1.1: Descriptive Statistics

https://milnepublishing.geneseo.edu/...017/02/958.png

https://milnepublishing.geneseo.edu/...e35759_fmt.png

https://web.archive.org/web/20230328...e35759_fmt.png

A **population** is the group to be studied, and population data is a collection of **all** elements in the population. For example:

- All the fish in Long Lake.
- All the lakes in the Adirondack Park.
- All the grizzly bears in Yellowstone National Park.

A **sample** is a subset of data drawn from the population of interest. For example:

- 100 fish randomly sampled from Long Lake.
- 25 lakes randomly selected from the Adirondack Park.
- 60 grizzly bears with a home range in Yellowstone National Park.

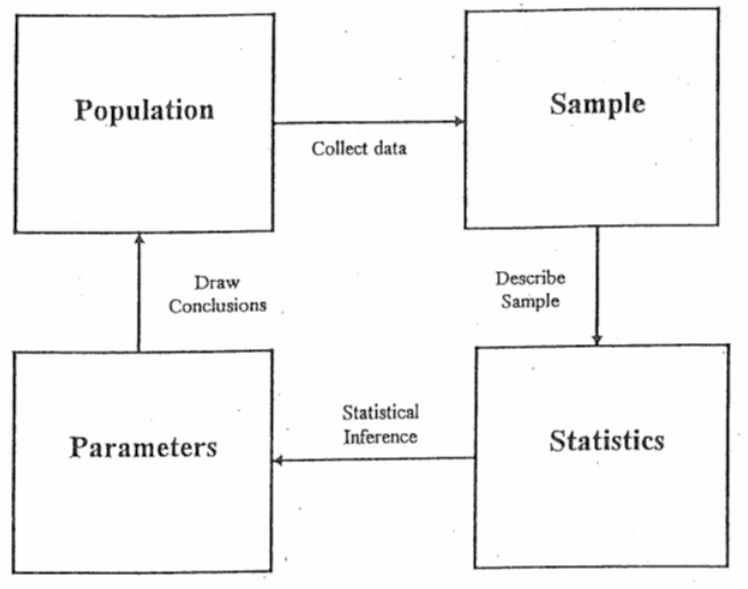


***Figure 1**. Using sample statistics to estimate population parameters.*

Populations are characterized by descriptive measures called parameters. Inferences about **parameters** are based on sample **statistics**. For example, the population mean (μ) is estimated by the sample mean ($\bar{x}$). The population variance (σ^2) is estimated by the sample variance (s^2).

Variables are the characteristics we are interested in. For example:

- The length of fish in Long Lake.
- The pH of lakes in the Adirondack Park.
- The weight of grizzly bears in Yellowstone National Park.

Variables are divided into two major groups: qualitative and quantitative. Qualitative variables have values that are attributes or categories. Mathematical operations cannot be applied to qualitative variables. Examples of qualitative variables are gender, race, and petal color. **Quantitative** variables have values that are typically numeric, such as measurements. Mathematical operations can be applied to these data. Examples of quantitative variables are age, height, and length. **Quantitative** variables can be broken down further into two more categories: discrete and continuous variables. Discrete variables have a finite or countable number of possible values. Think of discrete variables as “hens”. Hens can lay 1 egg, or 2 eggs, or 13 eggs… There are a limited, definable number of values that the variable could take on.

Continuous variables have an infinite number of possible values. Think of continuous variables as “cows”. Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk … There are an almost infinite number of values that a continuous variable could take on.

Example 1.1.1:

Is the variable qualitative or quantitative?

a. Species
b. Weight
c. Diameter
d. Zip Code

Solution

(qualitative quantitative, quantitative, qualitative)

Descriptive Measures

Descriptive measures of populations are called parameters and are typically written using Greek letters. The population mean is μ (mu). The population variance is σ^2 (sigma squared) and population standard deviation is σ (sigma). Descriptive measures of samples are called statistics and are typically written using Roman letters. The sample mean is $\bar{x}$ (x-bar). The sample variance is s^2 and the sample standard deviation is s. Sample statistics are used to estimate unknown population parameters. In this section, we will examine descriptive statistics in terms of measures of center and measures of dispersion. These descriptive statistics help us to identify the center and spread of the data.

Measures of Center

Mean

The arithmetic mean of a variable, often called the average, is computed by adding up all the values and dividing by the total number of values. The population mean is represented by the Greek letter μ (mu). The sample mean is represented by $\bar{x}$ (x-bar). The sample mean is usually the best, unbiased estimate of the population mean. However, the mean is influenced by extreme values (outliers) and may not be the best measure of center with strongly skewed data. The following equations compute the population mean and sample mean.

$$\mu = \frac{\sum x_i}{N} \tag{1.1.1}$$

$$\bar{x} = \frac{\sum x_i}{n} \tag{1.1.2}$$

where x_i is an element in the data set, N is the number of elements in the population, and n is the number of elements in the sample data set.

Example 1.1.2: mean

Find the mean for the following sample data set:

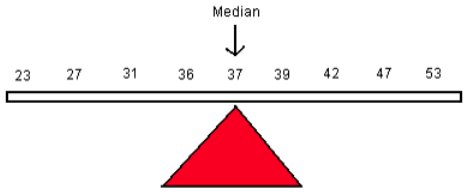
6.4, 5.2, 7.9, 3.4

Solution

$$\bar{x} = \frac{6.4 + 5.2 + 7.9 + 3.4}{4} = 5.725$$

Median

The median of a variable is the middle value of the data set when the data are sorted in order from least to greatest. It splits the data into two equal halves with 50% of the data below the median and 50% above the median. The median is resistant to the influence of outliers, and may be a better measure of center with strongly skewed data.



The calculation of the median depends on the number of observations in the data set.

To calculate the median with an odd number of values (*n* is odd), first sort the data from smallest to largest.

Example 1.1.3: Calculating Median with Odd number of values

Find the median for the following sample data set:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47, 51$$

Solution

The median is 39. It is the middle value that separates the lower 50% of the data from the upper 50% of the data.

To calculate the median with an even number of values (*n* is even), first sort the data from smallest to largest and take the average of the two middle values.

Example 1.1.4: Calculating Median with even number of values

Find the median for the following sample data set:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47$$

Solution

$$M = \frac{35 + 39}{2} = 37$$

Mode

The mode is the most frequently occurring value and is commonly used with qualitative data as the values are categorical. Categorical data cannot be added, subtracted, multiplied or divided, so the mean and median cannot be computed. The mode is less commonly used with quantitative data as a measure of center. Sometimes each value occurs only once and the mode will not be meaningful.

Understanding the relationship between the mean and median is important. It gives us insight into the distribution of the variable. For example, if the distribution is skewed right (positively skewed), the mean will increase to account for the few larger observations that pull the distribution to the right. The median will be less affected by these extreme large values, so in this

situation, the mean will be larger than the median. In a symmetric distribution, the mean, median, and mode will all be similar in value. If the distribution is skewed left (negatively skewed), the mean will decrease to account for the few smaller observations that pull the distribution to the left. Again, the median will be less affected by these extreme small observations, and in this situation, the mean will be less than the median.

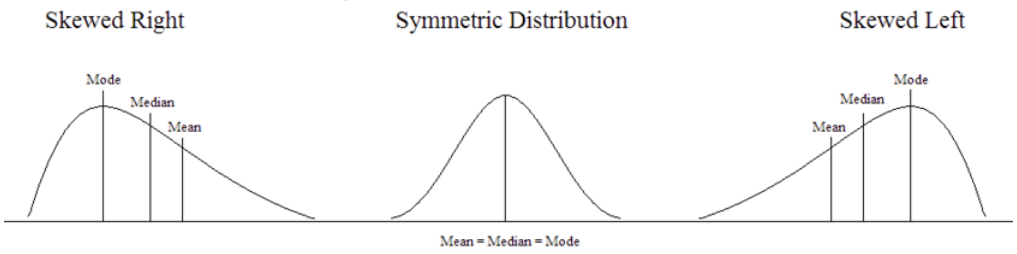


Figure 2. *Illustration of skewed and symmetric distributions.*

Measures of Dispersion

Measures of center look at the average or middle values of a data set. Measures of dispersion look at the spread or variation of the data. Variation refers to the amount that the values vary among themselves. Values in a data set that are relatively close to each other have lower measures of variation. Values that are spread farther apart have higher measures of variation.

Examine the two histograms below. Both groups have the same mean weight, but the values of Group A are more spread out compared to the values in Group B. Both groups have an average weight of 267 lb. but the weights of Group A are more variable.

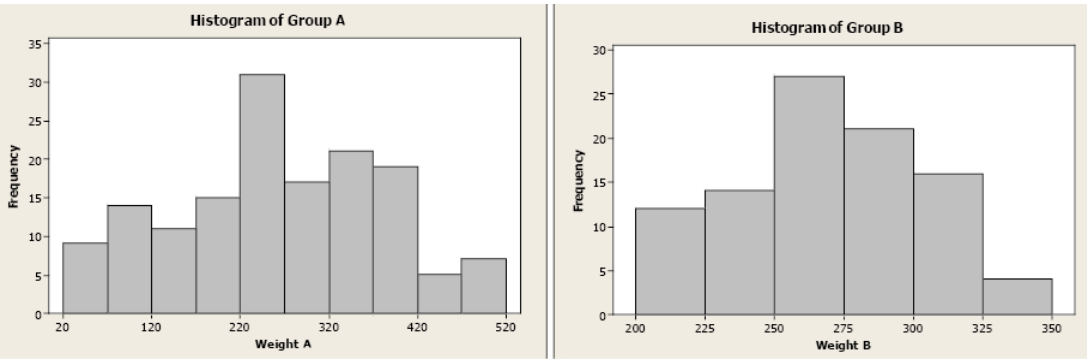


Figure 3. Histograms of Group A and Group B.

This section will examine five measures of dispersion: range, variance, standard deviation, standard error, and coefficient of variation.

Range

The range of a variable is the largest value minus the smallest value. It is the simplest measure and uses only these two values in a quantitative data set.

Example 1.1.5: Computing Range

Find the range for the given data set.

$$12, 29, 32, 34, 38, 49, 57$$

$$Range = 57 - 12 = 45$$

Variance

The variance uses the difference between each value and its arithmetic mean. The differences are squared to deal with positive and negative differences. The sample variance (s^2) is an unbiased estimator of the population variance (σ^2), with n-1 degrees of freedom.

Degrees of freedom: In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question.

The sample variance is unbiased due to the difference in the denominator. If we used “n” in the denominator instead of “n – 1”, we would consistently underestimate the true population variance. To correct this bias, the denominator is modified to “n – 1”.

Definition: population variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \tag{1.1.3}$$

Definition: sample variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} \tag{1.1.4}$$

Example 1.1.6: Computing Variance

Compute the variance of the sample data: 3, 5, 7.

Solution

The sample mean ($\bar{x}$) is 5. Then use Equation 1.1.4

$$s^2 = \frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation

The standard deviation is the square root of the variance (both population and sample). While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared. The standard deviation is a common method for numerically describing the distribution of a variable. The population standard deviation is σ (sigma) and sample standard deviation is *s*.

Definition:SAMPLE STANDARD DEVIATION

$$s = \sqrt{s^2} \tag{1.1.5}$$

Definition:POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\sigma^2} \tag{1.1.6}$$

Example 1.1.7:

Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

Solution

The sample mean ($\bar{x}$) is 5, using the definition of standard deviation

$$s = \sqrt{\frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1}} = \sqrt{4} = 2$$

Standard Error of Mean

Commonly, we use the sample mean $\bar{x}$ to estimate the population mean μ. For example, if we want to estimate the heights of eighty-year-old cherry trees, we can proceed as follows:

- Randomly select 100 trees
- Compute the sample mean of the 100 heights
- Use that as our estimate

We want to use this sample mean to estimate the true but unknown population mean. But our sample of 100 trees is just one of many possible samples (of the same size) that could have been randomly selected. Imagine if we take a series of different random samples from the same population and all the same size:

- Sample 1—we compute sample mean $\bar{x}$
- Sample 2—we compute sample mean $\bar{x}$
- Sample 3—we compute sample mean $\bar{x}$
- Etc.

Each time we sample, we may get a different result as we are using a different subset of data to compute the sample mean. This shows us that the sample mean is a random variable!

The sample mean ($\bar{x}$) is a random variable with its own probability distribution called the sampling distribution of the sample mean. The distribution of the sample mean will have a mean equal to μ and a standard deviation equal to $\frac{s}{\sqrt{n}}$

Note

The standard error $\frac{s}{\sqrt{n}}$ is the standard deviation of all possible sample means.

In reality, we would only take one sample, but we need to understand and quantify the sample to sample variability that occurs in the sampling process.

The standard error is the standard deviation of the sample means and can be expressed in different ways.

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \tag{1.1.7}$$

Note

\(s^2\) is the sample variance and *s* is the sample standard deviation

Example 1.1.8:

Describe the distribution of the sample mean.

A population of fish has weights that are normally distributed with μ = 8 lb. and s = 2.6 lb. If you take a sample of size n=6, the sample mean will have a normal distribution with a mean of 8 and a standard deviation (standard error) of $\frac{2.6}{\sqrt{6}}$ = 1.061 lb.

If you increase the sample size to 10, the sample mean will be normally distributed with a mean of 8 lb. and a standard deviation (standard error) of $\frac{2.6}{\sqrt{10}}$ = 0.822 lb.

Notice how the standard error decreases as the sample size increases.

The **Central Limit Theorem** (CLT) states that the sampling distribution of the sample means will approach a normal distribution as the sample size increases. If we do not have a normal distribution, or know nothing about our distribution of our random variable, the CLT tells us that the distribution of the $\bar{X}$'s will become normal as *n*increases. How large does *n* have to be? A general rule of thumb tells us that $n \geq 30$.

Note

The Central Limit Theorem tells us that regardless of the shape of our population, the sampling distribution of the sample mean will be normal as the sample size increases.

Coefficient of Variation

To compare standard deviations between different populations or samples is difficult because the standard deviation depends on units of measure. The coefficient of variation expresses the standard deviation as a percentage of the sample or population mean. It is a unitless measure.

Definition: CV of Population

$$CV = \frac{\sigma}{\mu} \times 100 \tag{1.1.8}$$

Definition: cv of sample

$$CV = \frac{s}{\bar{x}} \times 100 \tag{1.1.9}$$

Example 1.1.9:

Fisheries biologists were studying the length and weight of Pacific salmon. They took a random sample and computed the mean and standard deviation for length and weight (given below). While the standard deviations are similar, the differences in units between lengths and weights make it difficult to compare the variability. Computing the coefficient of variation for each variable allows the biologists to determine which variable has the greater standard deviation.

	Sample mean	Sample standard deviation
Length	63 cm	19.97 cm
Weight	37.6 kg	19.39 kg
	685.png	678.png

There is greater variability in Pacific salmon weight compared to length.

Variability

Variability is described in many different ways. Standard deviation measures point to point variability **within a sample**, i.e., variation among individual sampling units. Coefficient of variation also measures point to point variability but on a relative basis (relative to the mean), and is not influenced by measurement units. Standard error measures the **sample to sample variability**, i.e. variation among repeated samples in the sampling process. Typically, we only have one sample and standard error allows us to quantify the uncertainty in our sampling process.

Basic Statistics Example using Excel and Minitab Software

Consider the following tally from 11 sample plots on Heiburg Forest, where Xi is the number of downed logs per acre. Compute basic statistics for the sample plots.

ID					Order
1	25	625	-7.27	52.8529	4
2	35	1225	2.73	7.4529	6
4	55	3025	22.73	516.6529	10
5	15	225	-17.25	298.2529	2
6	40	1600	7.73	59.7529	8
7	25	625	-7.27	52.8529	5
8	55	3025	22.73	516.6529	11

ID					Order
9	35	1225	2.73	7.4529	7
10	45	2025	12.73	162.0529	9
11	5	25	-27.27	743.6529	1
Sum	20	400	-12.27	150.1819	3
	$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} X_i^2$	$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)$	$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$	

Table 1. Sample data on number of downed logs per acre from Heiburg Forest.

(1) Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{355}{11} 32.27 \tag{1.1.10}$$

(2) Median $= 35$

(3) Variance:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1} = \frac{2568.1519}{11-1} = 256.82$$
$$= \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1} = \frac{14025 - \frac{(355)^2}{11}}{11-1} = 256.82$$

(4) Standard deviation: $S = \sqrt{S^2} = \sqrt{256.82} = 16.0256$

(5) Range: $55 - 5 = 50$

(6) Coefficient of variation:

$$CV = \frac{S}{\bar{X}} \cdot 100 = \frac{16.0256}{32.27} \cdot 100 = 49.66\% \tag{1.1.11}$$

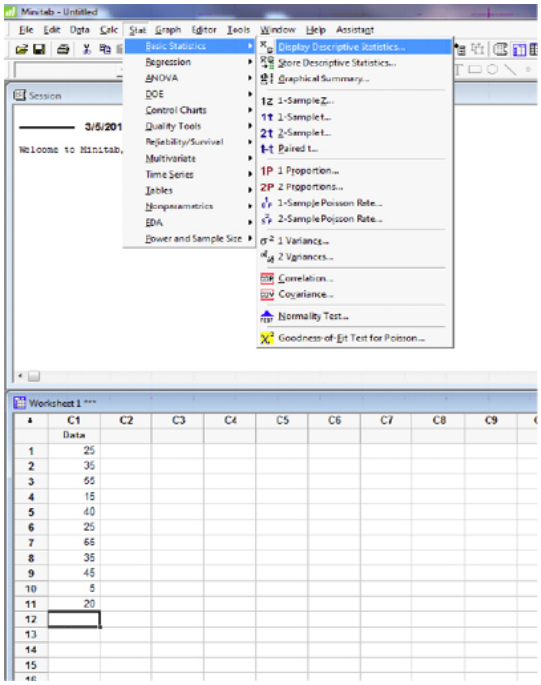
(7) Standard error of the mean:

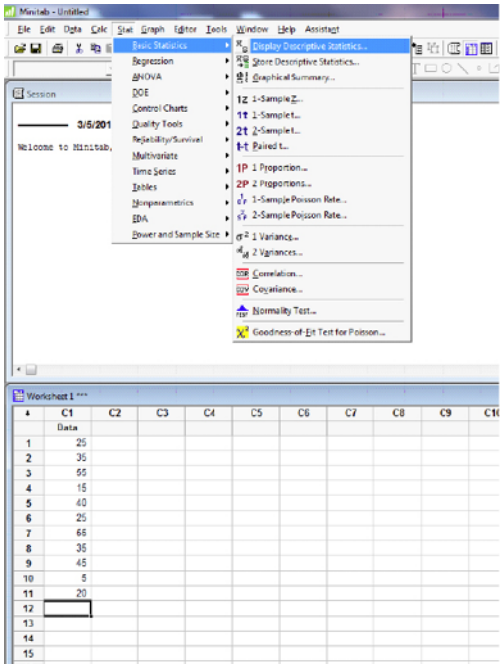
$$S_{\bar{X}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{256.82}{11}} = 4.8319$$
$$= \frac{S}{\sqrt{n}} = \frac{16.0256}{\sqrt{11}} = 4.8319$$

Software Solutions

Minitab

Open Minitab and enter data in the spreadsheet. Select STAT>Descriptive stats and check all statistics required.





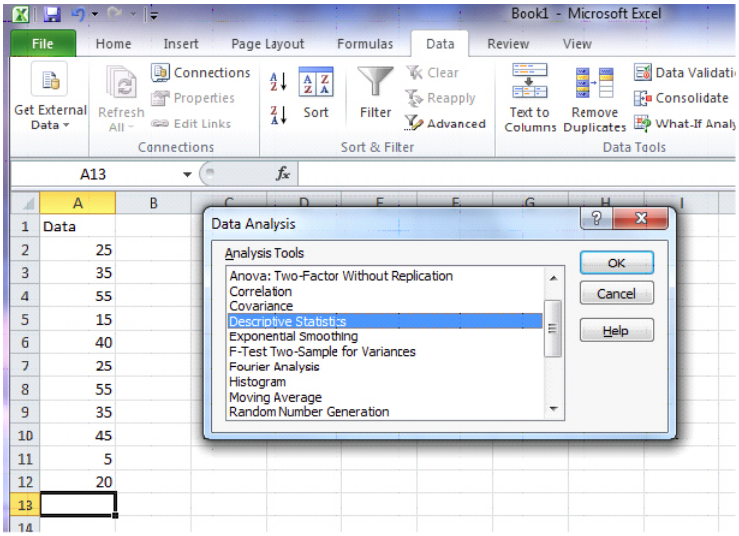
Descriptive Statistics: Data

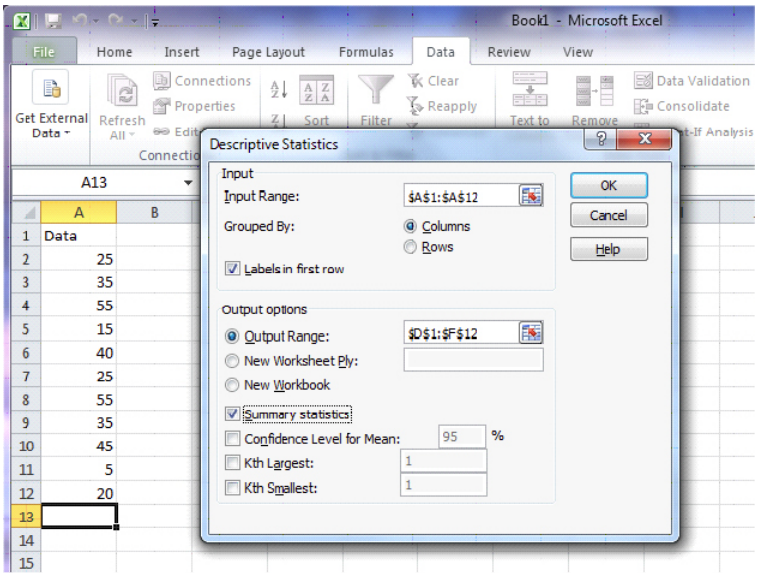
Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q1
Data	11	0	32.27	4.83	16.03	256.82	49.66	5.00	20.00

Variable	Median	Q3	Maximum	IQR
Data	35.00	45.00	55.00	25.00

Excel

Open up Excel and enter the data in the first column of the spreadsheet. Select DATA>Data Analysis>Descriptive Statistics. For the Input Range, select data in column A. Check “Labels in First Row” and “Summary Statistics”. Also check “Output Range” and select location for output.





Data	
Mean	32.27273
Standard Error	4.831884
Median	35
Mode	25
Standard Deviation	16.02555
Sample Variance	256.8182
Kurtosis	-0.73643
Skewness	-0.05982

Data	
Range	50
Minimum	5
Maximum	55
Sum	355
Count	11

Graphic Representation

Data organization and summarization can be done graphically, as well as numerically. Tables and graphs allow for a quick overview of the information collected and support the presentation of the data used in the project. While there are a multitude of available graphics, this chapter will focus on a specific few commonly used tools.

Pie Charts

Pie charts are a good visual tool allowing the reader to quickly see the relationship between categories. It is important to clearly label each category, and adding the frequency or relative frequency is often helpful. However, too many categories can be confusing. Be careful of putting too much information in a pie chart. The first pie chart gives a clear idea of the representation of fish types relative to the whole sample. The second pie chart is more difficult to interpret, with too many categories. It is important to select the best graphic when presenting the information to the reader.

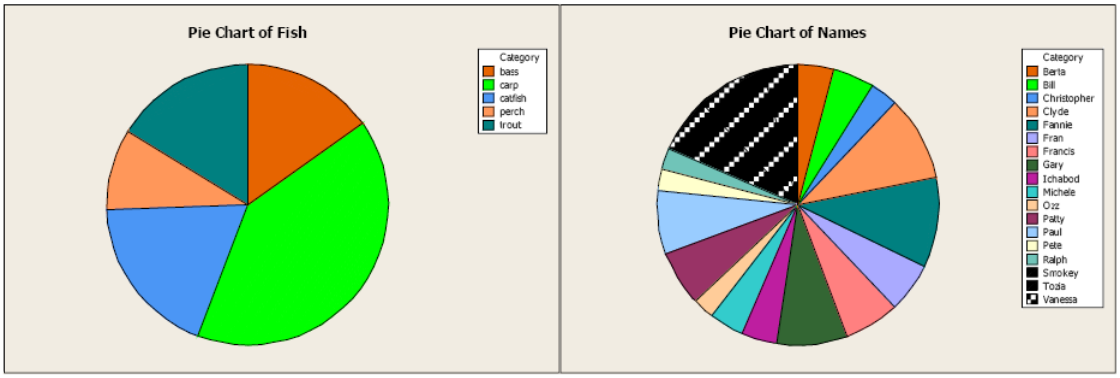


Figure 1.1.4:
Comparison of pie charts.
(Copyright; author via source)

Bar Charts and Histograms

Bar charts graphically describe the distribution of a qualitative variable (fish type) while histograms describe the distribution of a quantitative variable discrete or continuous variables (bear weight).

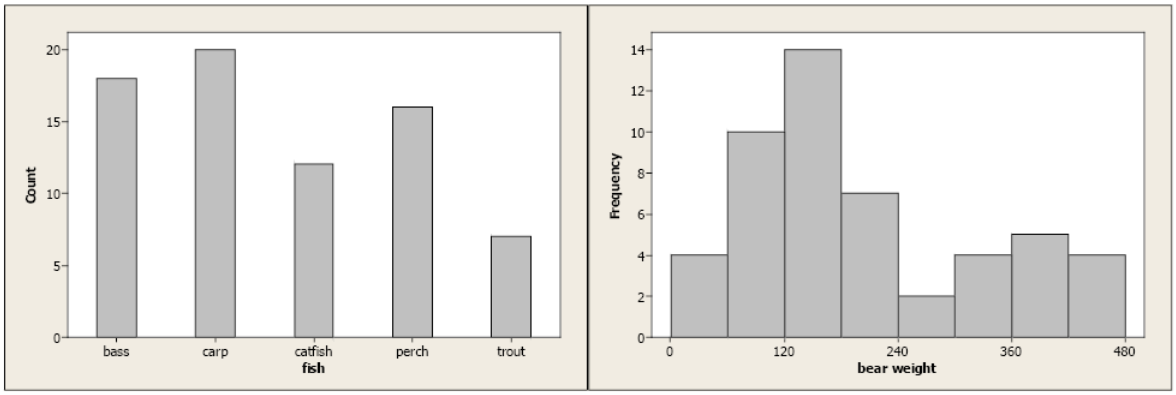


Figure 1.1.5:
Comparison of a bar chart for qualitative data and a histogram for quantitative data.
(Copyright; author via source)

In both cases, the bars' equal width and the y-axis are clearly defined. With qualitative data, each category is represented by a specific bar. With continuous data, lower and upper class limits must be defined with equal class widths. There should be no gaps between classes and each observation should fall into one, and only one, class.

Boxplots

Boxplots use the 5-number summary (minimum and maximum values with the three quartiles) to illustrate the center, spread, and distribution of your data. When paired with histograms, they give an excellent description, both numerically and graphically, of the data.

With symmetric data, the distribution is bell-shaped and somewhat symmetric. In the boxplot, we see that Q1 and Q3 are approximately equidistant from the median, as are the minimum and maximum values. Also, both whiskers (lines extending from the boxes) are approximately equal in length.

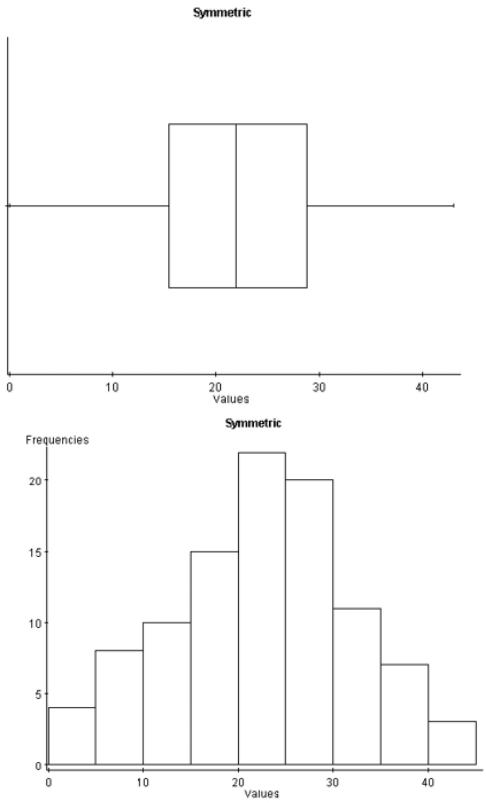


Figure 1.1.6: A histogram and boxplot of a normal distribution. (Copyright; author via source)

With skewed left distributions, we see that the histogram looks “pulled” to the left. In the boxplot, Q1 is farther away from the median as are the minimum values, and the left whisker is longer than the right whisker.

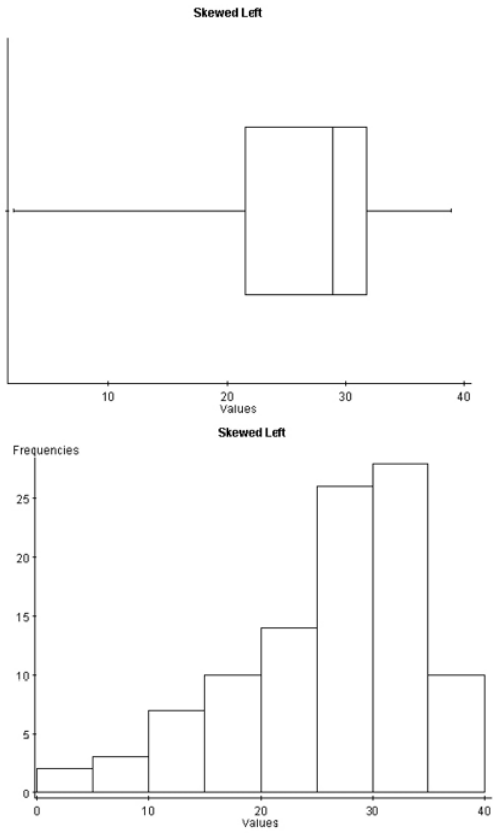


Figure 1.1.7:*A histogram and boxplot of a skewed left distribution.* (Copyright; author via source)

With skewed right distributions, we see that the histogram looks “pulled” to the right. In the boxplot, Q3 is farther away from the median, as is the maximum value, and the right whisker is longer than the left whisker.

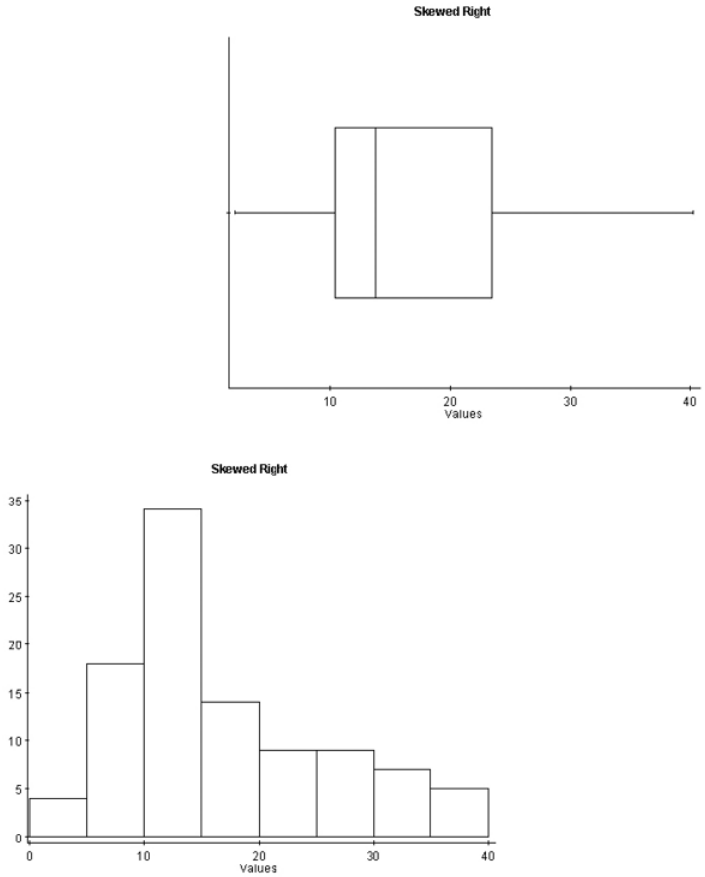


Figure 1.1.8: A *histogram and boxplot of a skewed right distribution.*(Copyright; author via source)
