Math 3338: Probability (Fall 2006)

Jiwen He

Section Number: 10853

`http://math.uh.edu/~jiwenhe/math3338fall06.html`

Chapter One
Overview and Descriptive Statistics (II)

1.3 Measures of Location

The Mean and The Median

DEFINITION

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

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

The numerator of $\bar{x}$ can be written more informally as Σx_i where the summation is over all sample observations.

DEFINITION

The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then,

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\[2ex] \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{ and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

Example 1.11. Windspan data

5H	9
6L	00122334
6H	5566799
7L	02
7H	55
8L	
8H	
9L	
9H	5

Figure 1.12 A stem-and-leaf display of the wingspan data

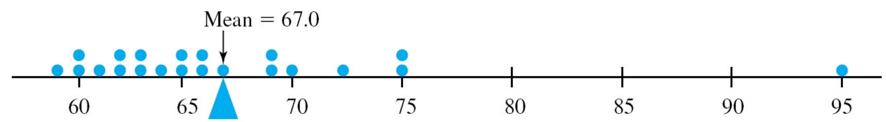

Figure 1.13 The mean as the balance point for a system of weights

- **Measure of the center:** $\bar{x}$ and $\tilde{x}$ provide a measure for the center of data set, but will not in general be equal: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1408}{21} = 67.0$ and $\tilde{x} = \left(\frac{n+1}{2}\right)^{\text{th}}$ ordered value $= 65$.
  - **Balance point:** $\bar{x}$ represents the average value of the observations in the sample. The point at $\bar{x}$ is the only point at which a fulcrum can be placed to balance the system of weights: $\sum (x_i - \bar{x}) = 0$.
  - **Middle point:** $\tilde{x}$ represents the middle value in the samele. It divides the data set into two parts of equal size.
- **Sensitivity to outliers:** $\bar{x}$ and $\tilde{x}$ are at opposite ends of a spectrum.
  - The mean $\bar{x}$ can be greatly affected by the presence of outliers. Without the outlier $x_{19} = 95$, $\bar{x} = 65.7$.
  - The median $\tilde{x}$ is insensitive to outliers. Without $x_{19} = 95$, $\tilde{x} = 65$.

Population Mean and Population Median

- **Population mean** μ: the average of all values in the population.
- **Population median** $\tilde{\mu}$: the middel value in the population.
- **Finite population:** $\mu = \dfrac{\text{sum of the } N \text{ population values}}{N}$.
- **Statistic inference:** use the sample mean $\bar{x}$ and the sample median $\tilde{x}$ to make an inference about the population mean μ and the population median $\tilde{\mu}$.
- **Measure of the center:** μ and $\tilde{\mu}$ will not generally be identical:

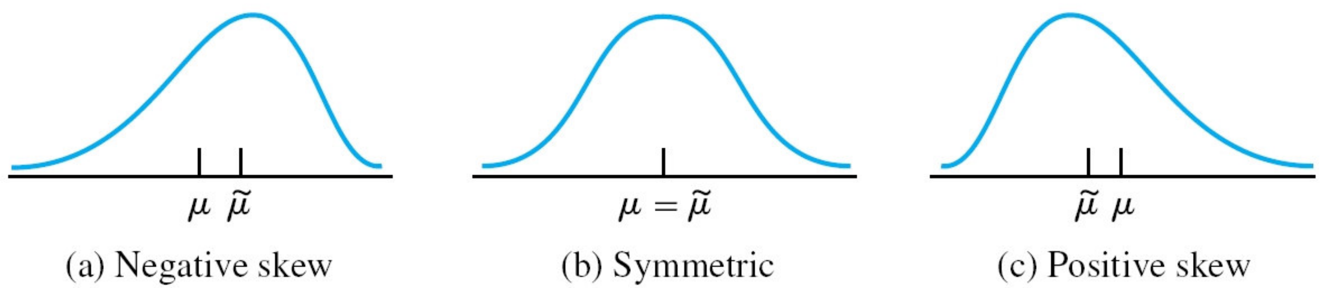

Figure 1.14 Three different shapes for a population distribution

- If the population distribution is positively ($\mu > \tilde{\mu}$) or negatively skewed ($\mu < \tilde{\mu}$), then $\mu \neq \tilde{\mu}$.

Quartiles, Percentiles, and Trimmed Means

- **Quartiles:** quartiles divide the data set into *four equal parts*, with the observations above the *third quartile* constituting the upper quarter of the data set, the *second quartile* being identical to the median, and the *first quartile* separating the lower quarter from the upper three-quarters.
- **Percentiles:** a data set (sample or population) can be even more finely divided using percentiles; the 99th percentile separates the highest 1% from the bottom 99%, and so on.
- **Trimmed mean and various sensitivity to outliers:**
  - **Median $\tilde{x}$:** computed throwing away as many values on each end as one can without eliminating everyting and average what is left.
  - **Mean $\bar{x}$:** computed throwing away nothing before averaging.
  - **Trimmed mean:** a compromise between $\bar{x}$ and $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.
  - Generally speaking, using a trimmed mean with a moderate trimming proportion (between 5 and 25%) will yield a measure that is neither as *sensitive* to outliers as the mean nor as *insensitive* as the median.

1.4 Measures of Variability

Measures of variability for sample data

- Fig. 1.16 shows dotplots of three samples with the same mean and median, yet the extent of spread about the center is different for all three samples.

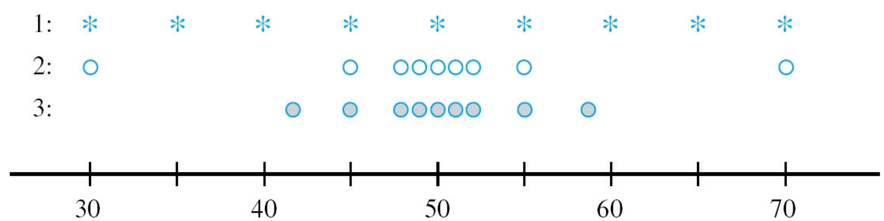

Figure 1.16 Samples with identical measures of center but different amounts of variability

- **Range:** the difference between the largest and smallest sample values.
- **Deviations from the mean:** obtained by subtracting $\bar{x}$ from each of the n sample observations: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$.

$$\text{sum of deviations} = \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

- **Variance:** denoted by s^2, is given by

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

Notice that the sum of squared deviations is divided by $n-1$ rather than n.

- **Standard deviation:** denoted by s, is given by $s = \sqrt{s^2}$.

Example 1.14. Postsurgical data

Table 1.3 Data for Example 1.14

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
154	23.62	557.904
142	11.62	135.024
137	6.62	43.824
133	2.62	6.864
122	−8.38	70.224
126	−4.38	19.184
135	4.62	21.344
135	4.62	21.344
108	−22.38	500.864
120	−10.38	107.744
127	−3.38	11.424
134	3.62	13.104
122	−8.38	70.224
$\sum x_i = 1695$	$\sum (x_i - \bar{x}) = .06$	$\sum (x_i - \bar{x})^2 = 1579.1$
$\bar{x} = \frac{1695}{13} = 130.38$		

$$s^2 = \frac{1579.1}{13 - 1} = 131.59, \quad s = \sqrt{131.59} = 11.47.$$

Motivation for s^2

- **Population variance:** when the population is finite,

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

  which is the average of all squared deviations from the pupulation mean.

- **Question:** why s^2 rather than the average squared deviation is used.
  - One could define s^2 as the average squared deviation of the sample x_i's about μ:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n},$$

    but μ is almost never known, so the sum of squared deviations about $\bar{x}$ must be used.
  - The x_i's tend to be closer to $\bar{x}$ than to μ, so to compensate for this the divisor $n-1$ is used rather than n.
  - We refer to s^2 as being based on $n-1$ degrees of freedom; recall that $\sum(x_i - \bar{x}) = 0$.

A computing formula for s^2

- **Rounding:** To guard against the effects of rounding, an alternative expression for s^2 is:

$$s^2 = \frac{S_{xx}}{n-1} \quad \text{where } S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$$

- **Example 1.15 - Remote sensing:**

Observation	x_i	x_i^2	Observation	x_i	x_i^2
1	15.2	231.04	9	12.7	161.29
2	16.8	282.24	10	15.8	249.64
3	12.6	158.76	11	19.2	368.64
4	13.2	174.24	12	12.7	161.29
5	12.8	163.84	13	15.6	243.36
6	13.8	190.44	14	13.5	182.25
7	16.3	265.69	15	12.9	166.41
8	13.0	169.00			
				$\sum x_i = 216.1$	$\sum x_i^2 = 3168.13$

$$S_{xx} = 3168.13 - \frac{(216.1)^2}{15} = 54.85, \quad s = \frac{54.85}{14} = 3.92.$$

Simplest Boxplots

DEFINITION Order the n observations from smallest to largest and separate the smallest half from the largest half; the median $\tilde{x}$ is included in both halves if n is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread** f_s, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

- **Five-number summary:** on which the simplest boxplot is based: smallest x_i, lower fourth, median, upper fourth, largest x_i.
- **Example 1.16:** Corrosion data

$$\underbrace{40\ 52\ 55\ 60\ 70\ 75\ 85\ 85\ 90\ 90}\ \ \overbrace{90\ 92\ 94\ 94\ 95\ 98\ 100\ 115\ 125\ 125}$$

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$

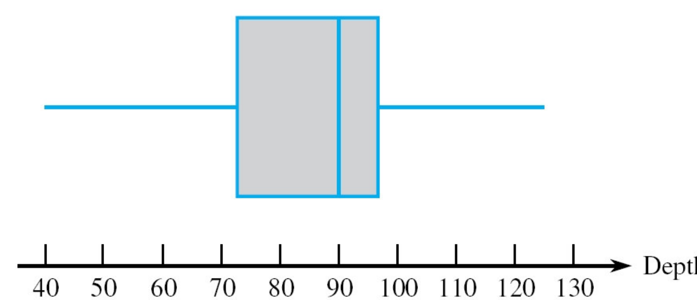

Figure 1.17 A boxplot of the corrosion data

Boxplots that show outliers

DEFINITION Any observation farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.

- **Example 1.17:** Pulse width data

5.3	8.2	13.8	74.1	85.3	88.0	90.2	91.5	92.4	92.9	93.6	94.3	94.8
94.9	95.5	95.8	95.9	96.6	96.7	98.1	99.0	101.4	103.7	106.0	113.5	

Relevant quantities are

$\tilde{x} = 94.8$ lower fourth $= 90.2$ upper fourth $= 96.7$

$f_s = 6.5$ $1.5f_s = 9.75$ $3f_s = 19.50$

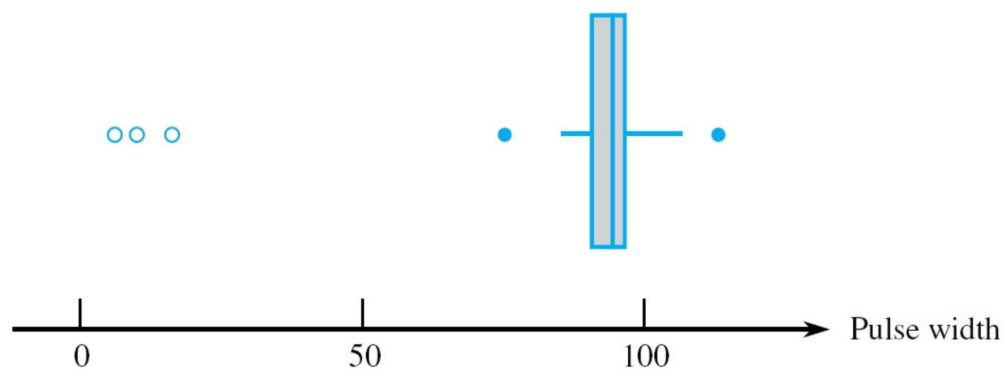

Figure 1.19 A boxplot of the pulse width data showing mild and extreme outliers

Comparative Boxplots - Example 1.18

1. Cancer		2. No cancer
9683795	0	95768397678993
86071815066815233150	1	12271713114
12302731	2	99494191
8349	3	839
5	4	
7	5	55
	6	
	7	Stem: Tens digit
HI: 210	8	5 Leaf: Ones digit

Figure 1.20 Stem-and-leaf display for Example 1.18

	$\bar{x}$	$\tilde{x}$	s	f_s
Cancer	22.8	16.0	31.7	11.0
No cancer	19.2	12.0	17.0	18.0

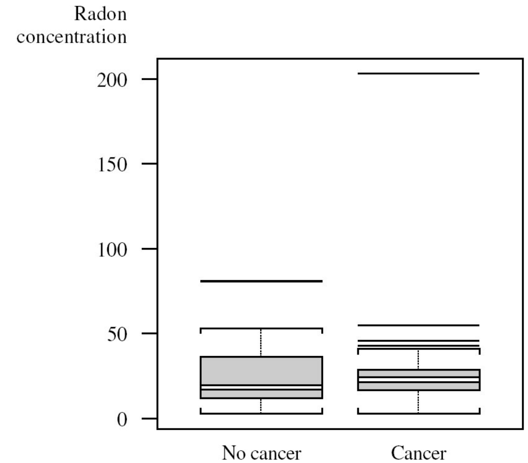

Figure 1.21 A boxplot of the data in Example 1.18, from S-Plus

- A comparative or side-by side boxplot is a very effective way to revealing similarities and differences between two or more data sets consisting of observations on the same variable.