

POL 602
Prof. Matthew Lebo
November 1st and 3rd, 2005
Estimation, Confidence Intervals, and Point Estimators

Estimation

As we have mentioned several times, the objective of statistics is to draw inference about a population based on a sample. Most statistical inference procedures involve either *estimation* or *hypothesis testing*. For the next few classes, we will be concerned with the former; later in the semester we will spend a lot of time on hypothesis testing as well.

Suppose we have a *target parameter*, which we will label in general terms θ , such as the average age of voters in the United States. If we wish to estimate this parameter using a sample from the population, such as the ANES, we can imagine two distinct types of estimates ($\hat{\theta}$).

We could try and find a single value, such as 45 years old. This is known as a *point estimate*. Alternatively, we could construct an interval that is intended to enclose the parameter of interest, such as 40-50 years old. This is called an *interval estimate*.

To find either type of estimate, we need a rule or a recipe or an algorithm that tells us how to calculate the value of an estimate based on the measurements contained in a sample. This is an *estimator*.

In our example, we are trying to calculate the population mean. One estimator could be a rule such as: take a sample, add up the age of every observation and divide it by the current year.

It should be obvious that this is not a good estimator for the population mean age, unless our sample size happens to be very close to the current year. A better rule is: take a sample, add up the age of every observation and divide by the sample size, n . This is most frequently expressed as the familiar formula:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

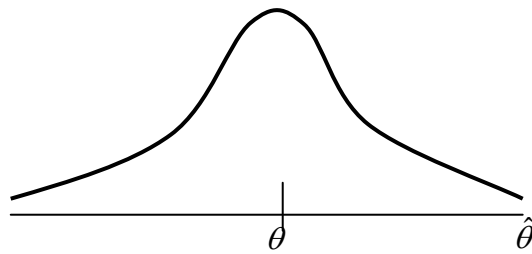
In this case, the sample mean \bar{Y} is our estimator for μ , the population mean.

This example brings up a point that we have discussed previously concerning the quality of estimates: we think that the sample mean is a *good* estimator of the population mean, but, in some way, the division by current year estimate of the sample mean is a *bad* estimator.

One of our main goals today is to formalize our understanding of what makes a particular estimator better than its competitors.

Last time we discussed the concept of the sampling distribution—either empirically or theoretically derived as the distribution of statistic (or estimator) over repeated samples.

Assume we are interested in estimating a population parameter θ and we have devised a formula—a point estimator $\hat{\theta}$. If we could take many samples and apply our estimator to the samples, we could construct a sampling distribution:



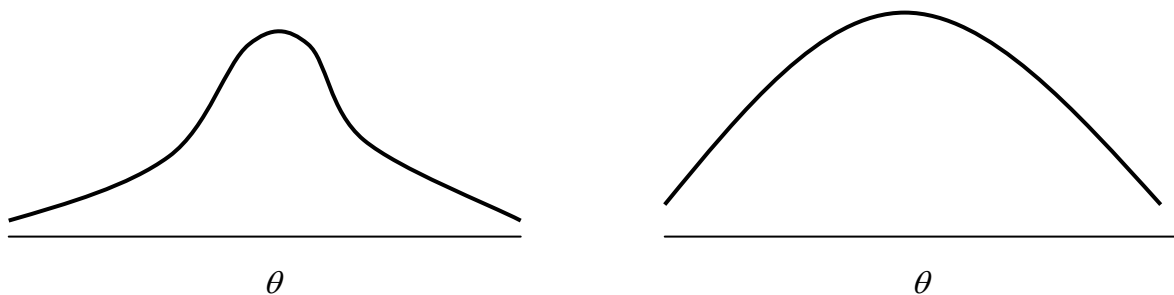
In this case (assuming a roughly normal sampling distribution), it looks as if the expected value of the sampling distribution is equal to the parameter of interest. This leads to our first important definition:

Let $\hat{\theta}$ be a point estimator for θ . If $E(\hat{\theta}) = \theta$, then the estimator is said to be *unbiased*. Otherwise, $\hat{\theta}$ is *biased*.

The bias of a point estimator is simply:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Next, let us consider the sampling distributions of two different unbiased estimators:



Which one would we prefer? The first one has a smaller variance than the second—the spread or dispersion of points is tighter around the expected value.

Since for the most part we are only going to be able to draw a small number (like 1) of samples, we would prefer, other things being equal, to have an estimator with a smaller variance to increase our likelihood of estimating the true population parameter accurately.

In this example, we are comparing the variances of two unbiased estimators (this does not always have to be the case; see below). When this is the case, we can use the ratio of their variances to define the *relative efficiency* of one unbiased estimator to another:

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}$$

which gives us the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$.

There is a simple way (in theory, not always in practice as we shall see) to take into account both bias *and* variance in comparing estimators: the *mean squared error* or (MSE).

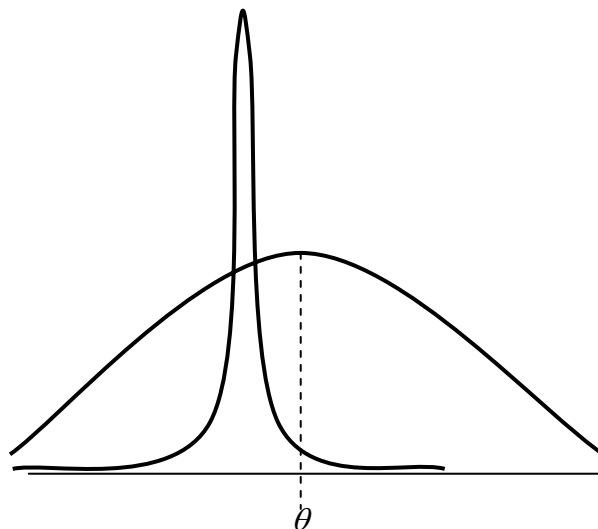
$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

which can be shown to equal:

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

or, the variance of the estimator plus the bias squared. It is a primary goal of estimation to choose an estimator that minimizes MSE. This is not always easy or even possible.

Consider:



One sampling distribution is unbiased but has a large variance. The other is clearly biased, but has very little dispersion. Which one you choose depends on your goals of inference, your tolerance for bias etc. [Note: this is discussed in Jeff's class as well].

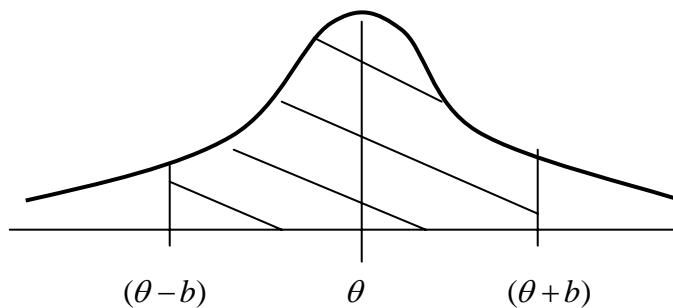
Yet another way to think about the goodness of a point estimator is to think about the difference between a single estimate and the population parameter. This is, itself a random variable that is closely related to the bias (bias is, after all, the expected value of this difference).

If we consider the absolute value of this difference, we call this quantity the *error of estimation*:

$$\varepsilon = |\hat{\theta} - \theta|$$

It is obvious that we would like this quantity to be as small as possible. As I mentioned, the error of estimation is a random variable (because θ -hat is a random variable), so we can not say how large or small it will be for a given estimate.

We can, however, make probability statements about it. For example, assume we have a sampling distribution for an estimator that looks like this:



If we pick two points near the tails of the distribution ($\theta \pm b$), we can say that the probability that the error of estimation is less than b is the area under the curve of the sampling distribution between the two points (the shaded area). That is,

$$P(|\hat{\theta} - \theta| < b) = P(\theta - b < \hat{\theta} < \theta + b)$$

This allows us to put a probabilistic bound on the error of estimation. This probability identifies the fraction of times, in repeated sampling, that the estimator falls within b units of the target parameter.

If b is small enough (depending on the particular application), this probability can thus serve as a measure of the goodness of a single estimate.

To think about this another way, what if we want to find the value of b such that $P(\varepsilon < b) < .90$? Mathematically we are looking for a value of b such that:

$$\int_{\theta-b}^{\theta+b} f(\hat{\theta})d\hat{\theta} = .90$$

If we don't know the sampling distribution of the estimator, we can still find an approximate bound on the error of estimation by expressing b as a multiple of the standard error of the estimator and using Tchebysheff's theorem.

Thus if $b = k\sigma_{\hat{\theta}}$, then the error of estimation has a probability of at least $1 - \frac{1}{k^2}$ of being less than $k\sigma_{\hat{\theta}}$. For example, if $k = 2$ standard deviations, we know that:

$$P(\varepsilon < 2\sigma_{\hat{\theta}}) \geq .75$$

In nature and in social applications, it turns out, this Tchebyseff result is conservative.

Actually, most random variables lie within two standard deviations of their means with a probability of about .95 (as per the empirical rule).

A final property of point estimators that I would like to discuss is *consistency*. An estimator is said to be consistent if, for any positive number δ :

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta} - \theta\right| \leq \delta\right) = 1$$

Or, the probability that the error of estimation is less than some arbitrarily small number is equal to 1 as the sample size approaches infinity. In other words, as the sample size increases, the estimator "converges in probability" to the quantity being estimated. This result is sometimes referred to as the *weak law of large numbers*.

A related theorem, useful for us, is that an unbiased estimator is a consistent estimator if:

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$$

[proof in Chapter 9 in the text].

For example, take the estimator \bar{Y} , the sample mean, for the population mean, μ .

We know from previous results that $E(\bar{Y}) = \mu$ and $V(\bar{Y}) = \frac{\sigma^2}{n}$.

Thus, as n increases towards infinity, the variance of the estimator approaches 0. \bar{Y} is thus a consistent estimator for the population mean.

Next I would like to present some common unbiased point estimators for quantities of practical interest to us as researchers in many applications. These are also presented in Table 8.1 on page 371 of the text.

For the most part, we will not derive these estimators or their standard errors (which come from theoretical models of their sampling distributions) although we will briefly discuss how one might derive such estimators in a few weeks. For now, I will simply discuss them intuitively.

The four target parameters we wish to estimate are: the population mean, a population proportion (sometimes called a *binomial parameter*), the difference in means between two populations, and the difference in proportions between two populations.

Target Parameter	Sample Size(s)	Point Estimator	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	\bar{Y}	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\bar{Y}_1 - \bar{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Note that the last column, the *standard error* is the standard deviation of the sampling distribution. Two other points bear mentioning regarding this table.

First, the expected values and standard errors are valid regardless of the form of the population probability density function used (i.e. we could be talking about the difference in means between two normals or two exponential distributions).

Second, the sampling distributions of all four estimators are approximately normal for large samples. This is obvious for the sample mean and sample proportion—this is just the Central Limit Theorem again. Similar results can be shown for the other two estimators.

This again raises the question, how large is “large” ? The text notes that these estimators approach normality in the sampling distribution when n is approximately 30 for the sample mean and difference of means, and that the necessary n is a function of p for the other cases: $n > 9 \frac{\max(p, q)}{\min(p, q)}$.

Example (8.2): Out of 1000 randomly-selected voters, 560 favor Jones. Estimate the fraction of voters favoring Jones in the population and place a 2 standard error bound on the error of estimation. To find the proportion, use the estimator:

$$\hat{p} = \frac{Y}{n} = \frac{560}{1000} = .56$$

Now in calculating the standard error, however, we have a problem—we don't know the actual value of p :

$$b = 2\sigma_{\hat{p}} = 2\sqrt{\frac{pq}{n}}$$

The solution is to use the estimated proportion instead (which, for a large n such 1000 is a reliable estimate):

$$b = 2\sigma_{\hat{p}} = 2\sqrt{\frac{pq}{n}} \approx 2\sqrt{\frac{(.56)(.44)}{1000}} = .03$$

What does this mean? The probability that the error of estimation is less than .03 is approximately .95 (from earlier discussion).

Thus, we can be confident that our estimate, .56, is within .03 of the true value with a probability of .95.

Example (8.3): Comparison of durability of two types of tire.

$$\begin{aligned} \bar{Y}_1 &= 26,400 & \bar{Y}_2 &= 25,100 \\ s_1^2 &= 1,440,000 & s_2^2 &= 1,960,000 \end{aligned}$$

Estimate the difference in mean miles to wear-out and place a 2 standard error bound on the error of estimation:

The point estimate is

$$(y_1 - y_2) = 26,400 - 25,100 = 1,300$$

miles before wearout.

We don't know, in this case, the variances of the populations (just as we didn't know the parameter p in the previous example). As the text points out, however, if both n 's are suitably large (say, over 30), we can use the sample variances as a reasonable estimate:

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{1,440,000}{100} + \frac{1,960,000}{100}} = 184.4$$

184.4 times 2 (we asked for 2 s.d.'s) is 368.8. Therefore, the difference in means is 1300 miles and we expect the error of estimation to be less than 368.8 miles with a probability of about .95.

Homework Problems: 8.17, 8.19, 8.20, 8.21, 8.23, 8.24, 8.25, 8.28, 8.29, 8.33.

Confidence Intervals

The table of estimators introduced last time consisted of only point estimators.

Recall that we also discussed the existence of *interval estimators* that are defined as (from the text): rules specifying the method for using the sample measurements to calculate two numbers that form the endpoints of the interval. There are two desirable characteristics of interval estimators:

1. They contain the target parameter, θ , of interest.
2. They are relatively narrow.

It is important to emphasize that, using frequentist logic, the size and location of an interval estimate are themselves random variables (just as point estimators are).

In any given sample, the estimate either contains the target parameter value or it does not.

Only in (hypothetical or empirical) repeated sampling can we say anything about the expected probability that a given interval estimate encloses θ .

Interval estimators are commonly referred to as *confidence intervals*. The upper and lower endpoints are called the *upper/lower confidence limits*.

The fraction of the time that, in repeated sampling, the intervals will contain the target parameter is called the *confidence coefficient*.

If we know that the confidence coefficient associated with our estimator is high, we can be confident that a confidence interval constructed with the results of a single sample will enclose θ .

We can summarize these words with the following relationship:

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

where $(1 - \alpha)$ is the confidence coefficient and the random interval between the lower and upper limits $[\hat{\theta}_L, \hat{\theta}_U]$ is called a *two-sided confidence interval*.

It is also possible to form a *one-sided confidence interval*:

$$P(\hat{\theta}_L \leq \theta) = 1 - \alpha$$

$$P(\theta \leq \hat{\theta}_U) = 1 - \alpha$$

these imply the confidence intervals: $[\hat{\theta}_L, \infty)$ and $(-\infty, \hat{\theta}_U]$, respectively.

Pivotal Method

NOTE: This is a largely theoretical exercise and one we can safely skip over to get to the more practical methods of building confidence interval.

The text describes a method which they call the *pivotal method* of finding C.I.'s that depends on finding a *pivotal quantity* with two characteristics:

1. It is function only of the sample data and the unknown parameter (i.e. the parameter of interest is the only unknown quantity).
2. Its probability distribution does not depend on the unknown parameter.

In other words, if we know the probability distribution of the pivotal quantity, we will be able to perform simple arithmetic operations (such as adding a constant to all terms) to form the desired confidence interval.

This is slightly easier to see in an example (8.4 in the book), although the example depends upon a method of finding probability distributions of random variables that we have not yet discussed:

Suppose we know that a random variable is distributed exponential with mean θ and we take a single sample, Y , from this distribution. Use this information to construct a confidence interval for θ with a confidence coefficient of .90.

From the beta distribution, the density of Y is:

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-y/\theta}, & y \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

To compute a confidence interval for the mean, we need a simple function of only the observed value and θ to serve as the pivotal quantity, whose probability distribution is independent of θ . This function is $U = \frac{Y}{\theta}$ which, itself, is distributed exponentially:

$$f_U(u) = \begin{cases} e^{-u}, & u > 0 \\ 0, & \text{elsewhere} \end{cases}$$

[Note: this is the magic hand-waving part. The text finds this distribution using the method of transformations (section 6.4) which we did not discuss. This is a method for finding density functions for functions of random variables that are either always increasing or always decreasing in a certain range. If we have a function $h(y)$, we find its density function by taking its inverse and multiplying it by the absolute value of the first derivative of its inverse. In this case:

$$u = h(y) = \frac{y}{\theta}$$

$$y = h^{-1}(u) = \theta u$$

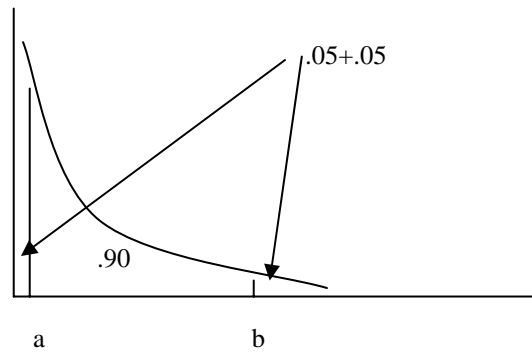
$$\frac{d(h^{-1}(u))}{du} = \theta$$

$$f_U(u) = f_Y(h^{-1}(u)) \left| \frac{d(h^{-1}(u))}{du} \right|$$

$$f_U(u) = \frac{1}{\theta} e^{-\theta u / \theta} |\theta| = e^{-u}$$

so the function U turns out to also be exponentially distributed.]

Anyway, to find the interval with a .90 confidence coefficient, we want to find two numbers a, b such that $P(a \leq U \leq b) = .90$. We can do this solving two integrals, corresponding to the areas:



$$P(U < a) = \int_0^a e^{-u} = .05$$

$$P(U > b) = \int_b^{\infty} e^{-u} = .05$$

So, $1 - e^{-a} = .05$ and $e^{-b} = .05$, thus $a = .051, b = 2.996$. We can thus say that:

$$\begin{aligned}
P(.051 \leq U \leq 2.996) &= .90 \\
P\left(.051 \leq \frac{Y}{\theta} \leq 2.996\right) &= .90 \\
P\left(\frac{.051}{Y} \leq \frac{1}{\theta} \leq \frac{2.996}{Y}\right) &= .90 \\
P\left(\frac{Y}{.051} \geq \theta \geq \frac{Y}{2.996}\right) &= .90 = P\left(\frac{Y}{2.996} \leq \theta \leq \frac{Y}{.051}\right)
\end{aligned}$$

In words, if we take a single draw from an exponential distribution, we can say that in 9/10 samples, the mean of the distribution (or the expected value) is between the observation/.051 and the observation/2.996.

For example, if we observe the value 4, 9/10 of the time in repeated samples, the actual mean will be between 78.4314 and 1.335. This is not exactly awe-inspiring precision, but you can't get something for nothing, and a single observation is close to nothing in the world of statistics.

Large-Sample Confidence Intervals

For many useful estimators, it is not necessary to find a pivotal quantity and derive its density function; they have been derived and have been in general use for a long time.

In this section, we will discuss large-sample confidence intervals for the point estimates we presented in table form last time: μ , p , $\mu_1 - \mu_2$, $p_1 - p_2$.

The derivation for the confidence intervals of all four of these is very simple. Since we are assuming large sample properties (recall the limits on n from last class: about 30 for means and the formula for proportions depending on p), we are further assuming that the central limit theorem holds.

That is, the estimators have an approximately normal sampling distribution. If we standardize, we get:

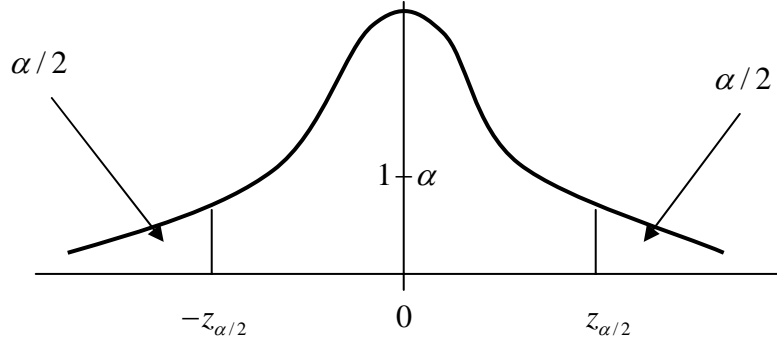
$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

where Z has an approximate standard normal distribution and can be used as the pivotal quantity (it depends only on the data and is distributed independently of θ) to develop confidence intervals.

What is the form of these confidence intervals? First, let us select two values in the tails of this sampling distribution, which we will label: $z_{\alpha/2}$ and $z_{-\alpha/2}$ such that

$$P(z_{-\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

Graphically, we are considering the area:



Substituting for Z yields:

$$P\left(z_{-\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

multiplying each term by the standard error:

$$P\left(z_{-\alpha/2}\sigma_{\hat{\theta}} \leq \hat{\theta} - \theta \leq z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

subtract theta-hat from each term:

$$P\left(z_{-\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta} \leq -\theta \leq z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta}\right) = 1 - \alpha$$

since the standard normal is symmetrical, $z_{-\alpha/2}$ is equal to $-z_{\alpha/2}$:

$$P\left(-z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta} \leq -\theta \leq z_{\alpha/2}\sigma_{\hat{\theta}} - \hat{\theta}\right) = 1 - \alpha$$

now, divide all terms through by -1 , not forgetting to “flip” the signs:

$$P\left(z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta} \geq \theta \geq -z_{\alpha/2}\sigma_{\hat{\theta}} + \hat{\theta}\right) = 1 - \alpha$$

and re-organize:

$$P\left(\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}} \geq \theta \geq \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

and, finally, we can just rewrite the inequality to read the other way:

$$P\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

Thus, the endpoints (or *lower and upper* bounds) for a $100(1 - \alpha)$ % confidence interval for a parameter theta are:

$$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$$

We can also construct one-sided confidence limits:

$$100(1 - \alpha) \% \text{ lower bound for } \theta = \hat{\theta} - z_{\alpha}\sigma_{\hat{\theta}}$$

$$100(1 - \alpha) \% \text{ upper bound for } \theta = \hat{\theta} + z_{\alpha}\sigma_{\hat{\theta}}$$

These results hold for all of the quantities of interest in the table that we presented earlier.

Example (8.7): 64 randomly-selected shoppers in a supermarket had a mean shopping time of 33 minutes with variance 256 minutes. Estimate the true population average shopping time with a confidence coefficient of $1 - \alpha = .90$.

Solution: we are interested in estimating μ . Our point estimate is thus:

$$\hat{\theta} = \bar{Y} = 33$$

The population variance is unknown, so we will again use the sample variance as an estimate (more on this later). Our confidence intervals thus have the form:

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}} \approx \bar{Y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

α , in this case, is equal to .10, so $\alpha/2 = .05$. From Appendix III, Table 4,

$$z_{.05} = 1.645$$

so the confidence intervals are given by: $33 \pm 1.645\left(\frac{16}{8}\right) = (29.71, 36.29)$.

Remember – this is the confusing part of frequentist statistics – that it is *not* true that there is a 90% chance that this interval contains the parameter’s true value. All we can

say in frequentist terms is that the estimator $\bar{Y} \pm 1.645\left(\frac{s}{\sqrt{n}}\right)$ contains μ in 9/10 samples

as long as n is at least 30.

In other words, saying we are confident “at the 90% level” means that *the procedure* used to find the confidence interval contains the true parameter 90% of the time.

In our single sample, the probability of the estimated interval including the parameter is either 0 or 1 – *it either does or it doesn't*.

The population mean is assumed to be a fixed value. If we want to make statements about the probability of this parameter being within a certain range, we must shift our philosophical foundation of probability to the subjective, or Bayesian, approach, which assumes that our data are fixed and our parameters are actually random variables.

Another example (8.8): Two brands of refrigerators, A and B, are each guaranteed for 1 year. In a sample of 50 type A fridges, 12 failed within the first year. Of a second sample of 60 B fridges, 12 also failed within the guarantee period. Estimate the difference in proportions with a confidence interval of .98.

Solution: the general confidence interval

$$\hat{\theta} \pm z_{\alpha/2} \sigma_{\hat{\theta}}$$

now has the specific form (from the point estimator table):

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

and we have the same problem as we had last time—we do not know the population p 's.

Again, we use our estimates and also read the z-score for .01 from the table (2.33):

$$(.24 - .20) \pm 2.33 \sqrt{\frac{(.24)(.76)}{50} + \frac{(.20)(.80)}{60}} = .04 \pm .1851 = (-.1451, .2251)$$

Note that this interval includes 0, which would be our estimate of no difference in proportions. This will be important when we talk about hypothesis testing later.

Some more examples....

Another example for means:

Let X equal the amount of orange juice (in grams per day) consumed by an American. Suppose it is known that the standard deviation of X is $\sigma = 96$. To estimate the mean of X , and orange growers' association took a random sample of $n = 576$ Americans and found that they consumed, on average, $\bar{X} = 133$ grams of orange juice per day. Thus, an approximate 90% confidence interval for μ is:

$$133 \pm 1.645 \left(\frac{96}{\sqrt{576}} \right) \text{ or } [133 - 6.58, 133 + 6.58] = [126.42, 139.58].$$

Note: If σ^2 is unknown and the sample size n is 30 or greater, we use the fact that the ratio $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ has an approximate normal distribution $N(0,1)$. This statement is true whether or not the underlying distribution is normal (CLT, again).

However, if the underlying distribution is badly skewed or contaminated with occasional outliers, most statisticians would prefer to have a larger sample size, say 50 or more.

Example For Proportions:

In a certain political campaign, one candidate has a poll taken at random among the voting population. The results are $y=185$ out of $n=351$ voters favor this candidate. Even though $y/n = 0.527$, should the candidate feel very confident of winning? An approximate 95% confidence interval for the fraction p of the voting population who favor this candidate is:

$$0.527 \pm 1.96 \sqrt{\frac{(0.527)(0.473)}{351}} \text{ or: } [0.475, 0.579]. \text{ There is a good possibility that } p < 0.5.$$

A Note on Experimental Design

A quick word about experimental design, which those of you in the psychology path will have an entire course on soon, and the rest of you will still need to know something about when you commission your own surveys or otherwise gather your own data.

One of the most important decisions is the sample size. We know that larger n is better, but also more costly. The concept of interval estimators or confidence intervals, allows us to make an educated guess about the required sample size for a given level of confidence. If the parameter we are interested in is θ , and we have a desired bound on the error of estimation B , for large samples, $z_{\alpha/2} \sigma_{\hat{\theta}} = B$. This is sufficient to allow us to find an approximate required sample size.

Example (8.9): the reaction of an individual to a psychological stimulus in an experiment may take one of two forms (dichotomous random variable), A or B. If the experimenter wants to estimate the probability (proportion in this case) that a subject reacts in manner A, how many people should be included in this experiment, assuming she wants a bound on the error of estimation of .04, with probability equal to .90 and that she thinks the proportion of A responses in the population is about .6.

Solution: $1 - \alpha = .90$, so $\alpha = .10$ and $\alpha/2 = .05$. From the table, we know that $z_{\alpha/2} = 1.645$. We thus can say that $1.645 \sigma_{\hat{p}} = .04$. Substituting the formula for the standard error (again, from the table):

$$1.645 \sqrt{\frac{.6(.4)}{n}} = .04, n = 406$$

so she needs an n of 406 to estimate the proportion with 90% confidence.

Note that this result depends on the estimate of the proportion. Often, we do not really have a good idea, *a priori*, of this quantity. However, we know that the maximum variance for \hat{p} occurs when $p=.5$. Thus, to find the most conservative guess about sample size, we can use $.5$. In this example, the required n would be 423. Thus, whatever the true value of p , an n of 423 is large enough to provide an estimate within $.04$ of p with probability $.90$.

Another example:

Suppose we think that the unemployment rate has been about 8% (that is, $p=.08$) but we wish to update our estimate and be very accurate about it – lets say we want to be 99% confident that the new estimate of p is within 0.001 of the true p . How many people should we sample?

$$\hat{p} \text{ will just be } \frac{y}{n} \text{ and the 99\% confidence interval will be: } \frac{y}{n} \pm 2.576 \sqrt{\frac{(y/n)(1-y/n)}{n}}$$

Though we don't know $\frac{y}{n}$ exactly before sampling, we do know, since it will be near

$$0.08 \text{ that: } 2.576 \sqrt{\frac{(y/n)(1-y/n)}{n}} \approx 2.576 \sqrt{\frac{(0.08)(0.92)}{n}} \text{ and we want this number to equal } 0.001.$$

$$\text{That is, } 2.576 \sqrt{\frac{(0.08)(0.92)}{n}} = 0.001.$$

$$\text{Or, } \sqrt{n} = 2576 \sqrt{0.0736} \text{ and } n \approx 488,394.$$

A couple lessons from this example: reliability and accuracy are costly! Reducing the desired accuracy to $.98$ and the accuracy to $.01$ our needed n drops to 3,982. Also, recognize the importance of sample size when given a single point estimate.

Small Sample Confidence Intervals

Now we want to consider the case of trying to estimate either the population mean or the difference in two means when n is less than about 30 and the population variance is unknown (this is not an uncommon case in political science).

In this case, we can no longer rely upon the central limit theorem to tell us that the sampling distribution is approximately normal.

If, however, we can assume that the underlying random variable of interest is distributed approximately normal (an unnecessary assumption for the central limit theorem, recall), we can find the sample mean and variance and construct a new standardized statistic:

$$T = \frac{\bar{Y} - \mu}{S / \sqrt{n}}$$

which is distributed t with $(n-1)$ degrees of freedom.

This is because the t distribution is actually a family of distributions with a different shape for each sample size—and it approaches the normal distribution as the sample size increases.

Note that the t distribution, or *Student's t*, is named for its discoverer, William Gosset who was a brewmeister for the Guinness brewery which stipulated that he not publish using his real name. This test was originally developed as part of beer quality control analysis! Gosset's basic question was how to judge if all the beer was palatable without drinking it all.

The density function of t is particularly ugly:

$$f(t) = \left(\frac{\Gamma(\nu+1)/2}{\sqrt{\pi\nu}\Gamma(\nu/2)} \right) \left(1 + \frac{t^2}{\nu} \right)^{-(\nu+1)/2}$$

but fortunately we can use a table to find areas under it instead of attempting to integrate—Appendix III, table 5, pg. 793.

Reading from this table, we can find values such that:

$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$$

so the confidence interval for the sample mean can be expressed as:

$$\bar{Y} \pm t_{\alpha/2} \left(\frac{S}{\sqrt{n}} \right)$$

Note too that one-sided confidence limits can be constructed as in the earlier discussion of large sample confidence intervals.

Example (8.11): a manufacturer of gunpowder has developed a new powder which was tested in 8 shells. The mean muzzle velocity was 2959 fps and the standard deviation was 39.1. Find a 95% confidence interval for the population average velocity assuming that muzzle velocities are approximately normally-distributed.

Solution: given this assumption, we can use the t -distribution to compute this confidence interval: $\bar{Y} \pm t_{\alpha/2} \left(\frac{S}{\sqrt{n}} \right)$. Using the table, we find $t(.025, 7 \text{ d.f.})=2.365$. So,

$$2959 \pm 2.365 \left(\frac{39.1}{\sqrt{8}} \right) = 2959 \pm 32.7$$

as the observed confidence interval for the population mean.

What if we are interesting in finding a small-sample confidence interval for the difference of two means of approximately normal populations?

In this case, we need to make a further assumption, that the variances of the two populations are equal ($\sigma_1^2 = \sigma_2^2$), but unknown.

To construct this confidence interval, we start with the pivotal quantity:

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

as the variances are assumed to be equal, we can rewrite this:

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

since this population standard deviation is unknown, we need to construct an estimator for it. We do this by pooling the sample data to create the pooled estimator:

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

which is, essentially, a weighted average of the variances of the two samples with the larger weight given to the sample of larger size.

We can then use this estimator to create a random variable W :

$$W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2}{\sigma^2} + \frac{\sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{\sigma^2}$$

which can be shown (see pg. 402 in the text for an extended discussion) to be the sum of two independently distributed chi-square variables. Thus, W itself is distributed chi-square with $n_1 + n_2 - 2$ degrees of freedom.

Finally, we can use the standard normal variable and this chi-square variable together to form a new random variable that is distributed t:

$$T = \frac{Z}{\sqrt{W/V}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2 \text{ d.f.})$$

Confidence intervals can thus be constructed using:

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\alpha/2}$ is taken from the table of values as usual.

Note that if the variances are not assumed to be equal, the problem is more difficult. It can be solved, however, if an auxiliary assumption about the ratio of the variances is made.

If, however, we have no information about the variances at all, this becomes difficult indeed. Estimation under these conditions is known as the *Berhens-Fisher Problem*, and there are no solutions that are universally accepted.

Example (8.12): To compare two different training programs for assembling a device in a factory, two groups of 9 employees were trained in the two different methods. At the end of the training, employees in the first group had a mean time of assembly of 35.22 minutes with a sample variance of 24.445. The second group had mean of 31.56 and variance 20.027.

Estimate the difference in means with a confidence coefficient of .95. Assume normally-distributed times with identical variances.

Solution: the first step is to compute the pooled estimate of variance:

$$s_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{8(24.445) + 8(20.027)}{9 + 9 - 2} = 22.236$$

and $s_p = 4.716$.

Next, find the appropriate value from the table: $t_{.025} = 2.120$ (16 d.f.). The confidence interval is thus:

$$(35.22 - 31.56) \pm (2.120)(4.716) \sqrt{\frac{1}{9} + \frac{1}{9}} = 3.66 \pm 4.71$$

note that this interval also contains 0. At this confidence level, it is thus not possible to distinguish the two training programs.

Another example:

Suppose that scores on a standardized test in mathematics taken by students from large and small schools are $\sim N(\mu_x, \sigma^2)$ and $N(\mu_y, \sigma^2)$ respectively where σ^2 is unknown. If a random sample of $n=9$ students from large high schools yielded $\bar{x}=81.31$, $s_x^2 = 60.76$ and a random sample of $m=15$ students from small high schools yielded $\bar{y} = 78.61$, $s_y^2 = 48.24$, the endpoints for a 95% confidence interval for $\mu_x - \mu_y$ are given by:

$$81.31 - 78.61 \pm 2.074 \sqrt{\frac{8(60.76) + 14(48.24)}{22}} \sqrt{\frac{1}{9} + \frac{1}{15}}$$

because $t_{0.025}(22) = 2.074$. The 95% confidence interval is [-3.65, 9.05].

Confidence Intervals for Variance

The last type of confidence interval I will present is an interval estimate for the population variance.

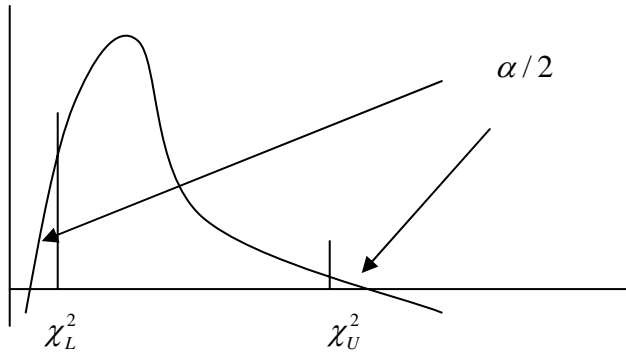
We are often interested not only in estimating means and proportions or differences between them, but in estimating the variance of a population. As in our previous discussions, we require a pivotal quantity. In this case, it can be shown that:

$$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2$$

with $(n-1)$ degrees of freedom. We can thus state that:

$$P \left[\chi_L^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_U^2 \right] = 1 - \alpha$$

since the chi-square distribution is not symmetric, we can not simply divide alpha by 2 and look up a single value in the chi-square table. To make our task relatively simple, by convention we find areas of $\alpha/2$ and $1 - (\alpha/2)$.



thus we obtain:

$$P \left[\chi_{\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-(\alpha/2)}^2 \right] = 1 - \alpha$$

solving for the variance and reordering the inequality yields:

$$P \left[\frac{(n-1)S^2}{\chi_{1-(\alpha/2)}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2}^2} \right] = 1 - \alpha$$

so, the 100(1-alpha)% confidence interval is thus:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{1-(\alpha/2)}^2} \right)$$

Example (8.13): An experimenter wants to check the validity of measurements, 4.1, 5.2 and 10.2. Estimate the true variance with a .90 confidence coefficient.

Solution: the sample variance is found to be 10.57. From Appendix III, Table 6, we find the two chi-square values for $\alpha/2 = .05$, 2 d.f.: $\chi_{.05}^2 = 5.991$, $\chi_{.95}^2 = .103$. Thus, the confidence interval is:

$$\left(\frac{2(10.57)}{5.991}, \frac{2(10.57)}{.103} \right) = (3.53, 205.24)$$

which is very wide, since n is very small.

Homework problems for confidence intervals: 8.42, 8.43, 8.45, 8.47, 8.48, 8.50, 8.51, 8.53, 8.54, 8.55, 8.59, 8.60, 8.62, 8.63, 8.69, 8.70, 8.71, 8.73, 8.77, 8.81, 8.82, 8.85, 8.87, 8.89.

Some Intuition for Confidence Intervals

Consider an archer shooting at a target. Suppose she hits the 10 cm bull's eye 95% of the time. That is, only one arrow of 20 misses.

Sitting behind the target (bravely) is our statistician who can't see the bull's eye, but can see the end of the arrow coming through the target.

The archer shoots a single arrow. Knowing the archer's skill level, the detective draws a circle with a 10 cm radius around the arrow. Now the statistician has *95% confidence* that his circle includes the center of the bull's eye.

He reasons that if he drew 10 cm radius circles around many arrows, his circles would include the center 95% of the time.

What does this mean given all that we have learned?

If p is the center of the bull's eye and \hat{p} s are the arrows, we know that the distribution of the arrows is nearly normal with a mean of p and a standard deviation $\sigma_{\hat{p}} = \frac{\sqrt{pq}}{\sqrt{n}}$.

Since the curve is normal, we can use the Z -transformation and a standard table to find the width of the interval within which 95% of the arrows hit. Since the Z table has 2.5% of the curve beyond -1.96 and another 2.5% of the curve beyond $+1.96$, we find the 95% area – the radius of the bull's eye – to be within 1.96 standard deviations.

This gives us: $P\left(-1.96 \leq \frac{\hat{p} - p}{\sigma_p} \leq 1.96\right) = .95$ which is equivalent to:

$$P\left(\hat{p} - 1.96\sigma_p \leq p \leq \hat{p} + 1.96\sigma_p\right) = .95$$

This equivalent to saying we are drawing circles around a lot of arrows and saying that 95% of them cover p .

Of course, we have a problem...we don't actually know the size of the bull's eye because we don't know p and the width is a multiple of σ_p .

So we cheat a little and use the **standard error** of \hat{p} : $SE_{\hat{p}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ in its place.

This is the best we can do and our consolation is that it can be theoretically justified as, on average, correct.

Now we can have a formula that relies only on information our sample gives us: $\hat{p} \pm 1.96SE_{\hat{p}} = .550 \pm (1.96)(.0157) = .550 \pm .031$

$$P(\hat{p} - 1.96SE_{\hat{p}} \leq p \leq \hat{p} + 1.96SE_{\hat{p}}) = .95.$$

If we take one sample and calculate \hat{p} and $SE_{\hat{p}}$ we can easily get confidence intervals.

For example, given a sample of 1000 voters, 550 of who disapprove of Bush, what can we say about the population?

$$\hat{p} = .55 \text{ and } SE_{\hat{p}} = \frac{\sqrt{(.55)(.45)}}{\sqrt{1000}} = .0157.$$

We have 95% confidence that p is within the range:

$$\hat{p} \pm 1.96SE_{\hat{p}} = .550 \pm (1.96)(.0157) = .550 \pm .031$$

Alternatively, and more correctly, we can say that if p is outside that range, there is only a 5% chance that we would draw such a sample.

Now, if we want to increase our confidence level, we can do two things.

First, we can increase the size of the circle we draw around the arrows, that is increase the area under the normal curve, giving us higher values of Z .

Second, we can improve our aim, by increasing the size of our sample.

Deriving Point Estimators

Next, I want to discuss, briefly, two methods that one can employ to find estimators for population parameters of interest. Up until this point, (with the possible exception of the first example in the section on confidence intervals) we have used only intuition to derive our estimators.

The first method we will discuss is the so-called *method of moments*, one of the oldest methods for deriving point estimators. Recall that the k^{th} moment of a random variable, taken about the origin is:

$$\mu'_k = E(Y^k)$$

the corresponding k th sample moment, then, is the average of the observed values raised to the k :

$$m'_k = \frac{1}{n} \sum_{i=1}^n Y_i^k$$

The method of moments assumes, somewhat intuitively, that the sample moment should be a good estimator of the population moment. The method is simple: set the population moment equal to the sample moment, and solve for the desired estimator.

Example (9.11): A random sample of n observations, Y_1, Y_2, \dots, Y_n , is selected from a population in which Y is distributed uniform on the continuous interval $(0, \theta)$. Use the method of moments to estimate the unknown endpoint of the interval (θ).

Solution: for a uniform (continuous) random variable, the first moment about the origin is equal to (from our discussion at the beginning of the section on continuous random variables):

$$\mu'_1 = E(Y^1) = \mu = \frac{\theta_2 + \theta_1}{2} = \frac{\theta + 0}{2} = \frac{\theta}{2}$$

The corresponding first sample moment is such the simple average:

$$m'_1 = \frac{1}{n} \sum_{i=1}^n Y_i^1 = \bar{Y}$$

setting the two quantities equal and solving:

$$\begin{aligned}\mu'_k &= m'_k \\ \frac{\theta}{2} &= \bar{Y} \\ \hat{\theta} &= 2\bar{Y}\end{aligned}$$

We should check to see if this estimator is unbiased:

$$E(\hat{\theta}) = E(2\bar{Y}) = 2E(\bar{Y}) = 2\mu = 2\frac{\theta}{2} = \theta$$

so the estimator is unbiased. Is it consistent [what is consistency—review]?

$$V(\hat{\theta}) = V(2\bar{Y}) = \frac{2^2\sigma^2}{n} = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}$$

[note: this comes from the variance of $\bar{Y} = \sigma^2/n$ result from the discussion on sampling distributions and from the variance of the uniform distribution: $\theta^2/12$.]

and, as n gets large, we can see that:

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$$

so we have derived a useful estimator that is both unbiased and consistent!

Note that these solutions also get at a question Everett asked a few classes ago – how can we know if an estimator is biased if we don't know the population parameters.

It can be shown, however, that there is another unbiased estimator that is more efficient.

This is illustrated in example 9.1, and the estimator is:

$$\left(\frac{n+1}{n}\right) \max(Y_1, Y_2, \dots, Y_n)$$

This is a general weakness of the method of moments—estimators produced are often not minimum variance unbiased estimators (MVUE's) and sometimes are even biased. On the other hand, the method is easy to understand and to apply.

Maximum Likelihood

An alternative method that leads more often to the derivation of minimum variance unbiased estimators is the method of *maximum likelihood*. Although in this course the examples we will examine are simple and straightforward, this method is central to the creation of many of the more complex estimators that you will study in the third course in this sequence.

The text presents a thought experiment that tries to provide the intuition behind the derivation of maximum likelihood estimators (MLE's). Assume we have an urn that contains three balls, which can be either red or white. We draw two balls from the urn, without replacement. If our sample consists of two red balls, what would be the best estimate of the number of red balls originally in the urn? In other words, if the number of red balls in the urn is a parameter, and we observe data that consists of two red balls,

what is the value of the parameter that maximizes the probability or *likelihood* of observing those data?

By inspection, we can see that there are only two ways we can get 2 red balls: either there are 2 red and 1 white in the urn, or there are 3 red in the urn (i.e. the parameter #red is equal to 2 or 3). What are the associated probabilities of observing the data given these parameter values? For #red=2,

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

and for #red=3,

$$\frac{\binom{3}{2}}{\binom{3}{2}} = 1$$

of these two possible values, the latter (#red=3) maximizes the probability of observing two red balls in our sample. This is thus the MLE for this problem.

More formally, we define the joint probability function (or joint density function) of observing n random variables conditional on k parameters as the likelihood function, L :

$$L(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_k)$$

The method of maximum likelihood requires that we choose as estimates those values of the parameters that maximize this function.

A slightly more involved example will illustrate this (3.10, pg. 105). Suppose that we survey 20 individuals and ask whether they favor or oppose the implementation of a new policy. If we find, in our sample, 6 in favor of the policy, what is the MLE of the true population proportion p ? [From our previous discussion, what have we been using as an estimate for p ?]

Solution: it is reasonable to model the process generating our data using the binomial distribution, which assumes y successes in n independent trials that have a probability of success p and of failure q . [Question, when would this not be reasonable?]

This provides an answer to the question concerning the relationship between the intuitive estimator of p and the binomial distribution. If we do this, we get:

$$P(Y = 6) = \binom{20}{6} p^6 (1-p)^{14}$$

This function requires only the observed data and parameters—it is a likelihood function.

What, then, is the estimate for p that maximizes this function? From calculus, how do we maximize any simple function? We take the first derivative and set it equal to zero. This finds a point where the rate-of-change, or slope of the function is equal to zero.

We then check this point against the endpoints to verify that it is a global maximum (and/or we use the second derivative test).

Looking at this function, however, it does not seem easy to take its derivative with respect to p . Luckily there are some very easy results from mathematics that will greatly simplify this problem.

First, since taking the natural (or any other) logarithm of a function is a simple *monotonic* transformation of that function, we can take the natural log of both sides. The value of p that maximizes the \ln of the function is thus the same as the value of p that maximizes the original function:

$$L = \binom{20}{6} p^6 (1-p)^{14}$$
$$\ln(L) = \ln \left[\binom{20}{6} p^6 (1-p)^{14} \right]$$

this allows to do something that greatly simplifies things.

What is the log of the product of n terms? The sum of the logs of the n terms:

$$\ln(L) = \ln \left[\binom{20}{6} \right] + \ln(p^6) + \ln[(1-p)^{14}]$$

what is another law of logs that can help us here? The log of something raised to an exponent equals the product of the exponent and the log of the base:

$$\ln(L) = \ln \left[\binom{20}{6} \right] + 6 \ln(p) + 14 \ln(1-p)$$

we can now take the first derivative of this function with respect to p , keeping in mind that 20 choose 6 is just a constant:

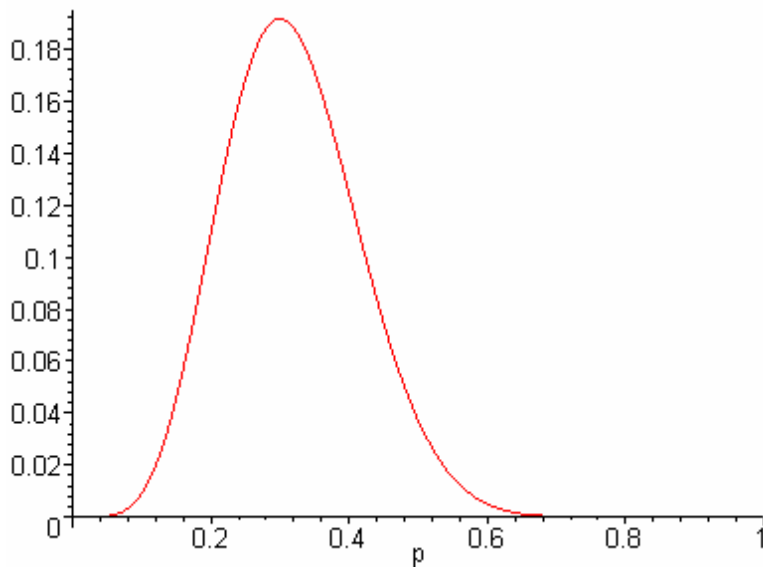
$$\frac{d \ln(L)}{dp} = \frac{6}{p} + \frac{14}{(1-p)}$$

setting this equal to zero and solving:

$$\begin{aligned}\frac{6}{p} - \frac{14}{1-p} &= 0 \\ 6 - 6p &= 14p \\ 6 &= 20p \\ p &= \frac{6}{20}\end{aligned}$$

so, $\hat{p}_{MLE} = 6/20$, the same result as we got using the intuitive estimator!

What does this example look like graphically? Even though it is difficult to take the derivative of the likelihood function without transforming it, we can still inspect a computer-drawn graph of it:



which clearly has a maximum at $6/20$ or $.3$.

More generally, in the case of binomial random variables (or estimating proportions),

$$\hat{p}_{MLE} = \frac{Y}{n}$$