Statistics Refresher

Courtesy of Dr. David Lane – Associate Professor
Rice University – Houston, Texas
On the web at http://onlinestatbook.com/rvls/

I. Statistics

The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity (such as the mean) calculated in a sample. Such statistics are used to estimate parameters.

The term "statistics" sometimes refers to calculated quantities regardless of whether or not they are from a sample. For example, one might ask about a baseball player's statistics and be referring to his or her batting average, runs batted in, number of home runs, etc. Or, "government statistics" can refer to any numerical indexes calculated by a governmental agency.

although the different meanings of "statistics" has the potential for confusion, a careful consideration of the context in which the word is used should make its intended meaning clear.

II. Descriptive Statistics

One important use of descriptive statistics is to summarize a collection of data in a clear and understandable way. For example, assume a psychologist gave a personality test measuring shyness to all 2500 students attending a small college. How might these measurements be summarized? There are two basic methods: numerical and graphical. Using the numerical approach one might compute statistics such as the mean and standard deviation. These statistics convey information about the average degree of shyness and the degree to which people differ in shyness. Using the graphical approach one might create a stem and leaf display and a box plot. These plots contain detailed information about the distribution of shyness scores.

Graphical methods are better suited than numerical methods for identifying patterns in the data. Numerical approaches are more precise and objective.

Since the numerical and graphical approaches compliment each other, it is wise to use both.

III. Inferential Statistics

Inferential statistics are used to draw inferences about a population from a sample. Consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night's sleep. Is the difference real or could it be due to chance? How much larger could the real difference be than the 12 points found in the sample? These are the types of questions answered by inferential statistics.

There are two main methods used in inferential statistics: estimation and hypothesis testing. In estimation, the sample is used to estimate a parameter and a confidence interval about the estimate is constructed.

In the most common use of hypothesis testing, a "straw man" null hypothesis is put forward and it is determined whether the data are strong enough to reject it. For the sleep deprivation study, the null hypothesis would be that sleep deprivation has no effect on performance.

Confidence Interval (1 of 2)

A confidence interval is a range of values computed in such a way that it contains the estimated parameter a high proportion of the time. The 95% confidence interval is constructed so that 95% of such intervals will contain the parameter. Similarly, 99% of 99% confidence intervals contain the parameter. If the parameter being estimated were μ, the 95% confidence interval might look like the following:

$12.5 \leq \mu \leq 30.2$

If other information about the value of the parameter is available, it should be taken into consideration when assessing the likelihood that the interval contains the parameter. As an extreme example, consider the case in which 1,000 studies estimating the value of μ in a certain population all resulted in estimates between 25 and 30. If one more study were conducted and if the 95% confidence interval on μ were computed (based on that one study) to be:

$35 \leq \mu \leq 45$

then the probability that μ is between 35 and 45 is very low, the confidence interval not withstanding.

It is natural to interpret a 95% confidence interval on the mean as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple. As just discussed, one problem is that the computation of a confidence interval does not take into account any other information you might have about the value of the population mean.

Confidence Interval (2 of 2)

What about situations in which there is no prior information? Even here the interpretation is complex. The problem is that there can be more than one procedure that produces intervals that contain the population parameter 95% of the time. Which procedure produces the "true" 95% confidence interval? although the various methods are equal from a purely mathematical point of view, the standard method of computing confidence intervals has two desirable properties: (1) each interval is symmetric about the point estimate and (2) each interval is contiguous. Recall from the introductory section in the chapter on probability that for some purposes, probability is best thought of as subjective. It is reasonable, although not required by the laws of probability, that one adopt a subjective probability of 0.95 that a 95% confidence interval as typically computed contains the parameter in question.

Confidence intervals can be constructed for any estimated parameter, not just μ. For example, one might estimate the proportion of people who could pass a training program or the difference between the mean for subjects taking a drug and those taking a placebo.

Inferential Statistics: Population

A population consists of an entire set of objects, observations, or scores that have something in common. For example, a population might be defined as all males between the ages of 15 and 18.

Some populations are only hypothetical. Consider an experimenter interested in the possible effectiveness of a new method of teaching reading. He or she might define a population as the reading achievement scores that would result if all six year olds in the US were taught with this new method. The population is hypothetical in the sense that there does not exist a group of students who have been taught using the new method; the population consists of the scores that would be obtained if they were taught with this method.

The distribution of a population can be described by several parameters such as the mean and standard deviation. Estimates of these parameters taken from a sample are called statistics.

Inferential Statistics: Sample

A sample is a subset of a population. Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available.

Inferential statistics generally require that sampling be random although some types of sampling (such as those used in voter polling) seek to make the sample as representative of the population as possible by choosing the sample to resemble the population on the most important characteristics.

See also: biased sample and random sample

IV. Variables (1 of 2)

A variable is any measured characteristic or attribute that differs for different subjects. For example, if the weight of 30 subjects were measured, then weight would be a variable.

Quantitative and Qualitative

Variables can be quantitative or qualitative. (Qualitative variables are sometimes called "categorical variables.") Quantitative variables are measured on an ordinal, interval, or ratio scale; qualitative variables are measured on a nominal scale. If five-year old subjects were asked to name their favorite color, then the variable would be qualitative. If the time it took them to respond were measured, then the variable would be quantitative.

Independent and Dependent

When an experiment is conducted, some variables are manipulated by the experimenter and others are measured from the subjects. The former variables are called "independent variables" or "factors" whereas the latter are called "dependent variables" or "dependent measures."

Variables (2 of 2)

For example, consider a hypothetical experiment on the effect of drinking alcohol on reaction time: Subjects drank either water, one beer, three beers, or six beers and then had their reaction times to the onset of a stimulus measured. The independent variable would be the number of beers drunk (0, 1, 3, or 6) and the dependent variable would be reaction time.

Continuous and Discrete

Some variables (such as reaction time) are measured on a continuous scale. There is an infinite number of possible values these variables can take on. Other variables can only take on a limited number of values. For example, if a dependent variable were a subject's rating on a five- point scale where only the values 1, 2, 3, 4, and 5 were allowed, then only five possible values could occur. Such variables are called "discrete" variables.

V. Parameters

A parameter is a numerical quantity measuring some aspect of a population of scores. For example, the mean is a measure of central tendency.

Greek letters are used to designate parameters. At the bottom of this page are shown several parameters of great importance in statistical analyses and the Greek symbol that represents each one. Parameters are rarely known and are usually estimated by statistics computed in samples. To the right of each Greek symbol is the symbol for the associated statistic used to estimate it from a sample.

Quantity	Parameter	Statistic
Mean	μ	M
Standard deviation	σ	s
Proportion	π	p
Correlation	ρ	r

Central Tendency

Measures of central tendency are measures of the location of the middle or the center of a distribution. The definition of "middle" or "center" is purposely left somewhat vague so that the term "central tendency" can refer to a wide variety of measures. The mean is the most commonly used measure of central tendency. The following measures of central tendency are discussed in this text:
Mean
Median
Mode
Trimean
Trimmed mean

For symmetric distributions, these measures are all the same. For skewed distributions, they can differ markedly.

Mean

Arithmetic Mean

The arithmetic mean is what is commonly called the average: When the word "mean" is used without a modifier, it can be assumed that it refers to the arithmetic mean. The mean is the sum of all the scores divided by the number of scores. The formula in summation notation is:

$\mu = \Sigma X/N$

where μ is the population mean and N is the number of scores.

If the scores are from a sample, then the symbol M refers to the mean and N refers to the sample size. The formula for M is the same as the formula for
μ.

$M = \Sigma X/N$

The mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in skewed distributions since it can be greatly influenced by scores in the tail. Therefore, other statistics such as the median may be more informative for distributions such as reaction time or family income that are frequently very skewed

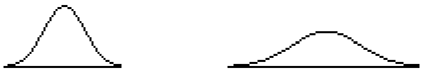
The sum of squared deviations of scores from their mean is lower than their squared deviations from any other number.

For normal distributions, the mean is the most efficient and therefore the least subject to sample fluctuations of all measures of central tendency.

The formal definition of the arithmetic mean is μ = E[X] where μ is the population mean of the variable X and E[X] is the expected value of X.

Normal Distribution

Normal distributions are a family of distributions that have the shape shown below.



Normal distributions are symmetric with scores more concentrated in the middle than in the tails. They are defined by two parameters: the mean (μ) and the standard deviation (σ). Many kinds of behavioral data are approximated well by the normal distribution. Many statistical tests assume a normal distribution. Most of these tests work well even if the distribution is only approximately normal and in many cases as long as it does not deviate greatly from normality.

The formula for the height (y) of a normal distribution for a given value of x is:

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

Standard Deviation and Variance (1 of 2)

The variance and the closely-related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667$$

.

The formula (in summation notation) for the variance in a population is

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

where μ is the mean and N is the number of scores.

When the variance is computed in a sample, the statistic

$$S^2 = \frac{\sum (X-M)^2}{N}$$

(where M is the mean of the sample) can be used. S^2 is a biased estimate of σ^2, however. By far the most common formula for computing variance in a sample is:

$$s^2 = \frac{\sum (X-M)^2}{N-1}$$

which gives an unbiased estimate of σ^2. Since samples are usually used to estimate parameters, s^2 is the most commonly used measure of variance. Calculating the variance is an important part of many statistical applications and analyses. It is the first step in calculating the standard deviation.

Standard Deviation

The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread.

An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standard deviations of the mean.

The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation.

(See next page for applications to risk analysis and stock portfolio volatility.)

Standard Deviation and Variance (2 of 2)

although less sensitive to extreme scores than the range, the standard deviation is more sensitive than the semi-interquartile range. Thus, the standard deviation should be supplemented by the semi-interquartile range when the possibility of extreme scores is present.

If variable Y is a linear transformation of X such that:

Y = bX + A,

then the variance of Y is:

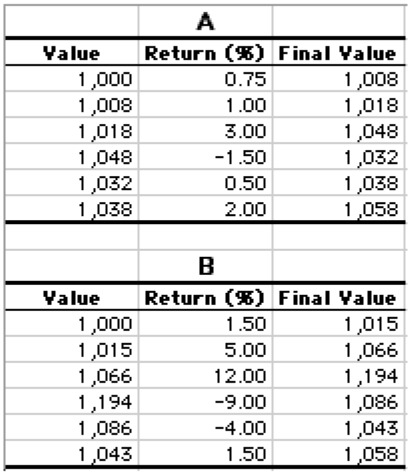
$$\sigma_y^2 = b^2\sigma_x^2$$

where σ_x^2 is the variance of X.

The standard deviation of Y is $b\sigma_x$ where σ_x is the standard deviation of X.

Standard Deviation as a Measure of Risk

The standard deviation is often used by investors to measure the risk of a stock or a stock portfolio. The basic idea is that the standard deviation is a measure of volatility: the more a stock's returns vary from the stock's average return, the more volatile the stock. Consider the following two stock portfolios and their respective returns (in per cent) over the last six months. Both portfolios end up increasing in value from $1,000 to $1,058. However, they clearly differ in volatility. Portfolio A's monthly returns range from -1.5% to 3% whereas Portfolio B's range from -9% to 12%. The standard deviation of the returns is a better measure of volatility than the range because it takes all the values into account. The standard deviation of the six returns for Portfolio A is 1.52; for Portfolio B it is 7.24.



	A	
Value	Return (%)	Final Value
1,000	0.75	1,008
1,008	1.00	1,018
1,018	3.00	1,048
1,048	-1.50	1,032
1,032	0.50	1,038
1,038	2.00	1,058

	B	
Value	Return (%)	Final Value
1,000	1.50	1,015
1,015	5.00	1,066
1,066	12.00	1,194
1,194	-9.00	1,086
1,086	-4.00	1,043
1,043	1.50	1,058

Sampling Distribution of a Proportion (1 of 4)

Assume that 0.80 of all third grade students can pass a test of physical fitness. A random sample of 20 students is chosen: 13 passed and 7 failed. The parameter π is used to designate the proportion of subjects in the population that pass (.80 in this case) and the statistic p is used to designate the proportion who pass in a sample (13/20 = .65 in this case). The sample size (N) in this example is 20. If repeated samples of size N where taken from the population and the proportion passing (p) were determined for each sample, a distribution of values of p would be formed. If the sampling went on forever, the distribution would be the sampling distribution of a proportion. The sampling distribution of a proportion is equal to the binomial distribution. The mean and standard deviation of the binomial distribution are:

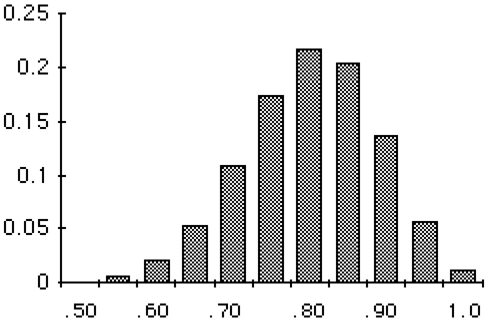
$\mu = \pi$

and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}$.

For the present example, N = 20, π = 0.80, the mean of the sampling distribution of p (μ) is .8 and the standard error of p (σ_p) is 0.089. The shape of the binomial distribution depends on both N and π. With large values of N and values of π in the neighborhood of .5, the sampling distribution is very close to a normal distribution.

Sampling Distribution of a Proportion (2 of 4)

The plot shown here is the sampling distribution for the present example. As you can see, the distribution is not far from the normal distribution, although it does have some negative skew.



Assume that for the population of people applying for a job at a bank in a major city, .40 are able to pass a basic literacy test required to get the job. Out of a group of 20 applicants, what is the probability that 50% or more of them will pass? This problem involves the sampling distribution of p with π = .40 and N = 20. The mean of the sampling distribution is π = .40. The standard deviation is:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}} = \sqrt{\frac{0.40(1-0.40)}{20}} = 0.11$$

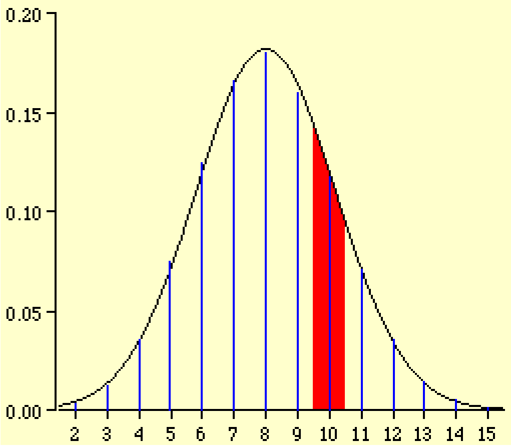
Sampling Distribution of a Proportion (3 of 4)

Using the normal approximation, a proportion of .50 is: (.50-.40)/.11 = 0.909 standard deviations above the mean. From a z table it can be calculated that 0.818 of the area is below a z of 0.909. Therefore the probability that 50% or more will pass the literacy test is only about 1 - 0.818 = 0.182.

Correction for Continuity

Since the normal distribution is a continuous distribution, the probability that a sample value will exactly equal any specific value is zero. However, this is not true when the normal distribution is used to approximate the sampling distribution of a proportion. A correction called the "correction for continuity" can be used to improve the approximation.

The basic idea is that to estimate the probability of, say, 10 successes out of 20 when π is 0.4, one should compute the area between 9.5 and 10.5 as shown below.



Sampling Distribution of a Proportion (4 of 4)

Therefore to compute the probability of 10 or more successes, compute the area above 9.5 successes. In terms of proportions, 9.5 successes is 9.5/20 = 0.475. Therefore, 9.5 = (0.475 - 0.40)/.11 = 0.682 standard deviations above the mean. The probability of being 0.682 or more standard deviations above the mean is 0.247 rather than the 0.182 that was obtained previously.The exact answer calculated using the binomial distribution is 0.245. For small sample sizes the correction can make a much bigger difference than it did here.

Binomial distribution (1 of 3)

When a coin is flipped, the outcome is either a head or a tail; when a magician guesses the card selected from a deck, the magician can either be correct or incorrect; when a baby is born, the baby is either born in the month of March or is not. In each of these examples, an event has two mutually exclusive possible outcomes. For convenience, one of the outcomes can be labeled "success" and the other outcome "failure." If an event occurs N times (for example, a coin is flipped N times), then the binomial distribution can be used to determine the probability of obtaining exactly r successes in the N outcomes. The binomial probability for obtaining r successes in N trials is:

$$P(r) = \frac{N!}{r!(N-r)!}\pi^{r}(1-\pi)^{N-r}$$

where P(r) is the probability of exactly r successes, N is the number of events, and π is the probability of success on any one trial. This formula for the binomial distribution assumes that the events:

1. are dichotomous (fall into only two categories)
2. are mutually exclusive
3. are independent and
4. are randomly selected

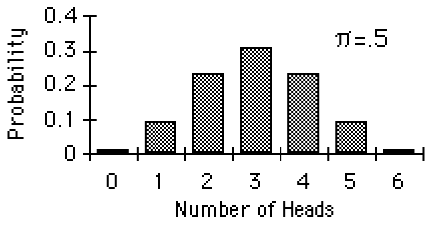
Consider this simple application of the binomial distribution: What is the probability of obtaining exactly 3 heads if a fair coin is flipped 6 times?

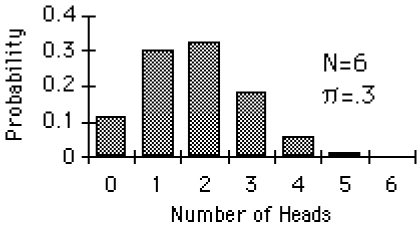
Binomial Distribution (2 of 3)

For this problem, N = 6, r = 3, and π = 0.5. Therefore,

$$P(3) = \frac{6!}{3!\,(6-3)!}\,(0.5)^{3}\,(1-0.5)^{6-3}$$
$$= \frac{6 \text{ x } 5 \text{ x } 4 \text{ x } 3 \text{ x } 2}{(3 \text{ x } 2)(3 \text{ x } 2)}\,(0.125)(0.125) = 0.3125.$$

Two binomial distributions are shown below. Notice that for π = 0.5, the distribution is symmetric whereas for π = 0.3, the distribution has a positive skew.





Binomial distribution (3 of 3)

Often the cumulative form of the binomial distribution is used. To determine the probability of obtaining 3 or more successes with n=6 and π = 0.3, you compute P(3) + P(4) + P(5) + P(6). This can also be written as:

$$\sum_{i=3}^{6} P(r_i)$$

and is equal to 0.1852 + 0.0595 + 0.0102 + 0.0007 = 0.2556. The binomial distribution can be approximated by a normal distribution (click here to see how). Click here for an interactive demonstration of the normal approximation to the binomial.

Standard Error

The standard error of a statistic is the standard deviation of the sampling distribution of that statistic. Standard errors are important because they reflect how much sampling fluctuation a statistic will show. The inferential statistics involved in the construction of confidence intervals and significance testing are based on standard errors. The standard error of a statistic depends on the sample size. In general, the larger the sample size the smaller the standard error. The standard error of a statistic is usually designated by the Greek letter sigma (σ) with a subscript indicating the statistic. For instance, the standard error of the mean is indicated by the symbol: σM.

Correlation

The correlation between two variables represents the degree to which variables are related. Typically the linear relationship is measured with either Pearson's correlation or Spearman's rho. It is important to keep in mind that correlation does not necessarily mean causation. For example, there is a high positive relationship between the number of fire fighters sent to a fire and the amount of damage done. Does this mean that the fire fighters cause the damage? Or is it more likely that the bigger the fire, the more fire fighters are sent and the more damage that is done. In this example, the variable "size of the fire" is the causal variable, correlating with both the number of fire fighters sent and the amount of damage done.