Margin of Error

The data collected by sampling will vary with the sample collected. Repeated random samples vary in a regular way since random numbers are used to choose the sample.

It is desirable to take many samples from a population to estimate how trustworthy our results are.

The margin of error says how close to the truth about the population the sample results would be in 95% of all samples drawn by the method used to draw one sample.

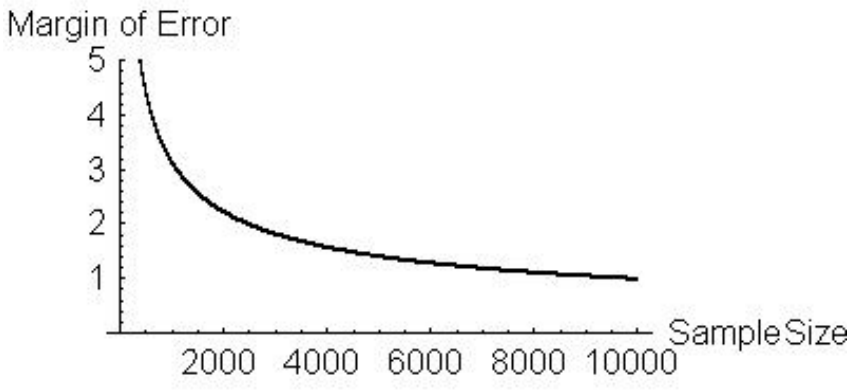
In other words, the margin of error is a measure of how precisely the sample results reflect the beliefs of the entire population.

We say we have 95% confidence that the truth about the population lies within the margin of error.

Finding the margin of error exactly is difficult work. There is a quick method for estimating the margin of error:

$$\text{margin of error} \sim \frac{100}{\sqrt{\text{sample size}}}$$

The following graph is online in an Excel file.



Since there is a square root in the quick estimate, to cut the margin of error in half we have to increase the sample size by a factor of 4! As the sample size increases, the reduction in the margin of error gets smaller.

The margin of error only estimates the error of chance in using simple random sampling techniques. The margin of error calculation does not account for bias in a sampling technique (for example, a voluntary response sample or a convenience sample).

Definitions

Individuals are the objects described by a data set, and may represent people, animals, things, quantities, etc.

Variables are characteristics of the individuals. The variable can be different for different individuals.

For example, the individuals might be students in a class, and the variable might be the final exam score. Other variables would be a quiz score, an assignment score, or a test score.

A set of n individuals would have the n observation values for a given variable: $x_1, x_2, x_3, \ldots, x_n$.

A distribution of a variable tells us what values the variable takes and how often it takes these values.

A distribution can be displayed by a histogram (good for large or small numbers of individuals) or a stemplot (good for small numbers of individuals) (and other ways we shall see shortly).

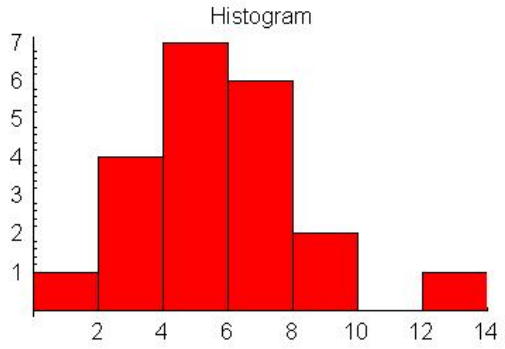
Constructing Histograms

A histogram is a graphical representation of data: the height of the bars in a histogram represent the number of times data was collected in the range at the base of the bar over a sample, or population, or even a number of different samples!

Consider the following set of data: 1, 2, 3, 5, 5, 6, 7, 8, 9, 12, 5, 6, 3, 4, 5, 6, 7, 4, 3, 5, 6

The information we need to construct a histogram can be represented in a table. The width of the base of the bar is a choice you make.

Bar width	Elements from list in width	Bar height (Number of elements in width)
0-2	$0 \leq 1 < 2$	1
2-4	$2 \leq 2, 3, 3, 3 < 4$	4
4-6	$4 \leq 5, 5, 5, 4, 5, 4, 5 < 6$	7
6-8	$6 \leq 6, 7, 6, 6, 7, 6 < 8$	6
8-10	$8 \leq 8, 9 < 10$	2
10-12	$10 \leq$ none < 12	0
12-14	$12 \leq 12 < 14$	1



If the data set given represents choices from a list, then it may be more appropriate to create a histogram with the number of times each number is listed. In that case, we would label the base of the histogram bars with the number it represents.

Histograms work well for large data sets, but you usually have to take some care to choose the bar width (sometimes called bin width) wisely to adequately convey the information in your data set to your audience.

There is an example of this online.

Constructing Stemplots

1. Separate each observation into a stem consisting of all but the final digit, and a leaf which is the final digit (leaves only have one digit!).
2. Write the stems in a vertical column with smallest at top. Draw a vertical line to the right of the column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Student #	Student Name	Test Score	Stem	Leaf
1	Mary	75	7	5
2	Mark	64	6	4
3	Ronnie	86	8	6
4	Tony	72	7	2
5	Al	77	7	7
6	Gayle	90	9	0
7	Sam	29	2	9
8	Fred	86	8	6
9	Beth	74	7	4
10	Amy	88	8	8
11	Rudy	91	9	1

Here is the stemplot of the data:

0	
1	
2	9
3	
4	
5	
6	4
7	2457
8	668
9	01

Note that a stemplot of the data would <u>not be</u>:

2	9
6	4
7	2457
8	668
9	01

You would lose the immediate recognition of 29 as a data element that is separate from the others.

Stemplots work well with small to mid-size data sets, and they contain all the data in the data set (a histogram does not give the audience access to the raw data).

Properties of Distribution

The distribution has certain properties that we can use to describe the look of the distribution. Usually, these properties can be easily described in very general terms from a histogram or stemplot.

The properties are typically:

- groupings
  Are the values of the variables closely grouped together?
  Is there one or more values of the variable that is far away from the other values of the variables?

- overall shape
  How many peaks are there?
  Is the distribution periodic (does it repeat itself)?
  Is the distribution symmetric (is one side like the other)?

- center
  Is there essentially one peak in the distribution?
  Where does this peak occur?
  We can describe the center of the distribution in terms of the average value, or the middle value.

- spread
  How wide is the distribution?

When describing a distribution:

- Look for outliers and try to explain why they are outliers.
- Concentrate on main feature, ignoring the outliers if you have been able to explain their existence to your satisfaction.
- Look for rough features, not precise variations.

These general descriptions can be converted to more formal descriptions using the following terms:

- An outlier is a value of a variable that falls outside the overall pattern.
- A distribution is symmetric if the left and right sides of the distribution are approximately mirror images of each other.
- A distribution is skewed to the right if the right side (or larger values) of the distribution extends much farther out than the left side (or smaller values).
- A distribution is skewed to the left if the left side (or smaller values) of the distribution extends much farther out than the right side (or larger values).
- The center of a distribution can be measured by calculating the mean (sometimes called the average) or the median (sometimes called the middle) of the distribution.
  Mean $= \bar{x} = \dfrac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$.
  Median: The number such that half the observations are smaller and half are larger.
  1. Arrange the observations from smallest to largest.
  2. If n is odd, the median is the observation in the middle of the list.
  3. If n is even, the median is the average of the two center observations in the list.
- Spread is described by quartiles and standard deviation.

Quartiles

We use quartiles to measure the spread of a distribution. They are similar to the median which is used to measure center.

The first quartile Q_1 is the number such that 25% of the observations are below it and 75% of the observations are above it.

The Median is the second quartile, the number such 50% of the observations are below it and 50% of the observations are above it.

The third quartile Q_3 is the number such that 75% of the observations are below it and 25% of the observations are above it.

The quartiles break the distribution up into quarters (hence the name).

We can combine these numbers with the minimum and maximum of the distribution to create the five-number summary of the distribution, which describes the distribution in some detail (center, spread).

$$\text{Minimum} \quad Q_1 \quad \text{Median} \quad Q_3 \quad \text{Maximum}$$

Example Here is a data set we can explore: $198, 167, 202, 190, 247, 238, 354, 234$. There are $n = 8$ observations.

The first thing we want to do is order the list from highest to lowest:

354 Maximum

247

Third Quartile $= Q_3 = (247 + 238)/2 = 242.5$

238

234

Median $= (234 + 202)/2 = 218$

202

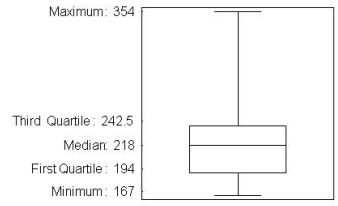
198

First Quartile $Q_1 = (198 + 190)/2 = 194$

190

167 Minimum

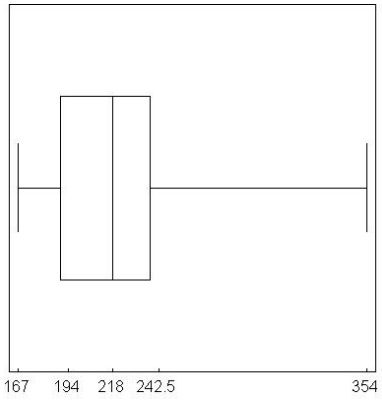
If we take the information we have above we can construct a box plot which visually represents the five number summary:



The box represents the spread of the middle half of the distribution; notice the median is NOT in the middle of the box.

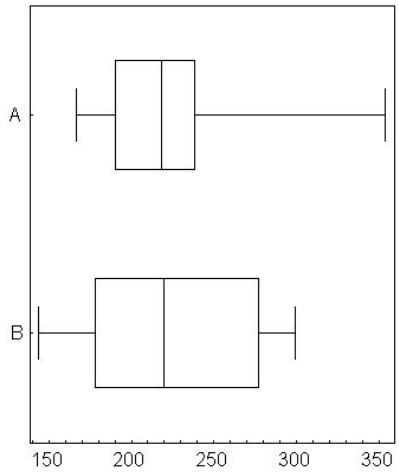
In this representation, the vertical distances are important, and the horizontal distances meaningless.

A boxplot could also be produced so the horizontal distances are the important ones:



Boxplots are useful for comparing more than one distribution at a time.

Example The boxplots for two distributions are given below.



We can instantly see that distribution B has a much larger spread than distribution A, although the medians of the two distributions are very close together, and the difference between the maximum and minimum values in distribution A is larger than for distribution B.

Variance and Standard Deviation

There are other ways of describing the center and spread of a data set, which are more often reported. They involve the mean and standard deviation.

The standard deviation is a measure of how far the observations are from their mean. The standard deviation is the square of the variance.

$$\text{Mean } = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Variance } = s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

$$\text{Standard Deviation } = s = \sqrt{s^2}$$

You can use a calculator to determine the standard deviation if you need it, so don't worry about memorizing this formula.

Note that the standard deviation has the same units as the mean.

Summary

- graphical description of distributions include:
  - histograms
  - stemplots
  - boxplots
- two methods of numerically describing center and spread of a distribution:
  - five number summary (min, first quartile, median, third quartile, max), leads to boxplots
  - mean and standard deviation
- the mean and standard deviation are greatly affected by outliers.
- the mean and standard deviation do not display the skewness of the distribution.
- the mean and standard deviation are best used for symmetric distributions without outliers.
- skewed distributions are best described by the five number summary, since the boxplot easily displays information about the skewness of the distribution.

Example From our previous example about grades:

Student #	Student Name	Test Score
1	Mary	75
2	Mark	64
3	Ronnie	86
4	Tony	72
5	Al	77
6	Gayle	90
7	Sam	29
8	Fred	86
9	Beth	74
10	Amy	88
11	Rudy	91

Since this is a small data set, a stemplot is still a great place to start.

Our previous stemplot:

0	
1	
2	9
3	
4	
5	
6	4
7	2457
8	668
9	01

There is an outlier at 29. Maybe this student did not prepare well for the test, or did not have the background necessary to do well on the test. Since this is very far from the other data, we will ignore it when looking for other features of the distribution.

The distribution appears to be symmetric. Therefore, it can be well described by mean and standard deviation.

Mean: 80.3 Standard deviation: 9.10 (where we used the formulas and excluded the outlier 29). Note that the

Note: If we didn't exclude the outlier, we would find: Mean: 75.6 Standard Deviation: 17.7 which is quite a bit different!

We now have a graphical picture of the data, a measure of center, a measure of spread, and have said something about symmetry and outliers. We are done analyzing the distribution.