IES302 2011/2 Part II.1 Dr.Prapun

11 Descriptive Statistics

Descriptive > Inferential

Statistics is the science of data. An important aspect of dealing with data is organizing and summarizing the data in ways that facilitate its interpretation and subsequent analysis. This aspect of statistics is called ***descriptive statistics***.

Sample vs. Population

Definition 11.1. In most statistics problems, we almost always work with a ***sample*** of observations that has been selected from some larger ***population*** of observations.

(a) In general, a ***population*** is defined as a collection of persons, objects, or items of interest.

- The population can be a widely defined category, such as "all automobiles," or it can be narrowly defined, such as "all Ford Mustang cars produced from 2008 to 2010."
- A population can be a group of people, such as "all workers presently employed by Microsoft," or it can be a set of objects, such as "all dishwashers produced on March 13, 2012, by the General Electric Company at Louisville plant."
- The researcher defines the population to be whatever he or she is studying.
- When researchers gather data from the whole population for a given measurement of interest, they call it a **census**.

parameter = a number that characterizes the population

a subset of the population

(b) A sample is a portion of the whole and, if properly taken, is representative of the whole.

- Because of time and money limitations, researcher often prefer to work with a sample of the population instead of the entire population.
- For example, in conducting quality-control experiments to determine the average life of lightbulbs, a lightbulb manufacturer might randomly sample only 75 lightbulbs during a production run.

"Conceptual" population

We often find it useful to describe data features numerically. For example, we can characterize the location or central tendency in the data by the ordinary arithmetic average or mean. Because we almost always think of our data as a sample, we will refer to the arithmetic mean as the sample mean.

Definition 11.2. If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

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

read "x bar"

$$\overline{g(x)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

The variability or scatter in the data may be described by the sample variance or the sample standard deviation.

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Definition 11.3. If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the ***sample variance*** is

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

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

$$s = \sqrt{s^2}$$

11.4. The units of measurement for the sample variance are the *square* of the original units of the variable. The standard deviation has the desirable property of measuring variability in the original units of the variable of interest.

Example 11.5. Suppose a sample contains three observations: $x_1 = 1, x_2 = 3, x_3 = 4$.

Sample mean : $\bar{x} = \frac{1}{3}(1+3+4) = \frac{8}{3}$, $\sum_{i=1}^{3} x_i = 8$

Sample variance :

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

$= \frac{1}{2}\left((1-\frac{8}{3})^2 + (3-\frac{8}{3})^2 + (4-\frac{8}{3})^2\right)$

$= \frac{1}{2}\left(\frac{(-5)^2 + 1^2 + 4^2}{9}\right) = \frac{42}{2\times 9} = \frac{7}{3}$

$1^2+3^2+4^2 = 26$

② $s^2 = \frac{1}{n-1}\left(\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2\right)$

$= \frac{1}{2}\left(26 - \frac{1}{3}8^2\right)$

$= \frac{1}{2}\times\frac{78-64}{3} = \frac{14}{6} = \frac{7}{3}$

11.6. The sample mean and sample variance could be used as estimates of the population mean and the population variance, respectively.

For finite population, we have Population $a_1, a_2, \ldots, a_N$; Sample $x_1, x_2, \ldots, x_n$

Population mean : $\mu = \frac{1}{N}\sum_{i=1}^{N} a_i$ → estimated by → Sample mean $= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Population variance : $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i-\mu)^2$ → Sample variance $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

If μ is known, then use $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

Definition 11.7. Another useful measure of variability is the sample range which is simply the difference between the largest and smallest observations. In particular, if the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the ***sample range*** is

$$r = \max(x_i) - \min(x_i)$$

Although the range is easy to use and understand, it has the weakness of being able to recognize only the extreme values in the data.

Definition 11.8. The sample ***median*** is a measure of central tendency that divides the (sorted) data into two equal parts, half below the median and half above.

- If the number of observations is even, the median is halfway between the two central values.
- If the number of observations is odd, the median is the central value.

Example 11.9. Consider the data shown in Figure 11 for shipments of peanuts from a hypothetical U.S. exporter to five Canadian cities.

City	Peanuts (Thousands of Bags)
Montreal	64.0
Ottawa	15.0
Toronto	285.0
Vancouver	228.0
Winnipeg	45.0

Figure 11: The numbers of bags of peanuts (in thousands) shipped by the U.S. exporter to five Canadian cities.

ordered data: 15 45 64 228 285

median

Example 11.10. Ryder System, Inc. reported the following data for percentage return on average assets over an 8-year period[20]

Raw data: 2.8 7.0 1.6 0.4 1.9 2.6 3.8 3.8

mode = 3.8

ordered data: 0.4 1.6 1.9 2.6 2.8 3.8 3.8 7.0

median = 2.7 $\left(\frac{2.6+2.8}{2}\right)$

Definition 11.11. The sample ***mode*** is the most frequently occurring data value.

[20]Source: Ryder System, Inc., 2005, 2003, and 1998 Annual Reports.

Example 11.12. Consider again the 8 years of return-on-assets data reported by Ryder System, Inc.:

11.13. Depending on the data, there can be more than one mode. For example, if the leftmost data value in Example 11.12 had been 1.6 instead of 0.4, there would have been two modes, 1.6 and 3.8

11.14. Distribution Shape and Measures of Central Tendency: The relative values of the mean, median, and mode are very much dependent on the shape of the distribution for the data they are describing.

Skewness refers to the tendency of the distribution to "tail off" to the right or left, as shown in parts (b) and (c) of Figure 12. In examining the three distribution shapes shown in the figure, note the following relationships among the mean, median, and mode:

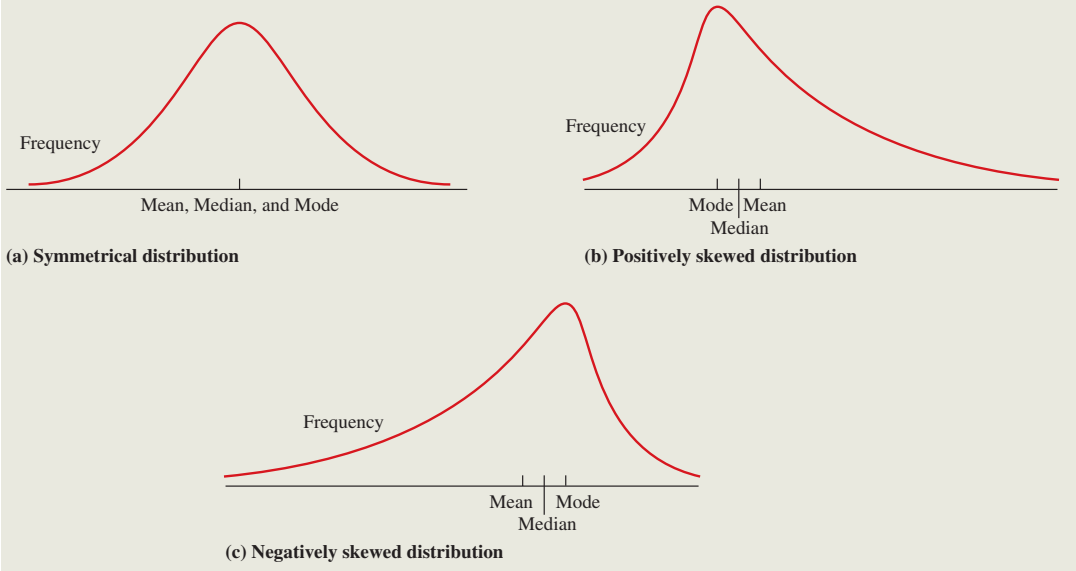


Figure 12: The shape of the distribution determines the relative positions of the mean, median, and mode for a set of data values.

(a) ***Symmetrical*** distribution: The mean, median, and mode are the same. This will be true for any symmetrical unimodal

distribution, such as this one. (When a distribution is bimodal, it will of course be impossible for the mean, median, and both modes to be equal.)

(b) ***Positively skewed*** distribution: The mean is greater than the median, which in turn is greater than the mode.

- Income distributions tend to be positively skewed, since there is a lower limit of zero, but practically no upper limit on how much a select few might earn. In such situations, the median will tend to be a better measure of central tendency than the mean.

(c) ***Negatively skewed*** distribution: The mean is less than the median, which in turn is less than the mode. Data that have an upper limit (e.g., due to the maximum seating capacity of a theater or stadium) may exhibit a distribution that is negatively skewed. As with the positively skewed distribution, the median is less influenced by extreme values and tends to be a better measure of central tendency than the mean.

Example 11.15. The incomes[21] of males who were 25 or older in 2000 were distributed as shown in Figure 13:

	Age Group				
Income	**25 to <35**	**35 to <45**	**45 to <55**	**55 to <65**	**65 or over**
under $10,000	8.4%	7.7%	8.4%	11.6%	18.7%
$10,000–under $25,000	29.9	21.0	17.7	24.4	45.2
$25,000–under $50,000	40.5	36.7	34.6	30.4	23.5
$50,000–under $75,000	13.8	19.6	20.3	16.7	6.4
$75,000 or over	7.4	15.0	19.0	16.9	6.3
	100.0	100.0	100.0	100.0	100.1*
Median (thousands)	$30.63	$37.09	$41.07	$34.41	$19.17
Mean (thousands)	$42.15	$53.48	$58.46	$61.05	$56.50

*Differs from 100.0 due to rounding.

Figure 13: Income distribution for male income-earners in the United States

[21]Source: Bureau of the Census, U.S. Department of Commerce, Statistical Abstract of the United States 2002, pp. 439440.

For each of these distributions, the mean exceeds the median, reflecting that incomes are skewed positively (toward larger incomes) for all of these age groups.

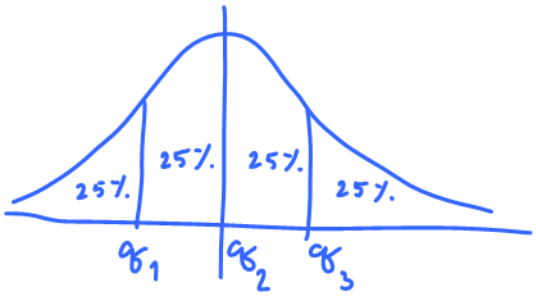
Definition 11.16. When an ordered set of data is divided into four equal parts, the division points are called ***quartiles***.

(a) The ***first*** or ***lower quartile***, q_1, is a value that has approximately 25% of the observations below it and approximately 75% of the observations above.

(b) The ***second quartile***, q_2, has approximately 50% of the observations below its value.

- The second quartile is exactly equal to the median.

(c) The ***third*** or ***upper quartile***, q_3, has approximately 75% of the observations below its value.



Example 11.17. Figure 14 shows the data on new privately owned housing[22].

Definition 11.18. The ***interquartile range*** (IQR) is defined as

$q_3 - q_1$ Quartile deviation : $\frac{q_3 - q_1}{2}$

It can also be used as a measure of variability.

Definition 11.19. The ***box plot*** (or box-and-whisker plot) is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.

[22]Source: Bureau of the Census, U.S. Department of Commerce, Construction Reports, series C20.

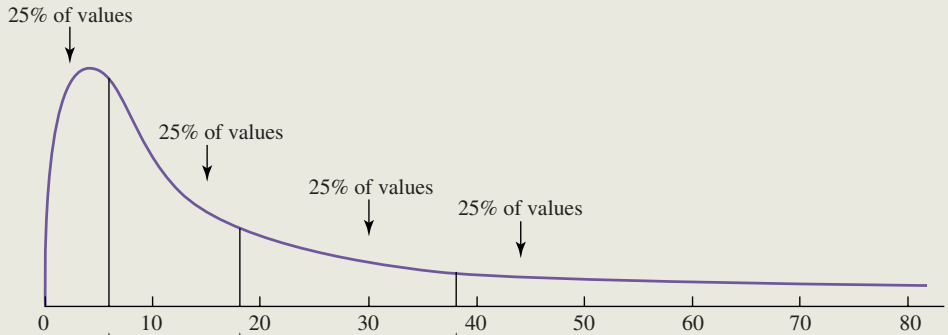
(a) Raw data (new housing starts, in thousands)

State	Starts	State	Starts	State	Starts	State	Starts	State	Starts
AL	17.2	HI	7.3	MA	39.2	NM	11.8	SD	2.5
AK	4.0	ID	4.3	MI	37.6	NY	61.9	TN	38.1
AZ	71.8	IL	38.7	MN	28.6	NC	70.7	TX	143.1
AR	9.9	IN	23.0	MS	8.8	ND	2.6	UT	16.5
CA	271.4	IA	5.2	MO	27.2	OH	33.0	VT	4.1
CO	32.8	KS	13.3	MT	2.0	OK	10.7	VA	64.1
CT	24.5	KY	13.8	NE	5.0	OR	11.3	WA	35.5
DE	4.6	LA	18.8	NV	14.0	PA	43.6	WV	1.5
FL	202.6	ME	8.1	NH	17.8	RI	5.4	WI	20.2
GA	73.1	MD	42.1	NJ	55.0	SC	32.8	WY	1.2

(b) Data arranged from smallest to largest

1.2	5.2	13.8	28.6	43.6
1.5	5.4	14.0	32.8	55.0
2.0	7.3	16.5	32.8	61.9
2.5	8.1	17.2	33.0	64.1
2.6	8.8	17.8	35.5	70.7
4.0	9.9	18.8	37.6	71.8
4.1	10.7	20.2	38.1	73.1
4.3	11.3	23.0	38.7	143.1
4.6	11.8	24.5	39.2	202.6
5.0	13.3	27.2	42.1	271.4

(c) Quartiles



Interquartile range = $Q_3 - Q_1 = 38.8 - 6.8 = 32.0$

Quartile deviation = $\frac{Q_3 - Q_1}{2} = \frac{38.8 - 6.8}{2} = 16.0$

Figure 14: Raw data, ordered data, and quartiles for new privately owned housing starts in the 50 U.S. states

- The box encloses the interquartile range with the left (or lower) edge at the first quartile, q_1, and the right (or upper) edge at the third quartile, q_3.
- A line is drawn through the box at the second quartile (which is the 50th percentile or the median),
- A line, or whisker, extends from each end of the box. The lower whisker is a line from the first quartile to the smallest data point within 1.5 interquartile ranges from the first quartile. The upper whisker is a line from the third quartile to the largest data point within 1.5 interquartile ranges from the third quartile.
- Data farther from the box than the whiskers are plotted as individual points.
- A point beyond a whisker, but less than three interquartile ranges from the box edge, is called an outlier.
- A point more than three interquartile ranges from the box edge is called an extreme outlier.

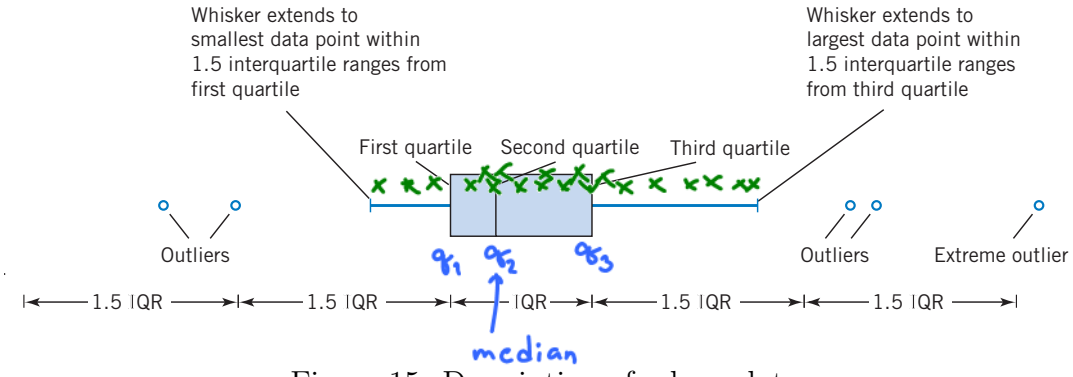


Figure 15: Description of a box plot.

Occasionally, different symbols, such as open and filled circles, are used to identify the two types of outliers. Sometimes box plots are called box-and-whisker plots.

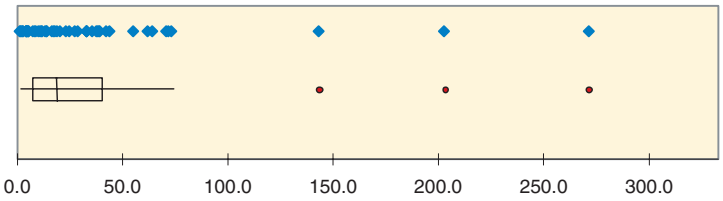


Figure 16: Example of a box plot.

11.20. Box plots are very useful in graphical comparisons among data sets, because they have high visual impact and are easy to understand. For example, Figure 17 shows the comparative box plots for a manufacturing quality index on semiconductor devices at three manufacturing plants. Inspection of this display reveals that there is too much variability at plant 2 and that plants 2 and 3 need to raise their quality index performance.

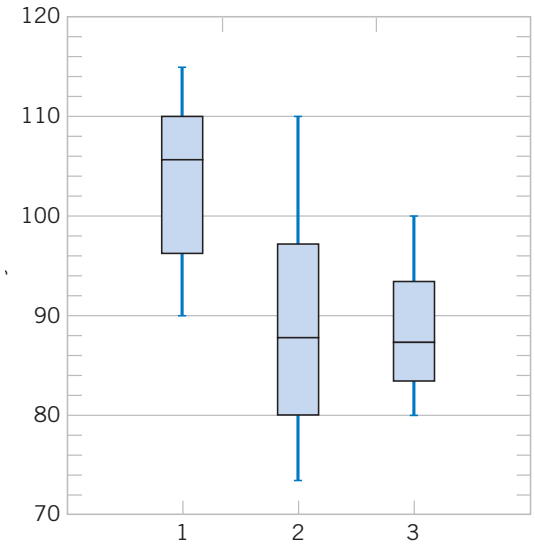


Figure 17: Comparative box plots.