2.9: Types of Statistical Analyses

2.9.1 Types of Statistical Analyses

Now that we understand the nature of our data, let's turn to the types of statistics we can use to interpret them. As mentioned at the end of chapter 1, there are 2 types of statistics: descriptive and inferential.

2.9.1.0.1 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 2. The table shows the average salaries for various occupations in the United States in 1999.

Salary 1999	Salary 2019	Occupation
$112,760	$175,310	pediatricians
$106,130	$155,600	dentists
$100,090	$126,240	podiatrists
$76,140	$97,152	physicists
$53,410	$80,750	architects
$49,720	$78,200	school, clinical, and counseling psychologists
$47,910	$56,640	flight attendants
$39,560	$59,670	elementary school teachers
$38,710	$65,170	police officers
$18,980	$28, 040	floral designers

Table 2. Average salaries for various occupations in 1999 and 2019 (median salaries reported by Bureau of Labor Statistics).

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 3. It shows the number of employed single young men to single young women for large metro areas in the US (reported in 2014). From this table we see that men outnumber women most in the San Jose, CA area, and women outnumber men most in the Memphis, TN area. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner between the ages of 25-34 years old! (These data come from Pew Research)

Highest Ratios of Employed Single Men to Single Women (25-34 y/o)	Men per 100 Women	Lowest Ratios of Employed Single Men to Single Women (25-34 y/o)	Men per 100 Women
1. San-Jose-Sunnyvale-Santa Clara, CA	114	1. Memphis, TN-MS-AR	59

2. Denver-Aurora-Lakewood, CO	101	2. Jacksonville, FL	70
3. San Diego-Carlsbad, CA	99	3. Detroit-Warren-Dearborn, MI	71
4. Minneapolis-St. Paul-Bloomington, MN-WI	98	4. Charlotte-Concord-Gastonia, NC-SC	73
5. Seattle-Tacoma-Bellevue, WA	96	5. Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	74
6. San Francisco-Oakland-Hayward, CA	93	6. Kansas City, MO-KS	75
7. Washington-Arlington-Alexandra, DC-VA-MD-WV	92	7. Nashville-Davidson-Murfreesboro-Franklin, TN	77
8. Los Angeles-Long Beach-Anaheim, CA	91	8. Miami-Fort Lauderdale-West-Palm Beach, FL	78
9. Pittsburgh, PA	90	9. New Orleans-Metairie, LA	78
10. Orlando-Kissimmee-Sanford, FL	90	10. Cincinnati, OH-KY-IN	78

Table 3. Number of employed, 25-34 year old ratio of men to women in large metro areas of the U.S. (Pew Research, 2014)

These descriptive statistics may make us ponder why there are ratio differences in these large metropolitan areas. 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).

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
2008	Constantina Dita-Tomescu	Romania	2:26:44
2012	Tiki Gelana	ETH	2:23:07
2016	Jemima Sumgong	Kenya	2:24:04

2020	Peres Jepchirchir	Kenya	2:27:20
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
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
2008	Samuel Wanjiru	Kenya	2:06:32
2012	Stephen Kiprotich	Uganda	2:08:01
2016	Eliud Kipchoge	Kenya	2:08:44
2020	Eliud Kipchoge	Kenya	2:08:38

Table 4. Winning Olympic marathon times.

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 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 or head cooks is $25.66[1]. 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.

2.9.1.0.1 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.