.