Chapter 3-numerical statistics

numerical Measures for Summarizing Data

Types: 1. Measures of Central Tendency
2. Measures of Variability
3. Measures of Relative Location

Measures of Central Tendency: 1. Mean 3. MODE
2. Median 4. TRIMMED MEAN

The **arithmetic mean** of a set of n measurements $y_1, y_2, ..., y_n$ is equal to the sum of the measurements divided by n. The mathematical notation for the arithmetic mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The **median** of a set of n measurements $y_1, y_2, ..., y_n$ is the value that falls in the middle position when the measurements are ordered from the smallest to the largest.

Rule for Calculating the Median:

1. Order the measurements from the smallest to the largest.
2. a. If the sample size is odd, the median is the middle measurement.
   b. If the sample size is even, the median is the average of the two middle measurements.

Example: A random sample of six measurements was taken from a population. These measurements were: $y_1=7$, $y_2=1$, $y_3=10$, $y_4=8$, $y_5=4$, and $y_6=12$. What is the sample mean and sample median for these data?

Calculations for the Sample Mean

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6}{6} = \frac{(7 + 1 + 10 + 8 + 4 + 12)}{6} = \mathbf{7}$$

Calculations for the Sample Median

Ordered Sample

$y_2=1, \quad y_5=4, \quad y_1=7, \quad y_4=8, \quad y_3=10, \quad y_6=12$

median = (7+8)/2 = **7.5**

Consider the following sample.

4	18	36	39	41	42	43	44	44	45
46	47	48	49	49	50	51	53	54	60

Below is the stem and leaf plot for these data.

Stem	Leaf
0	4
0	
1	
1	8
2	
2	
3	
3	69
4	12344
4	567899
5	0134
5	
6	0

The sample mean is 43.15 and the sample median is 45.5.

Which one gives you a "better" measure of location?

Can you identify any point which looks like it may be an incorrect value? If so, which one? A value like this is often referred to as an "extreme value" or **outlier**.

Remove the point you identified as a possible incorrect value and recalculate the sample mean and sample median.

Statistics that are not greatly influenced by extreme values are called "**robust statistics**". Is the sample mean or sample median a robust statistic? Which one?

Measurements of Variability
1. Range
2. Variance
3. Standard Deviation

The **sample range** is the difference between the largest and smallest measurements.

The **deviation** of an observation y_i from the sample mean is equal to

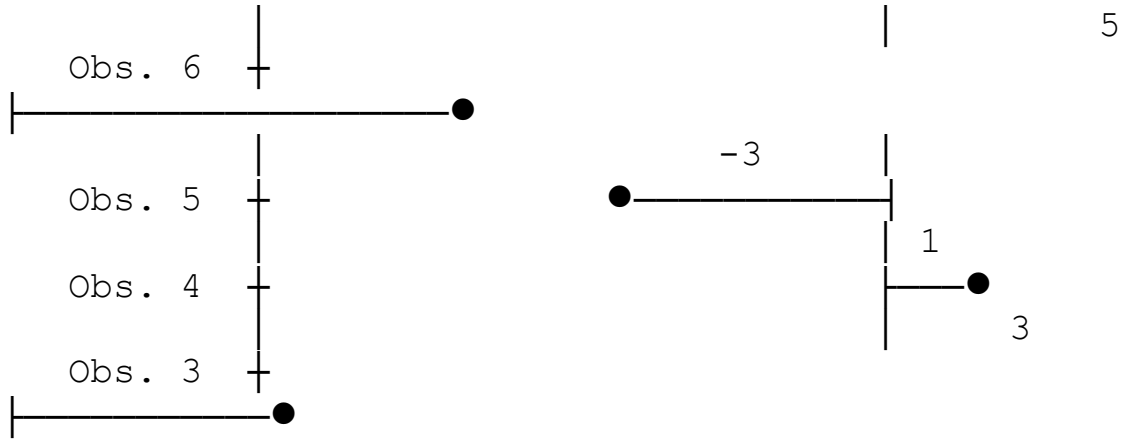
$$(y_i - \bar{y})$$

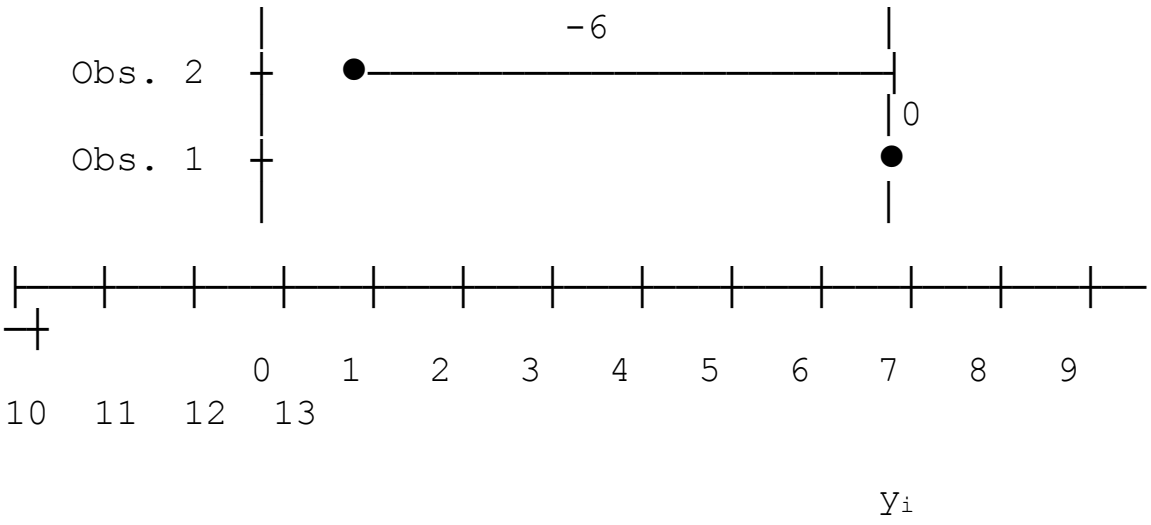
Deviations that are negative lie to the left of the sample mean and deviations that are positive lie to the right of the sample mean. Also, notice that the larger the squared deviation, the further away the observation is from the mean.

Example: A random sample of six measurements was taken from a population. These measurements were: y_1=7, y_2=1, y_3=10, y_4=8, y_5=4, and y_6=12. How much do these observations deviate from the sample mean?

Deviations

$\bar{y} = 7$




-6
Obs. 2
0
Obs. 1
0 1 2 3 4 5 6 7 8 9
10 11 12 13
y_i

The **sample variance** is a measure similar to the average of the squared deviations. The formula for the sample variance is:

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

The **sample standard deviation (s)** is the positive square root of the sample variance.

Example: A random sample of six measurements was taken from a population. These measurements were: $y_1=7$, $y_2=1$, $y_3=10$, $y_4=8$, $y_5=4$, and $y_6=12$. What is the sample variance and the sample standard deviation for these data?

Obs.	y	$y-\bar{y}$	$(y-\bar{y})^2$
1	7	0	0
2	1	−6	36
3	10	3	9
4	8	1	1
5	4	−3	9
6	12	5	25
			80

Obs.	y	y^2
1	7	49
2	1	1
3	10	100
4	8	64
5	4	16
6	12	144
	42	374

$$s^2 = \frac{\sum_{i=1}^{n}(y-\bar{y})^2}{n-1} = \frac{80}{5} = 16$$

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{374-\frac{(42)^2}{6}}{5} = \frac{80}{5} = 16$$

The sample standard deviation is: $s = \sqrt{s^2} = \sqrt{16} = 4$

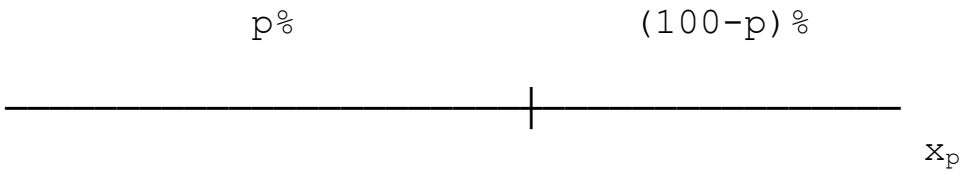
Why do we need the standard deviation?
The units of measurement for the mean and variance are different. For example, consider the problem where our observations are the measurement unit of inches. Then the measurement unit for the mean would also be in inches, but the measurement unit for the variance would be in inches squared. By taking the square root of the variance, we now have the same measurement units for both the mean and the standard deviation.

3. Measures of Relative Standing

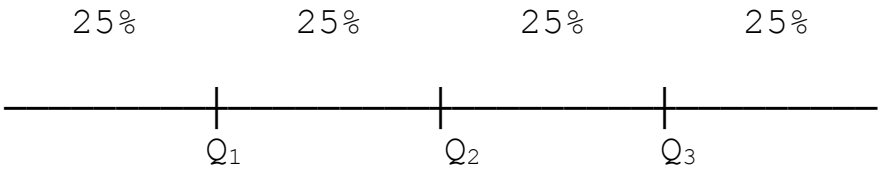
A measure of relative standing finds the position of a value relative to the other values in a set of observations.

Two Measures of Relative Standing
1. Percentile
2. Quartile

The **p^{th} percentile** is a value x_p such that p% of the measurements will fall below that value and (100-p)% of the measurements will fall above that value.



Quartiles divide the measurements into four parts such that 25% of the measurements are contained in each part. The first quartile (lower quartile) is denoted by Q_1, the second by Q_2, and the third (upper) by Q_3.



Jane took the SAT math test and received a score of 550. The form indicated that the score was the 70th percentile. What does that mean?

What percentage of the measurements should fall between the first and third quartiles?

Another measure of variability

The **interquartile range (IQR)** of a set of measurements is the difference between the upper and lower quartiles. It is the range of the middle 50% of the measurements.

The Box and Whisker Plot

The **box plot (box-and-whiskers plot)** is concerned with the symmetry of the data and incorporates measures of central tendency and location in order to study the variability of the measurements. So it can be used to describe both the behavior of the measurements in the middle and at the ends of the distribution. Below is an example of the stem and leaf and the box-and-whiskers plot for the life expectancy data.

Life Expectancies in 29 Developed Nations

```
     Stem Leaf                      #
Boxplot
       79 1                        1
|
       78
|
       77 146                      3
|
       76 335556                   6             +-
----+  75% Q3   76.5
       75 13589                    5             *-
----*  50% Med. 75.1
       74 1123889                  7             +-
-+--+  25% Q1   74.1
       73 2                        1
|
       72
|   + Mean 74.65
       71 005                      3
|
       70
       69 789                      3
0
         ----+----+----+----+
```

The box-and-whiskers plot can be expanded to provide information about observations that may be outliers or extreme values. This is done by constructing "fences" for the measurements. These fences are found using the following formulae:

Lower inner fence: $Q_1 - 1.5(IQR)$
Upper inner fence: $Q_3 + 1.5(IQR)$

Lower outer fence: $Q_1 - 3(IQR)$
Upper outer fence: $Q_3 + 3(IQR)$

A measurement beyond an inner fence is called a mild outlier, while a measurement beyond an outer fence is called an extreme outlier. Mild outliers are denoted by the symbol o and extreme outliers are denoted by the symbol *.