STAT 511

Overview and Descriptive Statistics

Prof. Michael Levine

January 14, 2019

Populations and Samples

- A **population** is a well-defined collection of objects.
- An example: all gelatin capsules of a particular type produced during a specified period
- When information is available for the entire population we have a **census**
- A subset of the population is a **sample**
- Example: a sample of bearings from a particular production run

Populations and Samples

- A **variable** is any characteristic whose value may change from one object to another in the population **Univariate** data consists of observations on a single variable
- An example: a type of transmission (automatic or manual) on each of ten automobiles recently purchased
- **Multivariate** data - observations made on more than two variables.
- An example: a record of systolic and diastolic blood pressure as well as the serum cholesterol level for each patient

Example of an Experiment: Coin Spinning

- It is claimed sometimes that the coin spinning produces outcomes different from tossing the same coin. In particular, one often hears that the likelihood of obtaining heads (H) is less than 50% in that case
- 1. Choose a penny. What is the chance of obtaining H by spinning that penny? Also, are two pennies equally likely to produce H when spun?
  2. Choose several pennies minted before 1982 and several pennies minted after 1982. As groups, are pre-1982 pennies and post-1982 pennies equally likely to produce H when spun? (Before 1982 - 95% copper and 5% zinc; after 1982 - 97.5% zinc and 2.5% copper)

Branches of Statistics

- Often, the first task after the data collection is to summarize the data
- This may involve using both graphical methods and calculation of numerical summary measures.
- **Descriptive Statistics** - summary and description of collected data.
- After obtaining a sample from the population, it is frequently necessary to draw some conclusion (make an inference) about the population as a whole
- **Inferential Statistics** - generalizing from a sample to a population

Randomization:Lanarkshire milk experiment

- Lanarkshire milk experiment: 5000 children received a daily supplement of 3/4 pint of raw milk, 5000 received 3/4 pint of pasteurized milk and $10,000$ children received no daily supplement
- Each child was weighed (while wearing indoor clothing) and measured for hight in February of 1930 (before the start of the study) and in June of 1930 (after the end of the study).
- The final observations of the control group exceeded those of the treatment groups by average amounts equivalent to 3 months growth of weight and 4 months of growth in height...Why?!

Randomization:Lanarkshire milk experiment

- Initially, the division into **treatment** vs. **control** was arbitrary, e.g. using the alphabet
- If the initial division produced groups with unbalanced numbers of well-fed or ill-nourished children, teachers were allowed to swap children between the two groups to produce (apparently) better balanced groups...Sounds interesting? It should!!
- Another culprit was the quality of clothing worn by children...

The pre-election polls of 1948

- Presidential elections of 1948 : Truman vs. Dewey (he of the two Hollywood films fame :"Smashing the Rackets" and "Racket Busters")
- Crossley poll: 50% vs. 45%; Gallup poll 50% vs. 44%; Roper poll: 53% vs. 38%
- The actual result: slightly less than 50% vs. slightly more than 45%...Why?!

Quota sampling

- First, one selects some important characteristics of the population, e.g. age, sex, race etc.
- Second, one attempts to obtain a sample that mimicks the general population with respect to these characteristics. For example, out of 15 people, there should be 7 men and 8 women. Out of 7 men, 3 have to be under 40 years of age etc...
- The problem here is nobody specifies how to choose *within* quotas. Commonly, interviewers select people with higher educational levels because they are easier to deal with.
- More educated voters tended to be more affluent and voted Republican in higher numbers...

Randomization in practice: Case I

- There is a need to test two versions of the final exam. Let's say, we have 40 students..
- Write the names of students on the slips of paper and put them in a jar
- **Sample without replacement** 20 names - those will make up Group 1
- The rest will make up the group 2

Randomization in practice: Case II

- Now imagine that we know in advance that the class comprises 30 freshmen and 10 nonfreshmen. We believe it is essential to have 3/4 freshmen in each group
- Create 40 slips of paper again but now separate freshman and nonfreshman slips.
- First, draw 15 slips out of 30 freshmen slips; 15 receive exam A, the rest-exam B
- Second, draw 5 slips out of 10 remaining slips; 5 receive exam A, the rest - exam B
- This is called **stratified random sampling**

Histograms of Discrete Data

- Determine the frequency and relative frequency for each value of x.
- Mark possible values of x on a horizontal scale
- Above each value, draw a rectangle whose height is the relative frequency of that value

Example

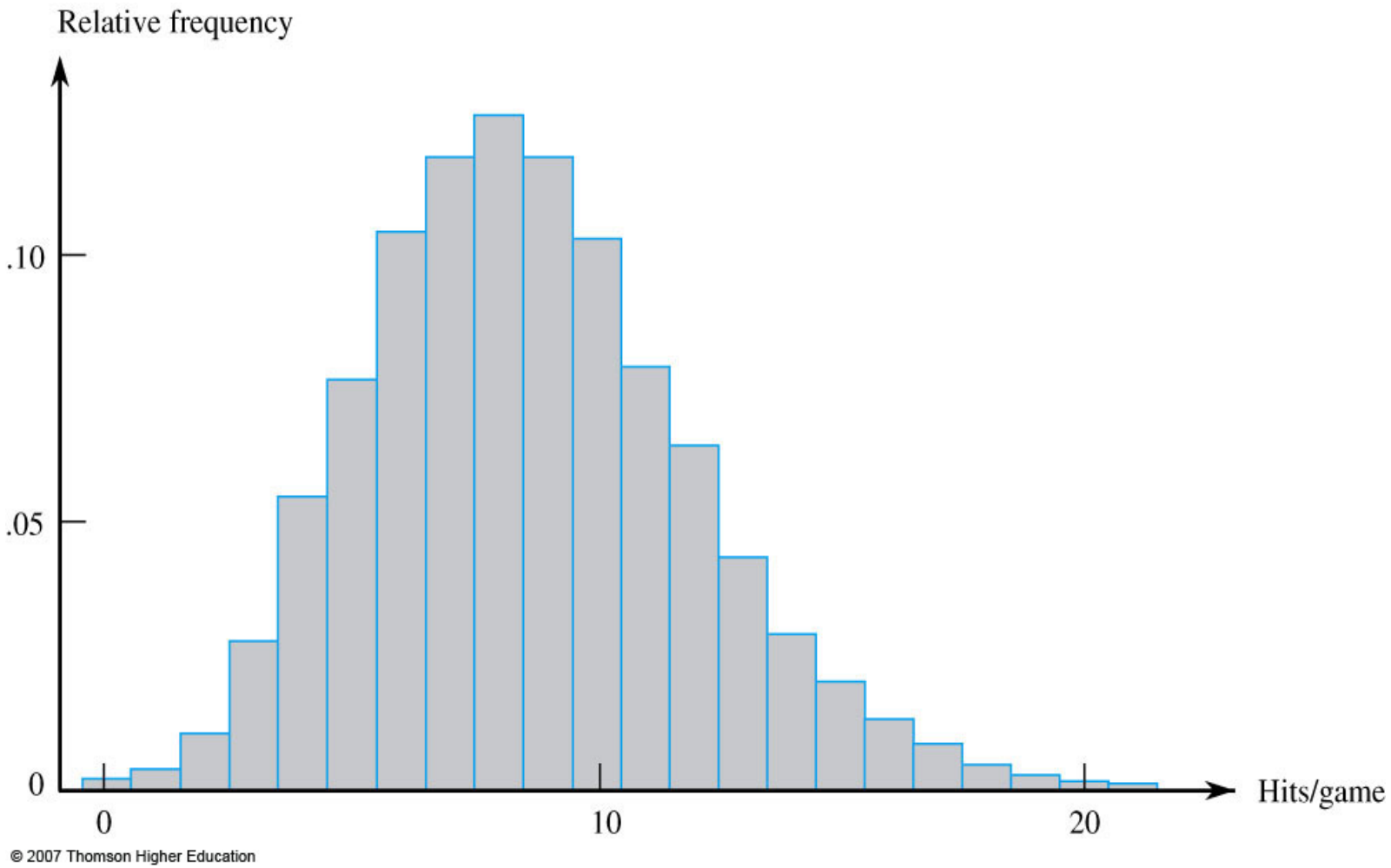
- Students from a small college were asked how many charge cards they carry. x is the variable representing the number of cards

x	# people	*Rel. Freq*
0	12	0.08
1	42	0.28
2	57	0.38
3	24	0.16
4	9	0.06
5	4	0.03
6	2	0.01

Example

- Number of hits per game for all nine-inning games that were played between 1989-1993 in a major league baseball
- The histogram seems to be unimodal but not symmetric

Figure :


Relative frequency
.10
.05
0
0
10
20
Hits/game
© 2007 Thomson Higher Education

Histograms for Continuous Data: Equal Class Widths

- The first step is splitting the data into a suitable number of **class intervals**
- As an example, let the variable of interest be the fuel efficiency of an automobile in mpg; the smallest observation is 27.8 and the largest is 31.4
- Commonly, an observation on the boundary placed in the interval to the *right* of the boundary: $27.5- < 28.0$, $28.0- < 28.5$, $28.5- < 29.0$ etc.
- Determine the (relative) frequency for each class. Then, above each class interval, draw a rectangle whose height is the (relative) frequency.

Histograms for continuous data with equal class widths: an example

- The data is a sample of adjusted consumption value during a particular period for 90 gas-heated homes in Wisconsin (in BTU's)
- The adjusted consumption is determined as the ratio:

$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather in degree days})\ (\text{house area})}$$

- The general rule of thumb is that the number of classes is approximately $\sqrt{\text{number of observations}}$.

Histograms for continuous data with equal class widths: an example

Figure :

Class	1−<3	3−<5	5−<7	7−<9	9−<11	11−<13	13−<15	15−<17	17−<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

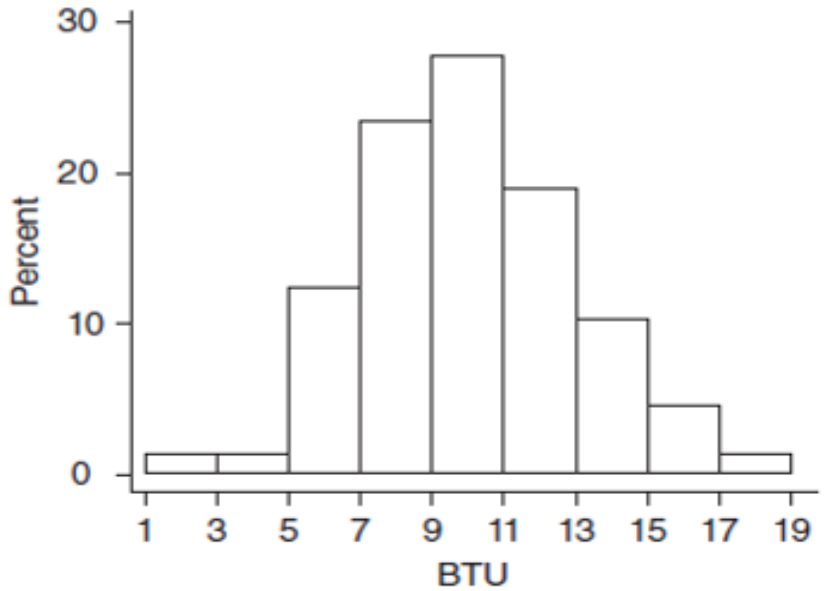


Figure 1.8 Histogram of the energy consumption data from Example 1.10

Histograms for Continuous Data: Unequal Widths

- Sometimes, unequal widths are called for, especially if the data is not uniformly concentrated over the range
- After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula:

$$\text{rectangle height} = \frac{\text{rel. frequency of the class}}{\text{class width}}$$

- The resulting heights are called *densities*. The vertical scale is called the *density scale*.

Histograms for continuous data with unequal class widths: an example

- The problem is the corrosion of the reinforcing steel...
- The data consists of 48 observations on measured bond strength (used for bonding glass-fiber-reinforced plastic rebars to concrete)
- The total area of all rectangles in a histogram drawn to **density scale** is always 1

The data

Figure :

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Class	2–<4	4–<6	6–<8	8–<12	12–<20	20–<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

Histograms for continuous data with unequal class widths: an example

Figure :

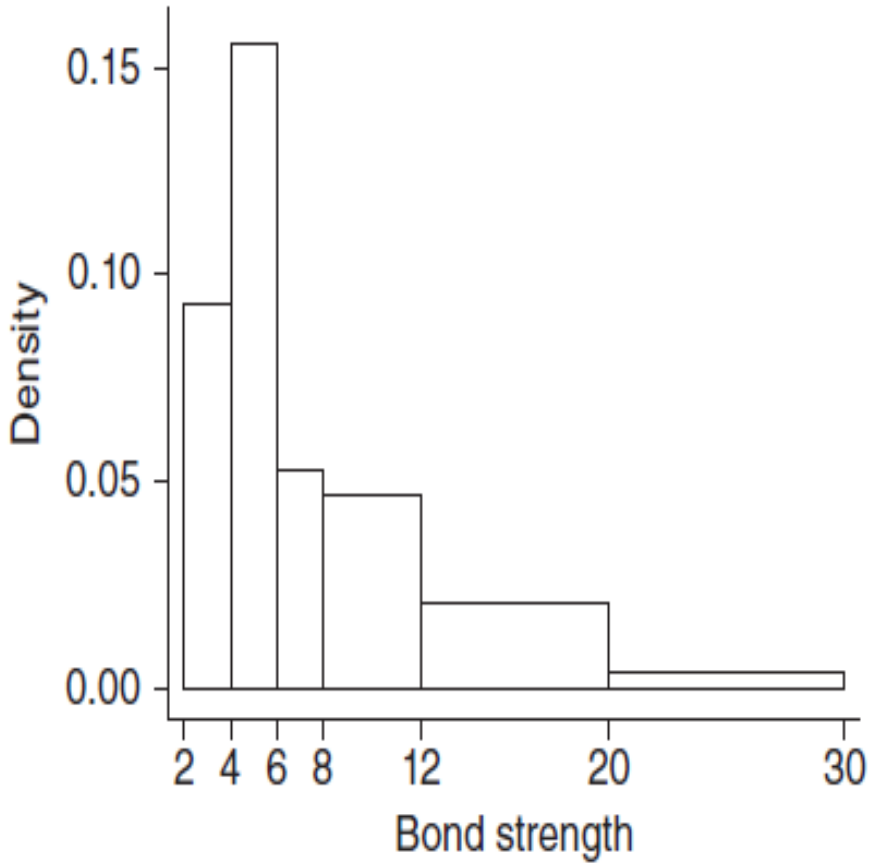
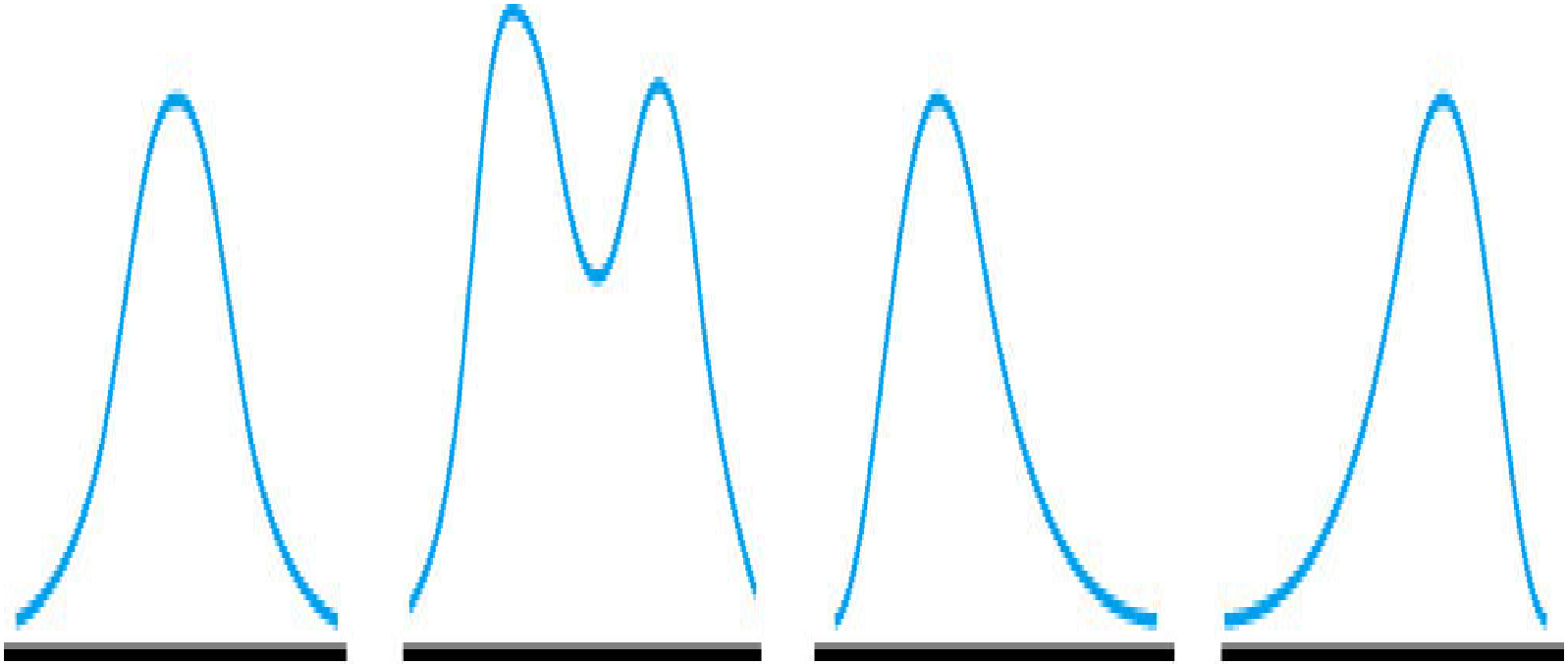


Figure 1.10 A Minitab density histogram for the bond strength data of Example 1.11

Histogram shapes

- Symmetric unimodal
- Bimodal
- Right-skewed
- Left-skewed

Figure :



(a) (b) (c) (d)

Sample mean

- The **sample mean** of the n numbers $x_1, \ldots, x_n$ is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Alternative notation:

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

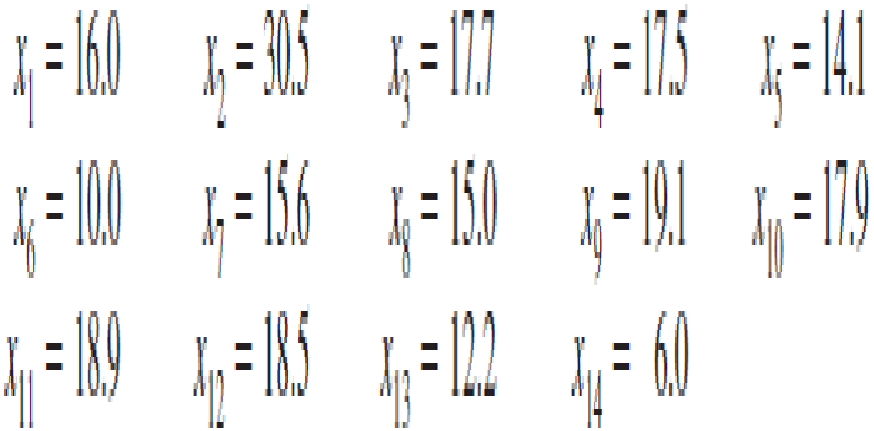
Example

- Recent years have seen growing commercial interest in the use of what is known as internally cured concrete.
- This concrete contains porous inclusions most commonly in the form of lightweight aggregate (LWA).
- The article Characterizing Lightweight Aggregate Desorption at High Relative Humidities Using a Pressure Plate Apparatus (J. of Materials in Civil Engr, 2012: 961969) reported on a study in which researchers examined various physical properties of 14 LWA specimens

Example

- Here are the 24-hour water-absorption percentages for the specimens

Figure :



- With the sum total $\sum x_i = 229.0$, the sample mean is

$$\bar{x} = \frac{229.0}{14} = 16.36$$

The sample mean is not robust!

- The mean value can be greatly affected by the presence of even a single outlier (unusually large or small observation).
- If a sample of employees contains nine who earn $50,000$ per year and one whose yearly salary is $150,000$, the sample mean salary is $60,000$; this value certainly does not seem representative of the data.

Sample median

- The **sample median** $\tilde{x}$ is the middle value in the set of data that has been arranged in ascending order. For an even number of data points the median is the average of the middle two.
- More precisely, suppose the number of observations n is odd. Then, the median $\tilde{x}$ is the observation number $\frac{n+1}{2}$.
- In the same way, if n is even, the median is defined as the average of $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th observations
- Median is a **robust** measure of the data center, unlike mean.
- The mean and the median are generally *not* the same

Median calculation example

- The following data give the concentration for a specific receptor for a sample of women with evidence of iron-deficiency anemia When ordered, they are 7.6 8.3 9.3 9.4 9.4 9.7 10.4 11.5 11.9 15.2 16.2 20.4 .
- As $n = 12$ is even, the median is $\frac{9.7+10.4}{2} = 10.05$.
- What would happen if the largest observation 20.4 was not there?

Population mean and median

- The **population mean** is defined as the sum of the N population values divided by N
- The sample mean is commonly used as a **point estimate** of the population mean
- The **population median** $\tilde{\mu}$ is defined as the "middle value" (in the same way as before) for the entire population. Again, sample median is commonly used as a point estimate of the population median.

Trimmed mean

- Often the median and the mean are just two extremes...How do we claim the middle ground? Consider the **trimmed mean.**
- The mean does not discard any observations while the median discards almost everything. We can discard a predetermined number or percentage of observations as an alternative.
- The following dataset gives the copper percentages in a sample of 26 Bidri artifacts

2.0	2.4	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	5.3	10.1

- The regular mean is $\bar{x} = 3.65$ and the median is $\tilde{x} = 3.35$. The difference is due to a large observation 10.1%

Trimmed mean

- 7.7% trimmed mean is the result of removing 2 smallest and 2 largest observations - $\bar{x}_{\text{tr}(7.7)} = 3.42$.
- The 10% trimmed mean is an appropriate weighted average of 7.7% trimmed mean (trimming two values at each end) and 11.5% trimmed mean (trimming three values at each end)

Measures of spread

- Variance measures the spread of the data
- The sample variance is

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

- The sample standard deviation is $s = \sqrt{s^2}$ and $n-1$ is referred to as the number of **degrees of freedom**

Example

- Consider the prefabricated plate example; there are $n = 11$ plate elements that have been subjected to a severe stress test. If x is the length of resulting cracks, $\sum_{i=1}^{n} x_i = 18.349$ and $\bar{x} = \frac{18.349}{11} = 1.6681$. Thus,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{10} = \frac{11.9359}{10} = 1.19359.$$

Computing the sample variance

- An alternative computing formula for the s^2 is based on the fact that

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

- Thus, we can write

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

- This formula needs to be used with the largest decimal accuracy possible

Some properties of standard deviation

- Let $x_1, \ldots, x_n$ be a sample of n observations and c any non-zero constant. Denote s_x^2 the sample variance of x's.
- 1. If $y_1 = x_1 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$
  2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$.

The fourth spread

- To define an alternative, again order n observations in a data set from smallest to largest. Then, the lower (upper) fourth is the median of the smallest (largest) half of the data; where the median is included in both halves if n is odd.
- Then, the **fourth spread** is defined as

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

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

Boxplots

- The simplest **boxplot** is a five-number summary : 1) minimum 2) lower fourth 3)median 4) upper fourth 5) maximum
- The "whiskers" mark the location of the smallest and the largest observations

Example

- The data consists of observations on the time until failure (1000s of hours) for a sample of turbochargers from one type of engine
- The five-number summary is as follows. smallest: 1.6 lower fourth: 5.05 median: 6.5 upper fourth: 7.85 largest: 9.0
- The plot indicates that there is a reasonable amount of symmetry in the middle 50% of the data, but overall values stretch out more toward the low end than toward the high end a negative skew.

Boxplot example

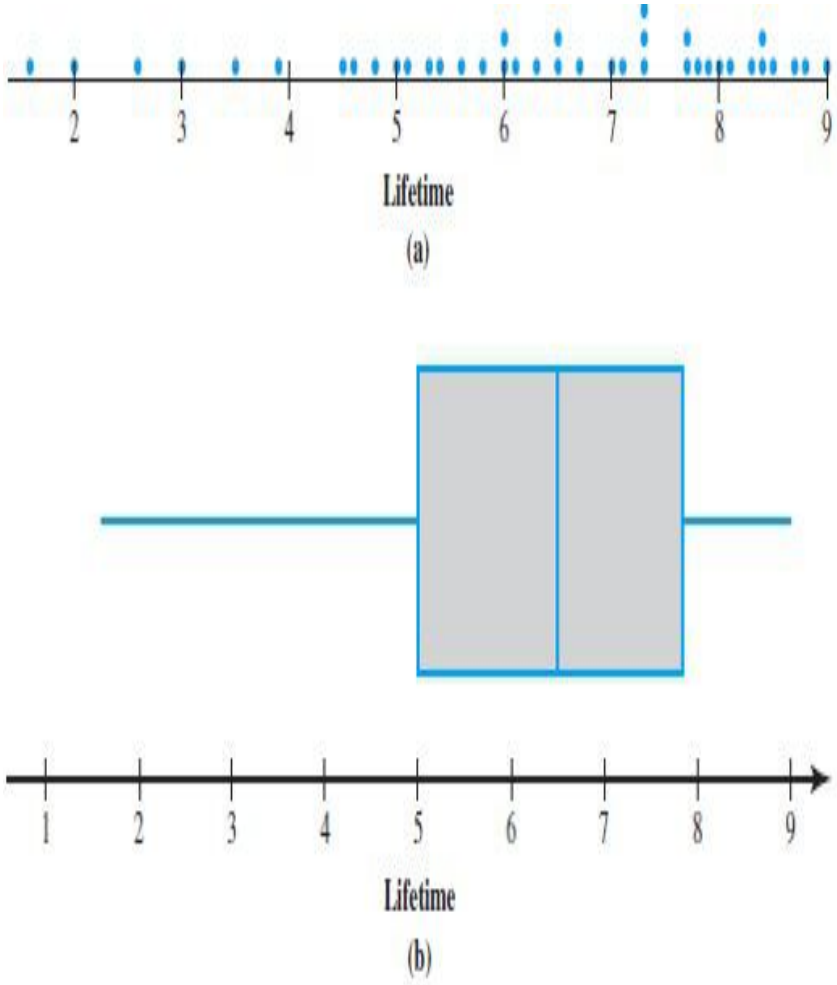


Figure 1.20 (a) Dotplot and (b) Boxplot for the lifetime data

Boxplots with outliers

- Many inferential procedures are based on the assumption that the population distribution is normal (a certain type of bell curve)
- Even a single extreme outlier in the sample warns the investigator that such procedures may be unreliable, and the presence of several mild outliers conveys the same message
- Typically, we represent each mild outlier by a closed circle and each extreme outlier by an open circle

Example

- The article Spurious Correlation in the USEPA Rating Curve Method for Estimating Pollutant Loads (J. of Environ. Engr., 2008: 610618) investigated various techniques for estimating pollutant loads in watersheds; the authors "discuss the imperative need to use sound statistical methods" for this purpose
- The data consists of the sample of TN (total nitrogen) loads (kg N/day) from a particular Chesapeake Bay location
- The whiskers extend out to the smallest observation, 9.69, on the left, and 312.45, the largest non-outlier observation, on the right
- There is some positive skewness in the middle half of the data (the median line is somewhat closer to the left edge of the box than to the right edge) and a great deal of positive skewness overall.

Boxplot example