Introduction to Elementary Statistics

2.2: Frequency Distributions and Histograms

- Stem-and-leaf plots often present adequate summaries, but they can get very big, very fast.

- Need other techniques for summarizing data.

- Frequency distributions and histograms are used to summarize large data sets.

Frequency Distribution: A listing, often expressed in chart form, that pairs each value of a variable with its frequency.

Ungrouped Frequency Distribution: Each value of *x* in the distribution stands alone.

Grouped Frequency Distribution: Group the values into a set of classes.

1. A table that summarizes data by classes, or class intervals.
2. In a typical *grouped* frequency distribution, there are usually 5-12 classes of equal width.
3. The table may contain columns for class number, class interval, tally (if constructing by hand), frequency, relative frequency, cumulative relative frequency, and class mark.
4. In an *ungrouped* frequency distribution each class consists of a single value.

Guidelines for constructing a frequency distribution:

1. Each class should be of the same width.

2. Classes should be set up so that they do not overlap and so that each piece of data belongs to exactly one class.

3. For problems in the text, 5-12 classes are most desirable. The square root of *n* is a reasonable guideline for the number of classes if *n* is less than 150.

4. Use a system that takes advantage of a number pattern, to guarantee accuracy.

5. If possible, an even class width is often advantageous.

Procedure for constructing a frequency distribution:

1. Identify the high (H) and low (L) scores. Find the range. Range = H - L.

2. Select a number of classes and a class width so that the product is a bit larger than the range.

3. Pick a starting point a little smaller than L. Count from L by the width to obtain the **class boundaries**. Observations that fall on class boundaries are placed into the class interval to the right.

Note:

1. The class width is the difference between the upper- and lower-class boundaries.

2. There is no *best* choice for class widths, number of classes, and starting points.

Example: The hemoglobin test, a blood test given to diabetics during their periodic checkups, indicates the level of control of blood sugar during the past two to three months. The data in the table below was obtained for 40 different diabetics at a university clinic that treats diabetic patients. Construct a grouped frequency distribution using the classes 3.7 - <4.7, 4.7 - <5.7, 5.7 - <6.7, etc. Which class has the highest frequency?

6.5	5.0	5.6	7.6	4.8	8.0	7.5	7.9	8.0	9.2
6.4	6.0	5.6	6.0	5.7	9.2	8.1	8.0	6.5	6.6
5.0	8.0	6.5	6.1	6.4	6.6	7.2	5.9	4.0	5.7
7.9	6.0	5.6	6.0	6.2	7.7	6.7	7.7	8.2	9.0

Solution:

Class	Frequency	Relative	Cumulative	Class
Boundaries	*f*	Frequency	Rel. Frequency	Mark, *x*
3.7 - <4.7	1	.025	.025	4.2
4.7 - <5.7	6	.150	.175	5.2
5.7 - <6.7	16	.400	.575	6.2
6.7 - <7.7	4	.100	.250	7.2
7.7 - <8.7	10	.250	.925	8.2
8.7 - <9.7	3	.075	1.000	9.2

The class 5.7 - <6.7 has the highest frequency. The frequency is 16 and the relative frequency is .40.

Histogram: A bar graph representing a frequency distribution of a quantitative variable. A histogram is made up of the following components:

1. A title, which identifies the population of interest.

2. A vertical scale, which identifies the frequencies in the various classes.

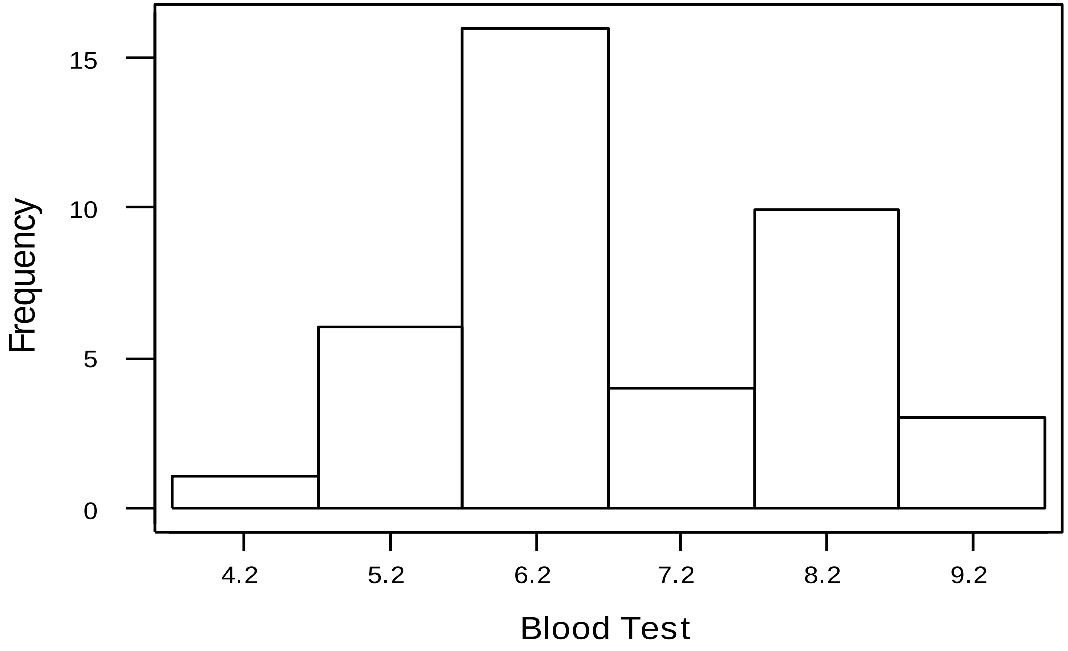
3. A horizontal scale, which identifies the variable *x*. Values for the class boundaries or class marks may be labeled along the *x*-axis. Use whichever method of labeling the axis best presents the variable.

Note:

1. The relative frequency is sometimes used on the vertical   scale.
2. It is possible to create a histogram based on class marks.

Example: Construct a histogram for the blood test results given in the previous example.

Solution:

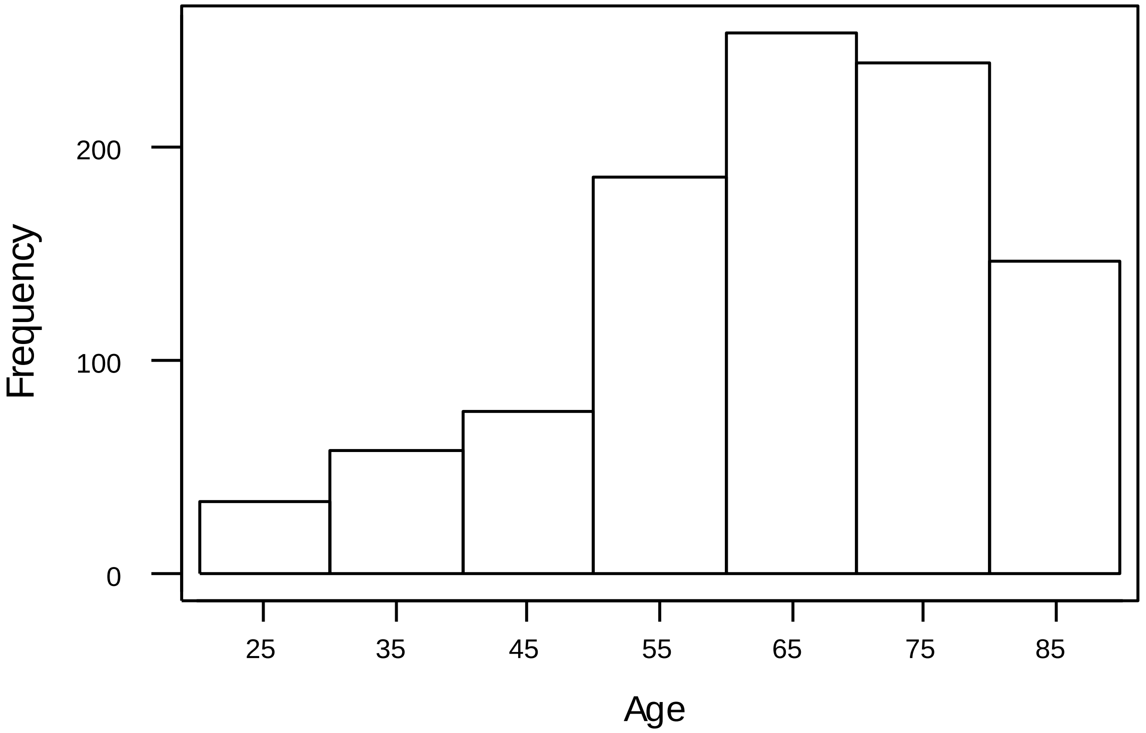


Example: A recent survey of Roman Catholic nuns summarized their ages in the table below.

Age	Frequency	Class Mark
20 up to 30	34	25
30 up to 40	58	35
40 up to 50	76	45
50 up to 60	187	55
60 up to 70	254	65
70 up to 80	241	75
80 up to 90	147	85

Construct a histogram for this age data.

Solution



Terms used to describe histograms:

Symmetrical: Both sides of the distribution are identical. There is a *line* of symmetry.

Uniform (rectangular): Every value appears with equal frequency.

Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail. (Positively skewed vs. negatively skewed)

J-shaped: There is no tail on the side of the class with the highest frequency.

Bimodal: The two largest classes are separated by one or more classes. Often implies two populations are sampled.

Normal: A symmetrical distribution is mounded about the mean and becomes sparse at the extremes.

Note:

1. The **mode** is the value that occurs with greatest frequency (discussed in Section 2.3).

2. The **modal class** is the class with the greatest frequency.

3. A **bimodal distribution** has two high-frequency classes separated by classes with lower frequencies.

4. Graphical representations of data should include a descriptive, meaningful title and proper identification of the vertical and horizontal scales.

2.3: Measures of Central Tendency

- Numerical values used to locate the middle of a set of data, or where the data is clustered.
- The term *average* is often associated with all measures of central tendency.

Mean: The type of average with which you are probably most familiar. The mean is the sum of all the values divided by the total number of values, *n*.

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

Note:

1. The population mean, μ, (lowercase mu, Greek alphabet), is the mean of all *x* values for the entire population.

2. We usually cannot measure μ but would like to estimate its value.

3. A physical representation: the mean is the value that balances the weights on the number

Example: The data below represents the number of accidents in each of the last 6 years at a dangerous intersection.

8, 9, 3, 5, 2, 6, 4, 5

Find the mean number of accidents.

Solution:

$$\bar{x} = \frac{1}{8}(8+9+3+5+2+6+4+5) = 5.25$$

Note: In the data above, change 6 to 26.

$$\bar{x} = \frac{1}{8}(8+9+3+5+2+26+4+5) = 7.75$$

The Mean can be greatly influenced by **outliers.**

Median: The value of the data that occupies the middle position when the data are ranked in order according to size.

Note:

1. Denoted by "*x* tilde" : $\tilde{x}$
2. The population median, M (uppercase mu, Greek alphabet), is the data value in the middle position of the entire population.

To find the median:

1. Rank the data.
2. Determine the *depth* of the median.

$$d(\tilde{x}) = \frac{n+1}{2}$$

3. Determine the value of the median.

Example: Find the median for the set of data
{4, 8, 3, 8, 2, 9, 2, 11, 3}.

Solution:

1. Rank the data: 2, 2, 3, 3, 4, 8, 8, 9, 11
2. Find the depth: $d(\tilde{x})=(9+1)/2=5$
3. The median is the fifth number from either end in the ranked data: $\tilde{x}=4$

Suppose the data set is {4, 8, 3, 8, 2, 9, 2, 11, 3, 15}.

1. Rank the data: 2, 2, 3, 3, 4, 8, 8, 9, 11, 15
2. Find the depth: $d(\tilde{x}) = (10+1)/2 = 5.5$
3. The median is halfway between the fifth and sixth observations:

Mode: The mode is the value of *x* that occurs most frequently.

Note: If two or more values in a sample are tied for the highest frequency (number of occurrences), there is **no mode**.

Midrange: The number exactly midway between a lowest value data *L* and a highest value data *H*. It is found by averaging the low and the high values.

$$\text{midrange} = \frac{L+H}{2}$$

Example: Consider the data set {12.7, 27.1, 35.6, 44.2, 18.0}.

The midrange is

$$\text{midrange} = \frac{L+H}{2} = \frac{12.7+44.2}{2} = 28.45$$

Note:

1. When rounding off an answer, a common rule-of-thumb is to keep one more decimal place in the answer than was present in the original data.
2. To avoid round-off buildup, round off only the final answer, not intermediate steps.