7.2.2: Descriptive versus Inferential Statistics

Now that we understand the nature of our data, let’s turn to the types of statistics we can use to interpret them. There are 2 types of statistics: descriptive and inferential.

Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word “data” refers to the information that has been collected from an experiment, a survey, an historical record, etc. (By the way, “data” is plural. One piece of information is called a “datum.”) If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time to give a full picture of the data. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics, which you'll be studying in another section. Here we focus on (mere) descriptive statistics. Some descriptive statistics are shown in Table $7.2.2.1$ The table shows the average salaries for various occupations in the United States in 1999.

Table $7.2.2.1$: Average salaries for various occupations in 1999.

Occupation	Salary
Pediatricians	$112,760
Dentists	$106,130
Podiatrists	$100,090
Physicists	$76,140
Architects	$53,410
School, clinical, and counseling psychologists	$49,720
Flight attendants	$47,910
Elementary school teachers	$39,560
Police officers	$38,710
Floral designers	$18,980

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table $7.2.2.2$ It shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table $7.2.2.2$: Number of Unmarried Men per 100 Unmarried Women in U.S. Metro Areas in 1990.

Cities with Mostly Men	Men per 100 Women	Cities with Mostly Women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69
4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70
6. State College, PA	113	6. Gadsden, AL	70

Cities with Mostly Men	Men per 100 Women	Cities with Mostly Women	Men per 100 Women
7. ClarksvilleHopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-SeasideMonterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) The following table shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Table $7.2.2.3$: Winning Olympic Marathon Times for Women.

Year	Winner	Country	Time
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20

Table $7.2.2.4$shows the same statistics, but for men.

Table $7.2.2.4$: Winning Olympic Marathon Times for Men

Year	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50
1900	Michel Theato	FRA	2:59:45
1904	Thomas Hicks	USA	3:28:53
1906	Billy Sherring	CAN	2:51:23
1908	Johnny Hayes	USA	2:55:18
1912	Kenneth McArthur	S. Afr.	2:36:54
1920	Hannes Kolehmainen	FIN	2:32:35
1924	Albin Stenroos	FIN	2:41:22
1928	Boughra El Ouafi	FRA	2:32:57
1932	Juan Carlos Zabala	ARG	2:31:36

Year	Winner	Country	Time
1936	Sohn Kee-Chung	JPN	2:29:19
1948	Delfo Cabrera	ARG	2:34:51
1952	Emil Ztopek	CZE	2:23:03
1956	Alain Mimoun	FRA	2:25:00
1960	Abebe Bikila	ETH	2:15:16
1964	Abebe Bikila	ETH	2:12:11
1968	Mamo Wolde	ETH	2:20:26
1972	Frank Shorter	USA	2:12:19
1976	Waldemar Cierpinski	E.Ger	2:09:55
1980	Waldemar Cierpinski	E.Ger	2:11:03
1984	Carlos Lopes	POR	2:09:21
1988	Gelindo Bordin	ITA	2:10:32
1992	Hwang Young-Cho	S. Kor	2:13:23
1996	Josia Thugwane	S. Afr.	2:12:36
2000	Gezahenge Abera	ETH	2:10.10
2004	Stefano Baldini	ITA	2:10:55

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written 2:44:22). The mean winning time for the second 13 races is 2:13:18. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can't answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table $7.2.2.3$and Table $7.2.2.4$leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how far they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

It is also important to differentiate what we use to describe populations vs what we use to describe samples. A population is described by a parameter; the parameter is the true value of the descriptive in the population, but one that we can never know for sure. For example, the Bureau of Labor Statistics reports that the average hourly wage of chefs is $23.87. However, even if this number was computed using information from every single chef in the United States (making it a parameter), it would quickly become slightly off as one chef retires and a new chef enters the job market. Additionally, as noted above, there is virtually no way to collect data from every single person in a population. In order to understand a variable, we estimate the population parameter using a sample statistic. Here, the term "statistic" refers to the specific number we compute from the data (e.g. the average), not the field of statistics. A sample statistic is an estimate of the true population parameter, and if our sample is representative of the population, then the statistic is considered to be a good estimator of the parameter.

Even the best sample will be somewhat off from the full population, earlier referred to as sampling bias, and as a result, there will always be a tiny discrepancy between the parameter and the statistic we use to estimate it. This difference is known as sampling error, and, as we will see throughout the course, understanding sampling error is the key to understanding statistics. Every observation we make about a variable, be it a full research study or observing an individual's behavior, is incapable of being completely representative of all possibilities for that variable. Knowing where to draw the line between an unusual observation and a true difference is what statistics is all about.

Inferential Statistics

Descriptive statistics are wonderful at telling us what our data look like. However, what we often want to understand is how our data behave. What variables are related to other variables? Under what conditions will the value of a variable change? Are two groups different from each other, and if so, are people within each group different or similar? These are the questions answered by inferential statistics, and inferential statistics are how we generalize from our sample back up to our population. Units 2 and 3 are all about inferential statistics, the formal analyses and tests we run to make conclusions about our data.

For example, we will learn how to use a t statistic to determine whether people change over time when enrolled in an intervention. We will also use an F statistic to determine if we can predict future values on a variable based on current known values of a variable. There are many types of inferential statistics, each allowing us insight into a different behavior of the data we collect. This course will only touch on a small subset (or a sample) of them, but the principles we learn along the way will make it easier to learn new tests, as most inferential statistics follow the same structure and format.

Summary

In simpler terms, we *infer* characteristics of the population based on characteristics of the sample.

Definition: Descriptive Statistics

Used to describe or summarize the data from the sample.

Definition: Inferential Statistics

Used to make generalizations from the sample data to the population of interest.

Let's practice!

Example 7.2.2.1

Use the following to decide which option is **inferential**.

- Target Population -- All Psychology majors in the U.S.
- Sample -- 30 students from a Research Methods course
- Data -- 18 want to become Clinical Psychologists (60%)

Which of the following statements is descriptive of the sample and which is making an inference about the population? *Why?*

1. 60% of American Psychology majors want to be clinical psychologists.
2. 60% of the students in the sample want to be clinical psychologists.

Solution

#1 is inferential because it is using the information from the sample of one class to infer about all Psychology majors in the U.S..

Your turn!

Exercise 7.2.2.1

Use the following to decide which option is **inferential**.

- Target Population -- California college students
- Sample -- 300 students from all 115 California community colleges
- Data – 150 are from the central valley, 150 are from outside of the valley

Which of the following statements is descriptive of the sample and which is making an inference about the population? *Why?*

1. 50% of California community college students are from the central valley.
2. 50% of the students in the sample are from the central valley.

Do you see anything wrong with this data?

Answer

#1 is inferential because it is using the information from the sample of 300 students to infer about all California college students.

If you know anything about the geography of California, you know that 50% of the state's population does not live in the central valley, so the sampling is suspect.