Chapter 6: Point Estimation

6.1 Descriptive Statistics

In Chapter 2, we considered probability distributions of random variables whose space ‘S‘ contains a countable number of outcomes: either a finite number of outcomes or outcomes that can be put into a one-to-one correspondence with the positive integers. Such a random variable is said to be of the discrete type, and its distribution of probabilities is of the **discrete type**.

Example 1. Examples for discrete type of data:

1.
2.
3.
4.

If, conceptually, the measurements could come from an interval of possible outcomes, we call them data from a distribution of the **continuous type** or, more simply, continuous-type data.

Example 2. Examples for continuous type of data:

1.
2.
3.
4.

Given a set of continuous-type data, we shall group the data into classes and then construct a **histogram** of the grouped data. This will help us better visualize the data. The following guidelines and terminology will be used to group continuous-type data into classes of equal length (these guidelines can also be used for sets of discrete data that have a large range).

How to make a Histogram:

1. Determine the ________________ and ________________ observations. The ________________ is the difference, ________________.

2. In general, select from ________________ to ________________ classes, which are non overlapping intervals, usually of equal length. These classes should cover the interval from the ________________ to the ________________.

3. Each interval begins and ends halfway between two possible values of the measurements, which have been rounded off to a given number of decimal places.

4. The first interval should begin about as much below the smallest value as the last interval ends above the largest.

5. The intervals are called ________________ and the boundaries are called ________________. We shall denote these k class intervals by ________________.

6. The ________________ are the smallest and the largest possible observed (recorded) values in a class.

7. The class mark is the ________________ of a class.

Frequency Table:

- A frequency table is constructed that lists the class intervals, the class limits, a tabulation of the measurements in the various classes, the frequency f_i of each class, and the class marks.
- A column is sometimes used to construct a relative frequency (density) histogram.
- With class intervals of equal length, a frequency histogram is constructed by drawing, for each class, a rectangle having as its base the class interval and a height equal to the frequency of the class.

– For the relative frequency histogram, each rectangle has an area equal to the relative frequency $\frac{f_i}{n}$ of the observations for the class.

That is, the function defined by

Problem 1. The weights in grams of 40 miniature Baby Ruth candy bars, with the weights ordered, are given below. Create a frequency table and a relative frequency histogram for the following data.

Data: Candy Bar Weights

20.5, 20.7, 20.8, 21.0, 21.0, 21.4, 21.5, 22.0, 22.1, 22.5

22.6, 22.6, 22.7, 22.7, 22.9, 22.9, 23.1, 23.3, 23.4, 23.5

23.6, 23.6, 23.6, 23.9, 24.1, 24.3, 24.5, 24.5, 24.8 ,24.8

24.9, 24.9, 25.1, 25.1, 25.2, 25.6, 25.8, 25.9, 26.1, 26.7

Empirical Distribution Suppose that we now consider the situation in which we actually perform a certain random experiment n times, obtaining n observed values of the random variable—say, $x_1, x_2, ..., x_n$.

Often the collection is referred to as a________________. It is possible that some of these values might be the same, but we do not worry about this at this time.

We artificially create a________________by placing the weight ________________ on each of these x-values.

Note that these weights are________________and sum to ________________, so we have a distribution we call the ________________, since it is determined by the data $x_1, x_2, ..., x_n$..

Mean, Variance and Standard Deviation The mean of the empirical distribution is which is the arithmetic mean of the observations $x_1, x_2, ..., x_n$.

We denote this mean by________________ and call it the sample mean (or mean of the sample $x_1, x_2, ..., x_n$). That is, the sample mean is

which is, in some sense, an estimate of________________if the latter is unknown.

The Sample variance is

The Sample Standard Deviation,

is a measure of ____.

Problem 2. Rolling a fair six-sided die five times could result in the following sample of n = 5 observations: $x_1 = 3, x_2 = 1, x_3 = 2, x_4 = 6, x_5 = 3$.

Find Mean, Variance and Standard deviation for this data. Explain your findings.

Note: If the histogram is "mound-shaped" or "bell- shaped," the following empirical rule gives rough approximations to the percentages of the data that fall between certain points. These percentages clearly are associated with the ________________________.

Empirical Rule: Let $x_1, x_2, \ldots, x_n$ have a sample mean x and sample standard deviation s. If the histogram of these data is "bell-shaped," then, for large samples,

How to make a relative frequency polygon?

1. Mark the midpoints at the top of each "bar" of the histogram.
2. Connect adjacent midpoints with straightline segments.
3. On each of the two end bars, draw a line segment from the top middle mark through the middle point of the outer vertical line of the bar.
4. Of course, if the area underneath the tops of the relative frequency histogram is equal to 1, which it should be, then the area underneath the relative frequency polygon is also equal to 1, because the areas lost and gained cancel out by a consideration of congruent triangles.

Problem 3: A manufacturer of fluoride toothpaste regularly measures the concentration of fluoride in the toothpaste to make sure that it is within the specification of 0.85 to 1.10 mg/g. Following table lists 100 such measurements.

0.98	0.92	0.89	0.90	0.94	0.99	0.86	0.85	1.06	1.01
1.03	0.85	0.95	0.90	1.03	0.87	1.02	0.88	0.92	0.88
0.88	0.90	0.98	0.96	0.98	0.93	0.98	0.92	1.00	0.95
0.88	0.90	1.01	0.98	0.85	0.91	0.95	1.01	0.88	0.89
0.99	0.95	0.90	0.88	0.92	0.89	0.90	0.95	0.93	0.96
0.93	0.91	0.92	0.86	0.87	0.91	0.89	0.93	0.93	0.95
0.92	0.88	0.87	0.98	0.98	0.91	0.93	1.00	0.90	0.93
0.89	0.97	0.98	0.91	0.88	0.89	1.00	0.93	0.92	0.97
0.97	0.91	0.85	0.92	0.87	0.86	0.91	0.92	0.95	0.97
0.88	1.05	0.91	0.89	0.92	0.94	0.90	1.00	0.90	0.93

Find the frequency table and draw the relative frequency histogram with relative frequency polygon.

Approximate mean and standard deviation

NOTE: In some situations, it is not necessarily desirable to use class intervals of **equal widths** in the construction of the frequency distribution and histogram. This is particularly true if the data are skewed with a very long tail.

Problem 4: The following 40 losses, due to wind-related catastrophes, were recorded to the near- est $1 million (these data include only losses of $2 million or more; for convenience,

they have been ordered and recorded in millions):

2	2	2	2	2	2	2	2	2	2
2	2	3	3	3	3	4	4	4	5
5	5	5	6	6	6	6	8	8	9
15	17	22	23	24	24	25	27	32	43

Find $h(x)$ and graph it.

STATISTICAL COMMENTS (Simpson's Paradox)

The relative frequency, $\frac{f}{n}$, is called a ________________ and is used to estimate a ________________, p, which is usually unknown.

Example, if a major league batter gets f = 152 hits in n = 500 official at bats during the season, then the relative frequency ________________ is an estimate of his ________________ and is called his batting ________________

Problem 5: Once while speaking to a group of coaches, one of the coaches made the comment that it would be possible for batter A to have a higher average than batter B for each season during their careers and yet B could have a better overall average at the end of their careers. While no coach spoke up, you could tell that they were thinking, "And that guy is supposed to know something about math."

Of course, the following simple example convinced them that the statement was true:

Suppose A and B played only two seasons, with these results:

	Player *A*			Player *B*		
Season	AB	Hits	Average	AB	Hits	Average
1	500	126	0.252	300	75	0.250
2	300	90	0.300	500	145	0.290
Totals	800	216	0.270	800	220	0.275