

Chapter 1

Def'n: Statistics:

- 1) are commonly known as numerical facts
 - 2) is a field of discipline or study
- Here, statistics is about variation.

3 main aspects of statistics:

- 1) Design ("Think"): Planning how to obtain data to answer questions.
- 2) Description ("Show"): Summarizing the obtained data.
- 3) Inference ("Tell"): Making decisions and predictions based on data.

Def'n: A population consists of all elements whose characteristics are being studied.

Ex1.1)

A sample is a portion of the population selected for study.

Ex1.2)

A parameter is a summary measure calculated for population data.

A statistic is a summary measure calculated for sample data.

Types of statistics:

Descriptive: methods to view a given dataset.

→

Inferential: methods using sample results to infer conclusions about a larger population.

→

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

- Qualitative (categorical): cannot assume a numerical value but classifiable into 2 or more non-numeric categories. →

- Quantitative (numerical): measured numerically.

- Discrete: only certain values with no intermediate values. →

- Continuous: any numerical value over a certain interval or intervals.

→

Chapter 2 – Categorical Data Graphs

Def'n: A frequency table (for qualitative data) is a listing of possible values for a variable, together with the # of observations for each value.

$$\text{Relative frequency} = \frac{f}{\sum f}$$

$$\text{Percentage} = \text{Relative frequency} \times 100\%$$

Table 2X0

Faculty	M	F	Frequency (f)	Relative frequency	Percentage (%)
Science					
Arts					
Business					
Phys. Ed. & Rec.					
Other					

Ex2.1) In percentage, how many students are in science? How many students are female?
How many students are female and in science? How many science students are female?
How many female students are in science?

Graphical Summaries

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

Look for: frequently and infrequently occurring categories.

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

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

A segmented bar chart uses a rectangular bar divided into segments that represent frequency or relative freq. of different categories.

Ex2.2) Draw appropriate graphs for data from Table 2X0.

Chapter 3 – Numerical Variable Graphs

Def'n: A stem-and-leaf display has each value divided into two portions: a stem and a leaf. The leaves for each stem are shown separately. (Values should be ranked.)

Look for:

- typical values and corresponding spread
- gaps in the data or outliers
- presence of symmetry in the distribution
- number and location of peaks

Ex3.1)

Note: *Dotplots* also exist (see p. 48 in textbook), but “replace” the values with dots.

Def'n: A histogram, like a bar graph, graphically shows a frequency distribution. The data here, however, is quantitative.

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

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

Cumulative Relative Frequency = (Cumul. freq. of a class) / (Total obs'ns in dataset)

Table 3X0 – Total earnings as of Sep. 1/2017

Worldwide Box Office (in millions)	Number of movies <i>f</i>	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			

Ex3.2)

NOTE: Dot and S-and-L plots are good for small data sets because data values are retained. Histograms are better for large data sets to condense the data.

Histogram shapes/traits: (corresponding figures drawn in class)

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

Def'n: A timeplot is a graph of data collected over time (or a *time series*).

Look for: - a *trend* over time, denoting a decrease or increase.

- a pattern repeating at regular intervals (a *cycle* or *seasonal variation*)

Ex3.3)

Chapters 3/4 – Summary measures (and one more graph)

Measures of Center

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

$$\text{Population mean: } \mu = \frac{\sum y_i}{N} \quad \text{Sample mean: } \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum y_i}{n}$$

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.

Note: median resistant to outliers, mean uses all observations.

Table 4X0 – Estimated provincial populations circa 2016 (in millions)

ON	QC	BC	AB	MB	SK	NS	NB	NL	PEI
13.983	8.326	4.752	4.253	1.318	1.151	0.950	0.757	0.530	0.149

Ex4.1)

Mathematical Characteristics of the Mean

Adding or subtracting a constant to all scores will alter the value of the mean by the value of the constant.

$$\frac{\sum (y_i + c)}{n} = \frac{\sum y_i + \sum c}{n} = \frac{\sum y_i}{n} + \frac{\sum c}{n} = \bar{y} + \frac{nc}{n} = \bar{y} + c$$

Multiplying or dividing all scores by a constant will alter the value of the mean by the value of the constant.

$$\frac{\sum (cy_i)}{n} = \frac{cy_1 + cy_2 + \dots + cy_n}{n} = \frac{c(y_1 + y_2 + \dots + y_n)}{n} = c \left(\frac{\sum y_i}{n} \right) = c\bar{y}$$

Ex4.2) Use data from Ex4.1 to see these work with the first two averages by a) adding 1 million to each province, and b) multiplying each province population by 1.05.

a) Avg. pop'n of all provinces:

Avg. pop'n from sample of 3 provinces:

b) Avg. pop'n of all provinces:

Avg. pop'n from sample of 3 provinces:

Comparing Mean and Median: (corresponding figures drawn in class)

1. Symmetric curve & histogram

2. Right-skewed: Median < Mean

3. Left-skewed: Mean < Median

Def'n: The mode is the most frequent value in a data set.

Ex4.3) Provinces →

Movies →

Measures of Spread

Def'n: Range = largest value – smallest value = max – min

Ex4.4) (from Table 4X0) range =

Deviations from the Mean:

Ex4.5) 1, 2, 4, 3

y_i	$y_i - \bar{y}$
1	$1 - 2.5 =$
2	$2 - 2.5 =$
4	$4 - 2.5 =$
3	$3 - 2.5 =$
	$\sum (y_i - \bar{y}) =$

Note that $\sum (y_i - \mu)$ and $\sum (y_i - \bar{y})$, or deviation of x from the mean, both equal zero.

Variance and Standard Deviation:

The most common measure of spread is standard deviation. Informally interpreted as the size of a “typical” deviation from the mean. Variance, however, must be calculated first.

The basic formulas for variance are:

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N} \quad s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

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

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

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

Finding the standard deviation only requires taking the *positive* square root of the variance.

Population: $\sigma = \sqrt{\sigma^2}$ Sample: $s = \sqrt{s^2}$

Ex4.6) 1, 2, 4, 3

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, $y_i - \bar{y}$, are freely determined. (Degrees of freedom apply only to samples.)

Mathematical Characteristics of the Standard Deviation

Adding or subtracting a constant to all scores will NOT alter the standard deviation.

Multiplying or dividing all scores by a constant will alter the standard deviation by the value of the constant.

Ex4.7) Compute standard deviation (population & sample) for 10, 20, 40, 30.

Measures of Position

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

Note: For odd n , EXCLUDE the median in each half when calculating Q_1 and Q_3 .

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

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

Ex4.8) 7.9 9.1 9.2 9.3 9.4 9.4 9.5 9.6 9.6 9.7

Boxplots:

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.

Step 2: Find the points beyond the boundaries: $1.5 \cdot \text{IQR}$ below Q_1 and $1.5 \cdot \text{IQR}$ above Q_3 , known as the lower & upper inner fences, respectively. These points are outliers.

Ex4.9) $1.5 \cdot \text{IQR} =$

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.

What to do with outliers?

Consider lower & upper outer fences at $3.0 \cdot \text{IQR}$ below Q_1 and $3.0 \cdot \text{IQR}$ above Q_3 .

Ex4.10) $3.0 \cdot \text{IQR} =$

Lower o.f. =

Upper o.f. =

A (*mild*) *outlier* is outside an inner fence but inside the outer fence.

A *far (or extreme) outlier* is outside either outer fence.

All textbooks are different for distinguishing outliers. Our textbook uses open circles for mild and asterisks, ‘*’, for far outliers. Overall, classifying outliers is important whereas drawing them a certain way is subjective.

Whiskers extend on each end to the most extreme observations that are *not* outliers.

Ex4.11)

Looking at center, spread, and skewness:

Approx. value of the center? Width of IQR? 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.

Chapter 5 – Standard Deviation as a Ruler

Empirical Rule applies only to a bell-shaped distribution.

1. About 68% of observations lie within 1σ of the mean.
2. About 95% of observations lie within 2σ of the mean.
3. About 99.7% of observations lie within 3σ of the mean.

Suppose we go further..., say, 6σ . Software produces a value of 99.999 999 803%, which means far less chance for “error” (the observations beyond 6σ from the mean).

Extra Measure of Position/Potential Outlier Identifier

$z\text{-score} = (\text{observation} - \text{mean}) / (\text{std. dev.})$

- $z\text{-score}$ tells us how many standard deviations the value is from the mean, positive OR negative
- more useful when distribution approximately normal.
- a potential outlier is more than 3σ from the mean.

Ex5.1) $\mu = 31.6$, $\sigma = 26.4$, $y = 50$