Chapter Three

Central Tendency and Variation

Central Tendency

Variation

Mean and Standard Deviation of Grouped Data

Percentiles and Quartiles

Measures of Central Tendency

Measures of central tendency show what data in a data set tend to center around.

Measure	Description	Commonly used for	{ 3, 4, 4, 10, 14, 43 }
Mode	most common value	nominal data	**4**
Median	middle value in a data set (or average of middle two values)	data sets expected to be skewed	$\frac{4+10}{2} = \mathbf{7}$
Mean	average of all values in a data set	most numerical data sets	$\frac{3+4+4+10+14+43}{6} = \mathbf{13}$
10% Trimmed Mean	average of all values in a data set except the highest and lowest 10% (other percents can be used as well)	data sets expected to be skewed	10% of $6 \approx 1$ $\frac{4+4+10+14}{4} = \mathbf{8}$

Measures of Variation

Measures of variation show how much variability there is within a data set.

Measure	Description	Example: { 1, 2, 3, 6 }
Sum of Squares (*SS*)	sum of the squared difference between each value and the mean	$\mu = 3$ $(1-3)^2 = 4$ $(2-3)^2 = 1$ $(3-3)^2 = 0$ $(6-3)^2 = 9$ **$SS = 4+1+0+9 = 14$**
Variance (σ^2)	average squared difference	**$\sigma^2 = 14 \div 4 = 3.5$**
Standard Deviation (σ)	square root of variance	**$\sigma = \sqrt{3.5} \approx 1.87$**
Coefficient of Variation (*CV*)	standard deviation divided by mean	**$CV \approx 1.87 \div 3 \approx 0.62$**

Methods to calculate standard deviation

Standard deviation, as well as many other statistics in this course, can be calculated a number of ways.

Method	Setup	Calculation	When used
Paper	Make a column for x, $x - \bar{x}$, and $(x - \bar{x})^2$.	Calculate μ, and use it to fill in the values in the other columns. Take the square root of the average of the last column.	initially, to understand what standard deviation actually represents, but rarely in a practical context
Calculator	Push `STAT`, choose `EDIT`, and enter the data into a list.	Push `STAT`, choose `CALC`, and choose `1-Var Stats`.	for relatively small data sets, such as most examples in this class
Spreadsheet	Do the same setup as on paper, but type in formulas instead of doing calculations.	Enter the data.	for large data sets, such as in most real-world contexts
Online	Read the directions of the particular website.	Submit the data.	for inconsequential data, when a calculator and computer are not available

Using samples to make population estimates

A common misunderstanding in statistics is the belief that sample statistics are values representing samples. Sample statistics are actually values representing population estimates. They are called sample statistics because they are calculated using sample data.

In many cases these are the same thing, but for standard deviation they are not.

Value	Meaning	Formula	Explanation	When used
population standard deviation	the standard deviation of all of the data	$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$	definition of standard deviation	when all of the data are known, such as each student's test score
sample standard deviation	an estimate of the standard deviation of all of the data, based on sample data	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	slightly higher than σ, to account for variation and outliers outside the sample	when only a sample is collected, such as in an experiment

Since the definition of statistics is using sample values to estimate population values, in most cases s is the correct value to calculate in a statistics course rather than σ.

Weighted Averages

A **weighted average** takes into account the importance of each category. A common use of weighted averages is when data are grouped into numerical ranges and not individually known. In the example below, college students graduated with an average debt estimated to be $940,000 ÷ 42 = **$22,381**.

College debt	Estimate *x*	# of Students *f*	Total *fx*
$0	$0	12	$0
$1 - $20,000	$10,000	14	$140,000
$20,001 - $50,000	$35,000	10	$350,000
$50,001 - $100,000	$75,000	6	$450,000
TOTAL		42	$940,000

In many cases, categories are given percentage weightings. A common use of this is college course grades or other rating systems. In the example below, John's semester grade is **87**.

Category	John's score *x*	Weighting *f*	Value *fx*
Paper I	90	20%	18
Paper II	100	20%	20
Midterm	84	25%	21
Final	80	35%	28
TOTAL		100%	87

Standard deviation of grouped data

Like mean, standard deviation can be estimated for grouped data. The data below are in \$1000's but otherwise are the same as above except for rounding. Using $\bar{x} \approx 22.4$ from before, they show that the sum of squares is $SS \approx 26{,}362$, making the variance $s^2 \approx \frac{26362}{41} \approx 643$ and the standard deviation $s \approx \sqrt{643} \approx 25.4$, that is, \$25,400.

Range	x	f	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
\$0	0	12	\$0	-22.4	500.9	6,011
\$0 - \$20	10	14	140	-12.4	153.3	2,146
\$20 - \$50	35	10	350	12.6	159.2	1,592
\$50 - \$100	75	6	450	52.6	2,768.8	16,613
TOTAL		42	940			26,362

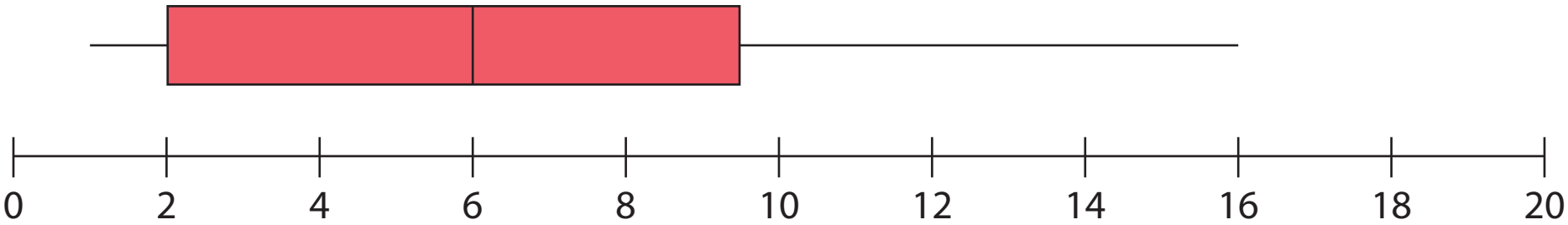
Quartiles

The **quartiles** of a data set divide it into four equal parts.

Quartile	Definition	{ 1, 1, 3, 5, 6, 7, 9, 10, 16 }
First (Q_1)	median of the values below the midpoint of the data set	median of {1, 1, 3, 5 }: **2**
Second (Q_2)	median of the data set	**6**
Third (Q_3)	median of the values above the midpoint of the data set	median of { 7, 9, 10, 16 }: **9.5**

A **box-and-whisker plot** shows the quartiles by showing a box from the first quartile to the third quartile and a line inside the box to mark the second quartile. Then whiskers are drawn from the box out to the lowest value and to the highest value.

It is important to draw the scale before placing the box or the whiskers.



The **range** is the total spread of the data, that is, the highest value minus the lowest value.

The **interquartile range** is the spread of the middle 50% of the data, that is, $Q_3 - Q_1$ (the size of the box).

Percentiles

Whereas quartiles divide a data set into quarters, **percentiles** divide a data set into hundredths, that is, percents. The p^{th} percentile can be thought of as the value below which are p% of the data in the data set.

There are different methods for calculating a p^{th} percentile, and in small data sets different methods may result in significantly different answers. One method is to find p percent of the sample size, add 0.5 to this, count this far into the data set, and, if this value is not an integer, add a proportional amount of the difference to the next number, such as adding .3 of the difference between 8 and the next number, 14, in the second example below.

Data set	index for 30^{th} percentile	30^{th} percentile
{ 4, 8, 14, 18, 24 }	30% of 5 is 1.5 $1.5 + 0.5 = 2$	the 2^{nd} number in data set is **8**
{ 4, 8, 14, 18, 24, 50 }	30% of 6 is 1.8 $1.8 + 0.5 = 2.3$	the "2.3^{rd}" number in the data set is $8 + 0.3(14 - 8) =$ **9.8**

Resistant Measures

An **outlier** is a value that is much further from the mean than most of the other data. Outliers can lead to misleading statistics, such as if four people's mile times are 5:00, 6:00, 6:00, and 31:00, making the average time 12:00.

A **resistant measure** is one that does not use outliers as part of its calculation, and thus is unaffected by outliers. Some examples are shown below.

Measure	{ 1, 2, 3, 4, 5 }	{ 1, 2, 3, 4, 500 }	Resistant
Median	3	3	yes
Mean	3	102	no
Standard Deviation	1.4	199	no