**What Is the Law of Large Numbers?**

The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as the sample becomes larger.

In a financial context, the law of large numbers indicates that a large entity which is growing rapidly cannot maintain that growth pace forever. The biggest of the blue chips, with market values in the hundreds of billions, are frequently cited as examples of this phenomenon.

KEY TAKEAWAYS

* The law of large numbers states that an observed sample average from a large sample will be close to the true population average and that it will get closer the larger the sample.
* The law of large numbers does not guarantee that a given sample, especially a small sample, will reflect the true population characteristics or that a sample that does not reflect the true population will be balanced by a subsequent sample.
* The law of large numbers indicates a bigger sample will represent a population mean, while the central tendency theorem states a bigger sample will represent a population's distribution.
* In business, the term "law of large numbers" is sometimes used in a different sense to express the relationship between scale and growth rates.
* As a company becomes bigger, it will experience difficulties maintaining percentage targets because the underlying dollars may become too large and unfeasible.

**Understanding the Law of Large Numbers:**

The law of large numbers can refer to two different topics. First, in statistical analysis, the law of large numbers can be applied to a variety of subjects. It may not be feasible to poll every individual within a given population to collect the required amount of data, but every additional data point gathered has the potential to increase the likelihood that the outcome is a true measure of the mean.

The law of large numbers does not mean that a given sample or group of successive samples will always reflect the true population characteristics, especially for small samples. This also means that if a given sample or series of samples deviates from the true population average, the law of large numbers does not guarantee that successive samples will move the observed average toward the population mean (as suggested by the Gambler's Fallacy).

Second, the term "law of large numbers" is sometimes used in business in relation to growth rates, stated as a percentage. It suggests that, as a business expands, the percentage rate of growth becomes increasingly difficult to maintain. This is because the underlying dollar amount is actually increasing even if the growth rate as a percentage is to remain constant.

**Law of Large Numbers and Statistical Analysis**

If a person wanted to determine the average value of a data set of 100 possible values, he is more likely to reach an accurate average by choosing 20 data points instead of relying on just two. This is because there is greater probability of the two data points being outliers or non-representative of the average, while there is lower probability in all 20 data points being non-representative.

For example, if the data set included all integers from one to 100, and sample-taker only drew two values, such as 95 and 40, he may determine the average to be approximately 67.5. If he continued to take random samplings up to 20 variables, the average should shift towards the true average as he considers more data points.

Law of Large Numbers and Central Limit Theorem

In statistical analysis, the law of large numbers is related to the central limit theorem. The central limit theorem states that as the sample size increases, the sample mean will be evenly distributed. This is often depicted as a bell-shaped curve where the peak of the curve depicts the mean and even distributions of sample data fall to the left and right of the curve.

In a related manner, the law of large numbers also states that data is refined as the sample grows. However, the law of large numbers more closely relates to the center of the bell curve. The law of large numbers indicates that as a sample size increases, the mean of the sample will more closely resemble the mean of the population. Therefore, the law of large numbers relates to the peak (the mean) of a curve, while the central limit theorem relates to the distribution of a curve.

Law of Large Numbers and Business Growth

In business and finance, this term law of large numbers is sometimes used colloquially to refer to the observation that exponential growth rates often do not scale. This is not actually related to the law of large numbers, but may be a result of the law of diminishing marginal returns or diseconomies of scale.

The same principles can be applied to other metrics, such as market capitalization or net profit. As a result, investing decisions can be guided based on the associated difficulties that companies with very high market capitalization can experience as they relate to stock appreciation. This concept is somewhat central to growth versus value stocks, as a company may find it to maintain its business strategy of rapid growth once it achieves market success.

Law of Large Numbers in Business Example

In fiscal year 2020, Tesla reported automotive sales (not gross sales) of $24.604 billion. The next year, the company reported $44.125 billion, an increase of roughly 79%. As electric vehicles are an emerging market and Tesla is beginning to finally experience economies of scale, the company is started to experience success very quickly.

The law of large numbers indicates that as Tesla continues to grow, it will become harder for the company to maintain this level of productivity. For example, assuming a steady growth rate of the next several years, it becomes quickly apparent that Tesla simply cannot maintain its current growth trajectory due to the underlying dollar values becoming unreasonable.

Law of Large Numbers and Insurance

The law of large numbers is also prominent in the insurance industry to calculate and refine projected risk. Imagine a situation where an insurance company is assessing how much to charge different customers for car insurance. Should the company have a small data set, it will not be able to adequately determine appropriate risk profiles.

As the insurance agency collects more data, it experiences the law of large numbers, they may soon find that young, male drivers are most likely to cause an accident. This larger sample becomes more representative of driving incidents, and the insurance company can arrive at more accurate conclusions about the appropriate insurance premiums to charge.

In addition, the law of large numbers allows insurance companies to deeply refine the criteria in which to assess premiums by analyzing what traits cause higher risk. For example,

Why Is the Law of Large Numbers Important?

In statistical analysis, the law of large numbers is important because it gives validity to your sample size. When working with a small amount of data, the assumptions you make may not appropriately translate to the actual population. Therefore, it is important to make sure enough data points are being captured to adequately represent the entire data set.

In business, the law of large numbers is important when setting targets or goals. A company may double its revenue in a single year. Should the company obtain only 50% growth in revenue the next year, it will have earned the same amount of money each of the last two years. Therefore, it is important to be mindful that percentages can be misleading as large dollar values escalate.

How Can Companies Overcome the Challenge of the Law of Large Numbers?

Companies often strive to overcome the challenge of the law of large numbers by acquiring smaller growth companies that can infuse scalable growth. They also attempt to become more efficient and utilize their size for manufacturing, ordering, or distribution benefits. Last, companies can be more attentive to dollar goals as opposed to percent goals.

What Is the Law of Small Numbers?

The law of small numbers is the theory that people underestimate the variability in small sample sizes. This means that when people study a sample size that is too small, they usually overestimate the population's value based on the incorrect sample size.

What Is the Law of Large Numbers in Psychology?

Similar to other examples above, the law of large numbers in psychology translates to how a larger number of trials often leads to a more accurate expected value. As more trials are performed, the closer the projection is to being a correct medical assessment.

**Why does variability matter?**

While the central tendency, or average, tells you where most of your points lie, variability summarizes how far apart they are. This is important because the amount of variability determines how well you can generalize results from the sample to your population.

Low variability is ideal because it means that you can better predict information about the population based on sample data. High variability means that the values are less consistent, so it’s harder to make predictions.

Data sets can have the same central tendency but different levels of variability or vice versa. If you know only the central tendency or the variability, you can’t say anything about the other aspect. Both of them together give you a complete picture of your data.

**Range**

The range tells you the spread of your data from the lowest to the highest value in the distribution. It’s the easiest measure of variability to calculate.

Interquartile range

The interquartile range gives you the spread of the middle of your distribution. The interquartile range is the third quartile (Q3) minus the first quartile (Q1). This gives us the range of the middle half of a data set.

Five-number summary

Every distribution can be organized using a five-number summary:

Lowest value

Q1: 25th percentile

Q2: the median

Q3: 75th percentile

Highest value (Q4)

These five-number summaries can be easily visualized using box and whisker plots.

For any distribution that’s ordered from low to high, the interquartile range contains half of the values. While the first quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.

Standard deviation

The standard deviation is the average amount of variability in your dataset.

It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation by hand:

List each score and find their mean.

Subtract the mean from each score to get the deviation from the mean.

Square each of these deviations.

Add up all of the squared deviations.

Divide the sum of the squared deviations by n – 1 (for a sample) or N (for a population).

Find the square root of the number you found.

Standard deviation formula for populations

If you have data from the entire population, use the population standard deviation formula:

FormulaExplanation\sigma =\sqrt{\dfrac{\sum{(X - \mu)^2}}{N}}

* \sigma = population standard deviation
* \sum = sum of…
* X = each value
* \mu = population mean
* N = number of values in the population

If you have data from a sample, use the sample standard deviation formula:

If you have data from a sample, use the sample standard deviation formula:

| **Formula** | **Explanation** |
| --- | --- |
| s =\sqrt{\dfrac{\sum{(X - \bar{x})^2}}{n - 1}} | * s = sample standard deviation * \sum = sum of… * X = each value * \bar{x} = sample mean * n = number of values in the sample |

Samples are used to make statistical inferences about the population that they came from.

When you have population data, you can get an exact value for population standard deviation. Since you collect data from every population member, the standard deviation reflects the precise amount of variability in your distribution, the population.

But when you use sample data, your sample standard deviation is always used as an estimate of the population standard deviation. Using n in this formula tends to give you a biased estimate that consistently underestimates variability.

Reducing the sample n to n – 1 makes the standard deviation artificially large, giving you a conservative estimate of variability.

While this is not an unbiased estimate, it is a less biased estimate of standard deviation: it is better to overestimate rather than underestimate variability in samples.

The difference between biased and conservative estimates of standard deviation gets much smaller when you have a large sample size.

**Variance**

The variance is the average of squared deviations from the mean. A deviation from the mean is how far a score lies from the mean.

Variance is the square of the standard deviation. This means that the units of variance are much larger than those of a typical value of a data set.

While it’s harder to interpret the variance number intuitively, it’s important to calculate variance for comparing different data sets in statistical tests like ANOVAs.

Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

### ariance formula for populations

| **Formula** | **Explanation** |
| --- | --- |
| \sigma^2 = \dfrac{\sum (X - \mu)^2}{N} | * \sigma^2 = population variance * \sum = sum of… * Χ = each value * \mu = population mean * Ν = number of values in the population |

### Variance formula for samples

| **Formula** | **Explanation** |
| --- | --- |
| s^2= \dfrac{\sum (X - \bar{x})^2}{n - 1} | * s^2 = sample variance * \sum = sum of… * Χ = each value * \bar{x} = sample mean * n = number of values in the sample |

**Biased versus unbiased estimates of variance**

An unbiased estimate in statistics is one that doesn’t consistently give you either high values or low values – it has no systematic bias.

Just like for standard deviation, there are different formulas for population and sample variance. But while there is no unbiased estimate for standard deviation, there is one for sample variance.

If the sample variance formula used the sample n, the sample variance would be biased towards lower numbers than expected. Reducing the sample n to n – 1 makes the variance artificially larger.

In this case, bias is not only lowered but totally removed. The sample variance formula gives completely unbiased estimates of variance.

So why isn’t the sample standard deviation also an unbiased estimate?

That’s because sample standard deviation comes from finding the square root of sample variance. Since a square root isn’t a linear operation, like addition or subtraction, the unbiasedness of the sample variance formula isn’t carried over the sample standard deviation formula.

What’s the best measure of variability?

The best measure of variability depends on your level of measurement and distribution.

Level of measurement

For data measured at an ordinal level, the range and interquartile range are the only appropriate measures of variability.

For more complex interval and ratio levels, the standard deviation and variance are also applicable.

Distribution

For normal distributions, all measures can be used. The standard deviation and variance are preferred because they take your whole data set into account, but this also means that they are easily influenced by outliers.

For skewed distributions or data sets with outliers, the interquartile range is the best measure. It’s least affected by extreme values because it focuses on the spread in the middle of the data set.

# **Describing Distributions**

Sample Size is the number of elements in a sample. It is referred to by the symbol n.

Be sure to use a lower case n for sample size. An uppercase N refers to Population Size, unless being used in the context of a normally distributed population.

Mode is the data element which occurs most frequently.

A useful mnemonic is to alliterate the words mode and most. Alliterations start with the same sound like: "seven slippery slimy snakes...".

A data set with only one mode is termed unimodal. Some data sets contain no repeated elements. In this case, there is no mode (or the mode is the empty set). It is also possible for two or more [nonadjacent] elements to be repeated with the same frequency. In these cases, there are two or more modes and the data set is said to be bimodal or multimodal. In the rare instance of a uniform or nearly uniform distribution, one where each element is repeated the same or nearly the same number of times, one could term it multimodal, but some authors invoke subjectivity by specifying multimodality only when separate, distinct, and fairly high peaks (ignoring fluctuations due to randomness) occur.

For binned data, such as occurs with a frequency table, the interval which contains the most items is the modal interval and the midpoint of this interval is considered the mode. The mode is rather unsophisticated, tends to provide little information, and does not readily lend itself to mathematical manipulation. It thus has limited value except when there are a large number of scores and it can help describe the distribution or when used for nominal variables.

The Median is the middle element when the data set is arranged in order of magnitude.

A useful mnemonic is to remember that the median is the grassy strip (in the rural area of the midwest where I come from) that divides opposing lanes in a highway. It is in the middle.

If there are an odd number of data elements, the median is a member of the data set. If there are an even number of data elements, the median is computed as the arithmetic mean of the middle two.

The median has other names, such as P50, which will be discussed below. The Hinkle textbook uses the symbol Mdn for median.

The Midrange is the arithmetic mean of the highest and lowest data elements.

Midrange is a type of average. Range is a measure of dispersion and will be discussed below. A common mistake is to confuse the two. Symbolically, midrange is computed as (xmax+xmin)/2

**The Best Average**

The ambiguity of the term average can lend to deception. Statisticians may often be cast as liars as a result. Note how advertisers may distort statistics to pursue their goals.

Some basic facts regarding averages are as follows.

Mean, median, and midrange always exist and are unique.

Mode may not be unique or may not even exist.

Mean and median are very common and familiar.

Mode is used less frequently; midrange is rarely used.

Only the mean is "reliable" in that it utilizes every data element.

The midrange, and also somewhat the mean, can be distorted by extreme data elements.

The mode is the only appropriate average for nominal data.

Round-off Rules

The mode, if it exists, and possibly the median are elements of the data set. As such, they should be specified no more accurately than the original data set elements.

The midrange and possibly the median are the arithmetic mean of two data set elements. One additional significant digit may be necessary to accurately convey this information.

The number of significant digits for the mean should conform to one of the following rules.

The significant digits should be no more than the number of significant digits in the sum of the data elements. Since the sample size (n) is an exact value, it has no affect on the number of significant digits obtained from the division. This is sometimes simplified as a rule of thumb by stating that the mean should be given to one more decimal place than the original data. However, this assumes the data set is small (n < 100) and that the data was recorded to a consistant precision.

The number of significant digits should be consistant with the precision obtained for the standard deviation.

It is not uncommon in science for results to be left in and interim calculations sometimes rounded to three significant digits, which is about all you could get out of a slide rule. Hence, this was commonly termed slide rule accuracy. In pre-calculator days, this also made hand calculations easier.

The important thing to remember is not to write down twelve decimal places without good reason, even though your calculator will often display such.

Presenting more than five significant digits is probably a joke and points will be deducted!

**Measures of Dispersion**

Another important characteristic of a data set is how it is distributed, or how far each element is from some measure of central tendancy (average). There are several ways to measure the variability of the data. Although the most common and most important is the standard deviation, which provides an average distance for each element from the mean, several others are also important, and are hence discussed here.

**Range**

Range is the difference between the highest and lowest data element.

Symbolically, range is computed as xmax-xmin. Although this is very similar to the formula for midrange, please do not make the common mistake of reversing the two. This is not a reliable measure of dispersion, since it only uses two values from the data set. Thus, extreme values can distort the range to be very large while most of the elements may actually be very close together. For example, the range for the data set 1, 1, 2, 4, 7 introduced earlier would be 7-1=6.

Recently it has come to my attention that a few books define statistical range the same as its more mathematical usage. I've seen this both in grade school and college textbooks. Thus instead of being a single number it is the interval over which the data occurs. Such books would state the range as [xmin,xmax] or xmin to xmax. Thus for the example above, the range would be from 1 to 7 or [1,7]. Be sure you do not say 1-7 since this could be interpretted as -6.

Hinkle defines range as (Highest score - Lowest score) + 1, where the +1 ensures that both extreme values are included. Although he notes the definition given above, he does note that this +1 definition is used throughout the book. The appropriateness of this modification increases as the level of measurement decreases.

**Standard Deviation**

The Standard deviation is another way to calculate dispersion. This is the most common and useful measure because it is the average distance of each score from the mean. The formula for sample standard deviation is as follows.

[s= the square root of (the sum of the

squares of the deviations from the mean divided by n-1)] sample standard deviation

Notice the difference between the sample and population standard deviations. The sample standard deviation uses n-1 in the denominator, hence is slightly larger than the population standard deviation which use N (which is often written as n).

[sigma= the square root of (the sum of the

squares of the deviations from the mean divided by N)] population standard deviation

It is much easier to remember and apply these formulae, if you understand what all the parts are for. We have already discussed the use of Roman vs. Greek letters for sample statistics vs. population parameters. This is why s is used for the sample standard deviation and [sigma] (sigma) is used for the population standard deviation. However, another sigma, the capital one ([Sigma]), appears inside the formula. It serves to indicate that we are adding things up. What is added up are the deviations from the mean: [x bar] - xi. But the average deviation from the mean is actually zero—by definition of the mean! Occasionally the mean deviation, using average distance or using the symbols for absolute value: |[x bar] - xi| is used. However, a better measure of variation comes from squaring each deviation, summing those squares, then taking the square root after dividing by one less than the number of data elements. This is very similar to a quadratic mean. The n-1 can be understood in terms of degrees of freedom—a topic we will have to cover for inferential statistics.

Another formula for standard deviation is also commonly encountered. It is as follows.

[s squared = the square root of

((n times the sum of the data elements square less the square

of the sum of the data elements) divided by n(n-1))]

Shortcut formula for standard deviation

This formula can be algebraically derived from the former and has two primary applications. First, calculators and computer programs often employ it because less intermediate results are necessary and it can be calculated in one pass through the data set. That is, you don't have to calculate the mean first and then find the deviations. Second, it is closely related to a formula which may be used to calculate the standard deviation for a frequency table. In general, the formulae are not used and we rely instead on calculators or computers.

**Variance**

Variance is the third method of measuring dispersion. Compare the two variance formulae with their corresponding standard deviation formulae, and we see that variance is just the square of the standard deviation. Statisticians tend to consider variance a primary measure and use it extensively (ANOVA, etc.), whereas scientists are very happy to use standard deviation exclusively. Personally, I have difficulty conceptualizing square points or square dollars.

[s squared = the sum of the

squares of the deviations from the mean divided by n-1] [sigma square = the sum of the

squares of the deviations from the mean divided by N]

Occasionally, the abbreviations SD for standard deviation and Var for variance will be seen.

**Range Rule of Thumb**

It can take some time to start to understand how these measures of variation may be useful. Consider the following scenerios. First, if a straight five points are added to everyone's score, the mean would increase five points, say from 70.8 to 75.8 but have no affect on the standard deviation. It remains, say, at 10.9. Second, if each test score was multiplied by .89 and then 21 points were added, not only does this move the mean from, say, 55.4 to 70.3, but it also reduced the standard deviation from, say, 15.0 to 13.5. This can be useful if the original test scores were very variable, and could easily have resulted in more D's and F's than your efforts justified. You might consider a third common way to adjust test scores, that of dropping the possible. Technically this doesn't change either the mean or the standard deviation, but it does effectively raise everyone's percentage. This doesn't help the lower scoring students nearly as much as it helps the top students.

A commonly given rule of thumb is that the range of a data set is approximately 4 standard deviations (4s). Thus the maximum data element will be about 2 standard deviations above the mean and the minimum data element about 2 standard deviations below the mean.

**More Round-off Information**

The standard deviation of a data set is often used in science as a measure of the precision to which a experiment has been done. It can also indicate the reproducibility of the result. Propagation of error dictates that intermediate values in your calculations should not be rounded. At least twice as many digits as will be used in the final answer should be retained.

It is rather meaningless to calculate the standard deviation for a data set of two elements.

Three is considered the smallest sample size where standard deviation is meaningful.

It is not uncommon for an experiment to involve millions of events and associated data. If you examine the standard deviation formula above, you will note that it depends inversely on the square root of n. We could thus expect to reduce the standard deviation of our answer by perhaps a thousand fold. It is the goal of many experiments to obtain very precise values, so great care is exercised to reduce systematic errors and also reduce the effect of random errors by increasing the repetitions.

Example: Consider a simple example of counting pennies where the outcomes 99, 100 and 101 are obtained. Find the mean and standard deviation.

Solution: We can easily calculate the mean as 100 and the standard deviation as 1.0.

Example: Consider further if this exercise were repeated 1000 times and 100 was obtained 991 times, 99 5 times and 101 4 times. Again, calculate the mean and standard deviation.

Solution: The mean is now 99.999 and the standard deviation is now 0.095. Here the additional precision is justified and the mean and standard deviation are given to the same 3 decimal place precision. It would be a mistake to report these results to only one more digit than the original data set, as in 100.0 and 0.1.

DO NOT USE a rounded s to obtain s2. Variance is the primary statistic, s is a derived quantity.

Standard deviation should be reported to at least one more decimal place than the data, or three significant digits.

Standard or z-Scores

We often find it useful to calculate how far, in standard deviations, a data element was from the mean. This is a very widely used procedure and this measure has the name z-score. It is also termed a standard score. Since many data sets have a somewhat normal distribution, it is a very helpful way to compare data elements from different populations—populations which may very well have differing means and standard deviations. However, we will be discussing the normal distribution tomorrow.

A typical example might be ACT and SAT scores. ACT scores range from 1 to 36 with a national mean of about 21.0 and standard deviation of about 4.7. SAT scores range from 200 to 800 (for each subtest) with a national mean of about 508 and standard deviation of about 111. Both ACTs and SATs appear to be approximately normally distributed. High school students often take both, perhaps several times and those from a particular school would represent a sample. This sample would have its own mean and standard deviation, but of course, these would be statistics, not parameters. (Our Math and Science Center students average about 1050 (total) when they take the SAT their eighth grade year and average over 1300 (total) when they take it their junior year. Our average ACT score (junior) is about 29.) The formulae used for z-score appear in two virtually identical forms, recognizing the fact that we may be dealing with sample statistics or population parameters. These formulae are as follows.

[z=(x-x bar)/s] z-score formulae [z=(x-mu)/sigma]

The following important attributes should be noted about z-scores.

Negative z-scores indicate a data element's position below the mean.

Positive z-scores indicate a data element's position above the mean.

z-scores should always be rounded to two decimal places.

IQs of 0 and 210 will be discussed in lesson 4 and z-scores of -6.67 and 7.33 should be obtained respectively, based on a population mean of 100 and a standard deviation of 15.

The population does not have to be normally distributed to calculate z-scores, but that is one of its primary applications.

In summary, z-scores provide a useful measurement for comparing

data elements from different data sets.

Ordinary or Unusual Scores

Now that we have defined z-score, we can define two more terms as follows.

Data elements more than 2 standard deviations away from the mean are termed unusual.

Data elements less than 2 standard deviations away from the mean are termed ordinary.

As you will recall, in a normally distributed population, 95% of the data will then be ordinary, so only 5% can be unusual. Chebyshev's theorem guarantees at least 75% of the data to be ordinary, so no more than 25% can be unusual.

Quartiles

Yet another method of measuring how a data set is distributed is to extend the concept of median and use smaller and smaller divisions. The first division we will examine is the quartile.

Note first how the median divides a population into two halves: a top half and a bottom half.

The top half consists of those data elements above the median, whereas the bottom half consists of those data elements below the median. If we subdivide each of these halves yet again, we have quartered the population and each of these division points are termed quartiles. Although one might occasionally speak of the bottom quartile, top quartile, etc., the term quartile technically refers to the three division points and not to the four divisions of the data.

Q1 is the term used for the median of the bottom half.

Q3 is the term used for the median of the top half.

Q2 is another term used for the median.

The precise definition specifies that at least 25% of the data will be less than or equal to Q1 and at least 75% of the data will be less than or equal to Q3. The terms upper (right) and lower (left) hinge are noted below and some software packages may not clearly differentiate between hinges and quartiles. All these measures of position assume the data is quantitative and can be put in numeric order.

Data are ranked when arranged in [numeric] order.

Since range is sensitive to outliers (defined below), sometimes the interquartile range is calculated. This range is the difference between the third and first quartiles: Q3-Q1. It is another measure of dispersion. Other common terms include: semi-interquartile range, (Q3-Q1)/2, another measure of dispersion, and midquartile or (Q1+Q3)/2, which is a measure of central tendancy (an average).

Hinges; Mild and Extreme Outliers

Another common term is hinge. There is a left or lower hinge and a right or upper hinge.

The upper hinge is the median of the upper half of all scores, including the median.

The lower hinge is the median of the lower half of all scores, including the median.

Outliers are extreme values in a data set. Sometimes the term outlier is applied to unusual values as defined above (Triola, 5th edition). More recently, outliers are defined in terms of the hinges or quartiles. Outliers are often differentiated as mild or extreme as defined below. The interquartile range or perhaps D = upper hinge - lower hinge is used. Generally, an outlier should be obvious and not borderline—right next to another element, but lying just outside some arbitrary line of demarcation.

Mild outlier are 1.5D to 3D beyond the corresponding hinge.

Hinkle terms these demarcation points 1.5 IQR beyond a hinge/quartile reasonable upper boundary (RUB) and reasonable lower boundary> (RLB). In a modified box plot (discussed below), the RUB and RLB replace the max and min, respectively, and outliers are noted with dots.

Extreme outlier are beyond the corresponding hinge by more than 3D.

Consider as an example the data set: {0, 2, 4, 5, 6, 3, 6, 1, 1, 50}. Obviously, 50 is a much larger number than any of the other elements. This outlier will cause the mean and variance to be much higher. Specifically, without 50, the mean is 3.1 and standard deviation 2.3, whereas with 50, the mean is 7.8 and standard deviation 15.0. Note that the quartiles are 1 and 6, whereas the hinges are 1.5 and 5.5 for the unmodified data set. For any of these definitions, 50 is way away from the other data and is an outlier. Outliers might be legitimate data values or errors. This 50 might really have been 5.0 and was miscoded (historically, punch card input was column sensitive) or poorly recorded in a lab book, with the decimal point extremely light or missing. 50 may also represent extreme extra credit on a 5 point quiz! It is not unusual to be tempted to omit such data values. It is not considered a good practice, but if such are omitted, be sure to clearly record that fact. You will have just crossed the line between objective and subjective science.

Deciles

Although not nearly as common as percentiles which follow below, deciles are yet another fractile which serve to partition data into approximately equal parts. Hence, just as there are three quartiles which divide a population into four parts, so too are there nine deciles dividing the population into ten parts. The deciles are termed D1 through D9.

D5 is another name for the median.

Stanine Scores

The term stanine is derived from standard nine and stanine scores range from 1 to 9 with 5 in the center. Except for 1 and 9, each stanine includes a band of scores one half a standard devaition wide. Thus stanine scores are standard scores with a mean of 5 and a standard deviation of 2. Test scores are commonly expressed using these single-digit scores which can help students and parents visualize where someone falls on the test scale.

Psycholgists and counselors frequently provide Norm-referenced interpretation of a scores for personality inventory and achievement tests. This typically means correlating a given score with a given percentile.

Percentiles

Percentiles are also like quartiles, but divide the data set into 100 equal parts. Each group represents 1% of the data set. There are 99 percentiles termed P1 through P99.

P50 is yet another term for median.

Other equivalents, such as P25=Q1, P75=Q3, P10=D1, etc., should also be obvious. Once again, the term percentile technically refers to the 99 division points, but is not uncommonly used to refer to the 100 divisions. For large data sets, one can calculate the locator L to help find a requested percentile. It is computed as follows.

[L=(k/100)n] Percentile Locator Formula

k is the percentile being sought and n, of course, is the number of elements in our data set. Usual conventions dictate that once L is obtained, it must be checked to see if it is a whole number. If it is a whole number, the value of Pk is the mean of the Lth data element and the next higher data element. If it is not a whole number, L must be rounded up to the next larger whole number. The value of Pk is then the Lth data element, counting from the lowest. There is an essential difference between rounding up and rounding off. If we round off [pi] we get 3. Whereas, if we round up [pi] we get 4. Hinkle gives a different formula which is applicable when the data is binned. Since percentiles are ordinal, a limited number of statistical operations are approriate for them.

The percentile rank of a score is a point on the percentile scale that gives the percent of scores at or below the specified score. When percentiles and scores are graphed in a cumulative frequency polygon or ogive, one can read a score on one axis and find percentile on the other or percentile on one and the corresponding percentile rank (a score) on the other.