

February 23, 2009

Chapter 5: Estimation

The probability distributions such as the normal, exponential, or binomial are defined in terms of parameters such as μ, σ , for the normal, p for the binomial, and θ for the exponential. If we know the values of these parameters, then we can answer probability questions as we saw in Chapter 4. However, in actual applications, the values of the parameters will not be known. We turn our attention now from probability to statistics. Statistics deals with the problem of estimating parameters and making inference about the values of parameters based on data. The problem of inference involves hypothesis testing which will be introduced in the next chapter.

1 Introduction to Estimation

This chapter deals primarily with the problem of estimating parameters of a probability distribution. The problem of estimating a parameter like the mean μ is more than just giving an estimate of the mean. We also need to provide a sense of the reliability of the estimate.

Random Samples. In order to estimate a parameter of a population of interest, we must have data from that population. A sample is usually obtained from the population. In order for the sample to be representative of the population and in order to make valid statistical inferences or estimations, the sample must be obtained in a scientifically valid way. There are many ways to do this and they all involve random sampling.

If the population is finite, then the simplest method is to collect a *simple random sample*.

Definition. A **Simple Random Sample** of size n is one where each subset of size n from the population has the same chance of being selected for the sample.

Entire books and courses are devoted to the problem of sampling design. Some other popular sampling methods are cluster sampling, two-stage sampling, stratified random sampling and systematic sampling. We will not go into details of these methods here.

In many problems, the population of interest is *effectively infinite*. For example, if we are interested in the population of adults that suffer from depression, we cannot hope to enumerate this population. Most of the examples we shall encounter deal with effectively infinite populations.

Randomized Clinical Trials.

One of the biggest advances in medical research involves the use of randomization when designing studies. Prior to this, different treatments were compared where there may have been little to no control over which patient received which treatment. Without some control over treatments received, it is very difficult to access the effectiveness of treatments.

For example, suppose there are two treatments for a particular illness. Suppose treatment A is more effective than treatment B but has more serious side effects. Sicker patients tend to take treatment A more than treatment B. If we compare the two treatments (A and B), it may appear that treatment B is doing

better than treatment A simply because most of the patients taking treatment A are worse off than those taking treatment B.

In a randomized clinical trial (RCT), the assignment of treatments to patients is done using some random mechanism (randomization). There will be numerous factors that affect how patients will respond to treatments and many of these factors we may not be able to control. If randomization is used, then the differences between these uncontrolled factors will hopefully be washed out. For example, if a large group of patients is randomly divided into two groups, it is unlikely the two groups will differ significantly before receiving a treatment in ways that impact the study under consideration. After treatments are administered to the two groups of patients and a statistically significant difference between treatments is found, then it stands to reason that the difference is most likely due to differences in the treatments and not other factors. Nonetheless, it is fairly common practice to report summary statistics (of ages, sex, race, baseline readings, etc.) for subjects in the different treatment groups for comparison purposes.

2 Estimating the Mean

This section at first may seem trivial. In order to estimate the mean μ of a population from a random sample X_1, X_2, \dots, X_n , simply use the sample mean \bar{X} . Is there anything else to say? How well does \bar{X} estimate the population mean? How close is \bar{X} to the population mean? The value of \bar{X} varies based on the random sample of data that was obtained. What we need to know when we use \bar{X} to estimate μ is: how does \bar{X} behave?

The answer to this question deals with a concept that is a bit hard to appreciate at first. If we have a large population with mean μ and we obtain a random sample and compute the sample mean \bar{X} , the value obtained by \bar{X} depends on the sample we obtained. Another sample would yield a different value for \bar{X} . Consider all the possible values that \bar{X} can assume for every possible random sample. When we collect a sample, we will observe one of these many values for \bar{X} . The main idea is that the value obtained by \bar{X} *varies* depending on the random sample obtained and the value is *random* because it is computed from a random sample. That is:

\bar{X} is a random variable.

This is the key conceptual point. Whenever a statistic is formed by plugging data from a random sample into a formula (e.g. \bar{X} , S^2 etc), the result is a random variable with a probability distribution. In order to know how \bar{X} performs as an estimator of μ , we need to know something about its probability distribution.

Definition. The **Sampling Distribution** of \bar{X} is the probability of distribution of \bar{X} as it varies over all possible values from all possible random samples of size n . Note: all statistics, not just \bar{X} will have sampling distributions.

Notation: Upper and Lower Case. Before the data are collected, we denote the random sample of points using upper-case: X_1, \dots, X_n , and the sample mean is denoted \bar{X} . However, once the data is collected and we have actual numerical realizations for X_1, \dots, X_n , then lower-case letters are used to denote the realized values: x_1, \dots, x_n and the observed sample mean is denoted \bar{x} .

In Chapter 4, we already determined some very important properties of the sample mean \bar{X} which we will repeat here:

1. $E[\bar{X}] = \mu$, expected value of \bar{X} is equal to the population mean. That is, \bar{X} is unbiased for μ .
2. $\text{Var}(\bar{X}) = \sigma^2/n$, where σ^2 is the population variance.
3. If the sample comes from a normal probability distribution, then the sample mean will also have a normal distribution: $\bar{X} \sim N(\mu, \sigma^2/n)$.

Note that because the variance of \bar{X} is σ^2/n , its standard deviation is σ/\sqrt{n} . The standard deviation of \bar{X} is known as the *standard error* of the mean. More generally, if θ represents a parameter of a population, we denote the estimator of θ by $\hat{\theta}$ (“theta-hat”). Because estimators are computed from a random sample, $\hat{\theta}$ will have a probability distribution and the standard deviation of this distribution is also known as the standard error of $\hat{\theta}$.

Definition. The **standard error** of an estimator $\hat{\theta}$ is its standard deviation. The standard error of the sample mean \bar{X} is

$$SE(\bar{X}) = \sigma/\sqrt{n}.$$

This formula requires σ which is unknown in practice and has to be estimated. Typically, σ is estimated using

$$s = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)}.$$

The *estimated standard error* is given by:

$$\text{Estimated Standard Error of } \bar{x}: \quad s/\sqrt{n}.$$

Notice that the standard error has \sqrt{n} in the denominator. Thus, as the sample size n gets larger, the variability of \bar{X} as an estimator of μ gets smaller. That is, \bar{X} gets more precise as n increases.

The Central Limit Theorem

Of course, we still have the problem of determining the probability distribution for \bar{X} . When obtaining a random sample, the underlying probability distribution for a variable of interest will not be known. We can plot the data and determine if it is bell-shaped or skew etc., but we will not know the probability density. This presents a problem because the exact distribution of \bar{X} depends on the underlying distribution. The most well-known theoretical result in probability and statistics is the *central limit theorem*. We may not know the exact distribution of \bar{X} , but if our sample size is large enough, it turns out that \bar{X} will behave approximately as a normal random variable:

The Central Limit Theorem. If X_1, X_2, \dots, X_n denotes a random sample from a population with mean μ and variance $\sigma^2 < \infty$, then the sampling distribution of \bar{X} is approximately normal for large n : \bar{X} is approximately $N(\mu, \sigma^2)$.

The central limit theorem is extremely useful because it tells us that the sample mean behaves like a normal random variable approximately if the sample size is large enough.

Question: How large does the sample size n have to be in order to guarantee a good normal approximation to the sampling distribution of \bar{X} ? In other words, what does “large n ” mean in the statement of the Central Limit Theorem? The answer depends on the underlying population.

- As we saw earlier, linear combinations of independent normal random variables will have a normal distribution. Thus, \bar{X} has an exact normal distribution for any sample size n if the population is normal.
- \bar{X} will be nearly normally distributed for relatively small sample sizes (e.g. $n > 5$) if the underlying distribution is close to being normal.
- If the underlying distribution deviates strongly from being normal (e.g. strongly skewed) then larger sample sizes are required for a good normal approximation.
- A general rule of thumb that is applied routinely in practice is that if $n \geq 30$, then the distribution of \bar{X} will be approximately normal. It is always a good idea to plot the data (e.g. histogram) to access if the underlying distribution deviates strongly from normality or not.
- The sampling distribution of a statistic (e.g. \bar{X}) can be accessed using the computer intensive *bootstrap* idea – see Section 3.

Example. In Chapter 4 we saw an example of data on rainfalls in Allen County, Ohio (recorded in inches). In that example, we saw that the exponential distribution appeared to give a good fit to the histogram. The exponential distribution is strongly skewed to the right. The probability density function for the exponential distribution is given by

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0.$$

In practice, the parameter θ needs to be estimated. A natural estimator to use is $\hat{\theta} = \bar{X}$, this is the *maximum likelihood estimator*, see Section 7.

For the sake of argument, suppose $\theta = 1/4$. One can show that the mean (or expected value) of the exponential distribution is θ . In our case, this means that the average rainfall is a quarter inch. Figure 1 illustrates the central limit theorem in this example. Figure 1 shows the sampling distribution of \bar{X} for various sample sizes $n = 1, 2, 10$, and 50. The underlying distribution is strongly skewed to the right as can be seen for the pdf when $n = 1$. The pdf of \bar{X} is also strongly skewed to the right for $n = 2$. However, for $n = 10$, the pdf of \bar{X} looks very similar to the familiar bell-shaped normal pdf and even more so for $n = 50$.

A couple other points to bear in mind from Figure 1:

- Each of the pdf’s for \bar{X} are centered over $\theta = 1/4$. This follows from the fact that \bar{X} is unbiased for μ .
- The spread in the distribution of \bar{X} gets smaller as n gets bigger. This is a consequence of the fact that the standard error of \bar{X} is σ/\sqrt{n} .

Suppose there is concern that the average amount of rain per rainfall is increasing. One can show that the probability of observing a single rainfall exceeding $1/2$ inch is $e^{-2} \approx 0.135$. Therefore, it is not too unusual to observe a rainfall with more than a half inch of rain (it will happen about 13-14% of the time).

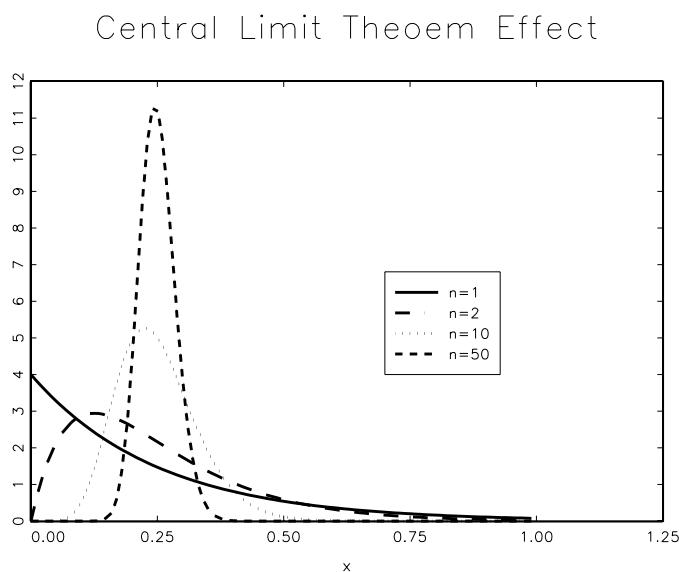


Figure 1: Illustration of the Central Limit Theorem. The underlying distribution is exponential. The pdf's are shown for the sampling distribution of \bar{X} for various sample sizes.

Suppose we observe $n = 50$ rainfalls, and the observed sample mean is $\bar{x} = 1/2$. How likely is it that the sample mean takes a value of $1/2$ or greater if the population mean is $1/4$? If the population average is $1/4$ and we observe $n = 50$ rainfalls, the average of these 50 rainfalls should take a value near $1/4$. Is a value of $\bar{x} = 1/2$ “near” $1/4$ or is it unlikely that the sample mean would take a value this far from the mean value? Note that for a single rainfall, there is a probability of 0.135 that it exceeds $1/2$. We can use the central limit theorem to answer this question. For the exponential distribution, one can show that the standard deviation is $\sigma = \theta = 0.25$. Recall, that in order to compute normal probabilities, we have to standardize first. Let us compute the (approximate) probability that \bar{X} exceeds $1/2$:

$$\begin{aligned}
 P(\bar{X} > 0.5) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{0.5 - \mu}{\sigma/\sqrt{n}}\right) \\
 &\approx P\left(Z > \frac{0.5 - 0.25}{0.25/\sqrt{50}}\right) \quad (\text{Central Limit Theorem}) \\
 &= 1 - \Phi(7.701) \\
 &\approx 0
 \end{aligned}$$

It is extraordinarily rare that a standard normal random would take a value exceeding 7.701 and hence it would be extremely unusual to observe a sample mean rainfall (based on a sample size of $n = 50$) of 0.5 or greater if the true average rainfall is $\mu = 0.25$. This is strong evidence that the population average μ does not equal $1/4$. For if $\mu = 1/4$, then it is very unlikely to observe a sample mean of $1/2$ or greater. The probability just computed $p \approx 0$ is an example of a p -value which we will define when we discuss hypothesis testing in the next chapter.

Mean or Median?

We have considered using \bar{X} as an estimator for the population mean μ up to this point. Recall that if

the distribution is symmetric, then the population median and mean are equal. In such cases we could use the sample median to estimate μ also. The sample median will be an unbiased estimator of μ if the distribution is symmetric. The question arises: which estimator should we use: the mean or median? For sample data, the mean and median will generally not be equal to each other, even when sampling from a symmetric distribution.

This question arises in many other settings as well. That is, we want to estimate a parameter θ and we may have a choice of many different estimators. How do we choose the “best” estimator? One criteria is to choose *unbiased* estimators: An estimator $\hat{\theta}$ is an unbiased estimator of a parameter θ if $E[\hat{\theta}] = \theta$. That is, the average value of our statistic $\hat{\theta}$ over all possible samples of size n will be exactly equal to θ .

If the sample mean and median are both unbiased for μ , which estimator should one use? In many examples, different estimators can give quite different values, even if they are both unbiased. The answer to the above question is: if one has a choice between several unbiased estimators, then use the one that has the smallest variance. That is, choose the estimator that is the most precise. We call such an estimator the *Uniformly Minimum Variance Unbiased* (UMVU) estimator.

If we are sampling from a normal distribution, the standard error of the sample mean is smaller than the standard error of the sample median. Suppose we have data measuring the body temperatures of healthy adult women and that the distribution of body temperatures follow a normal distribution. Then the sampling distributions of \bar{X} and the sample median would look like those shown in Figure 2. Note that the distribution for \bar{X} is more concentrated about the population mean than the distribution for the sample median. Also note that the sampling distribution of the sample median is “bell-shaped” as well. One can show that the sample median will typically have an asymptotically normal distribution (but with a larger variance than the sample mean).

There are examples of symmetric distributions where the sample median has a smaller standard error than the sample mean. These distributions will usually have “heavy tails” that produce outlying values. Recall that the sample mean can be greatly influenced by extreme observations whereas the sample median is not.

3 The Bootstrap

Many statistics are computed by doing some sort of averaging and by the central limit theorem (or variants of this theorem), the sampling distribution of these statistics will be approximately normal for large sample sizes. In many situations, the statistic of interest can become quite complicated and hence describing its sampling distribution can be very difficult. Much of the mathematical work in statistics has been directed towards determining the sampling distributions of complicated statistics. In many cases, the problem is too difficult to solve analytically. One of the great advances in statistics over the past 30 years is the introduction of the *bootstrap*. The bootstrap is a computer intensive method that can easily be applied with the wide accessibility of high speed computing. The bootstrap allows us to determine (approximately) the sampling distributions of statistics, even in cases where analytical solutions are not available.

The idea behind the bootstrap is fairly straightforward. Ideally, to determine the sampling distribution of a statistic, we would take repeated samples from the population and compute the statistic of interest for each sample. After doing this many times (say several thousand times), we would get a fairly clear picture

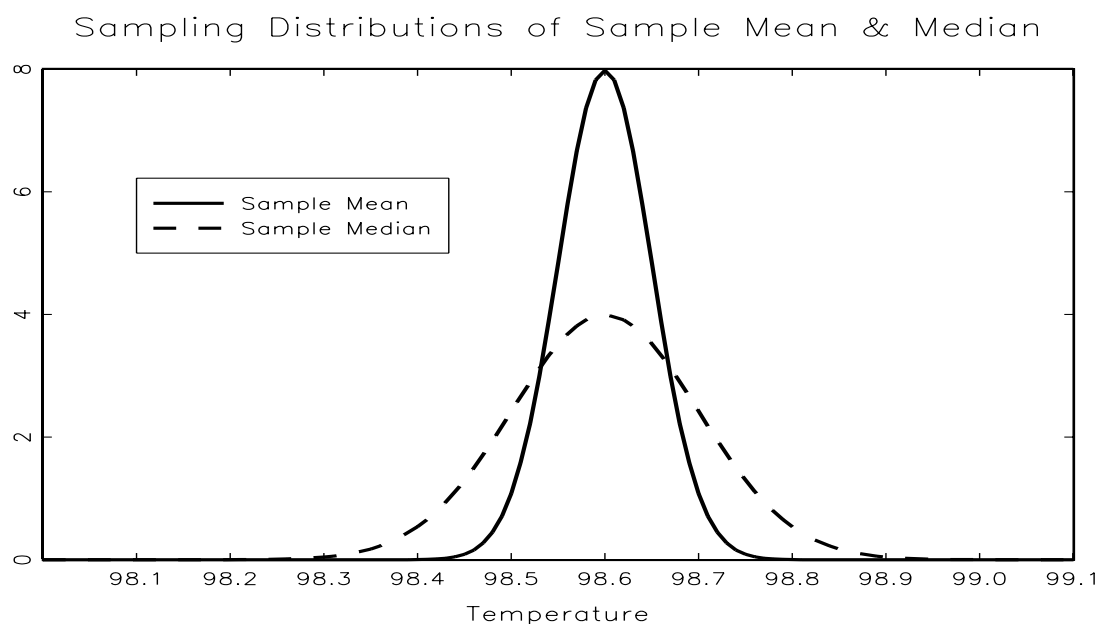


Figure 2: Sampling distributions for the sample mean and median. The distribution for the sample mean is more concentrated about the mean than the distribution for the sample median.

of how the statistic behaves, that is, we would gain insight into the statistic's sampling distribution. For instance, we could make a histogram of all the different realizations of the statistic from repeated sampling. However, in reality, we only take a single sample from the population. How then can we determine how the statistic behaves in repeated sampling? The idea of the bootstrap is to take repeated samples from your data set.

Consider a variable measured on a population of interest (e.g. body temperatures of adults) and the associated probability distribution of that variable. Once we have a sample of n observations, we can define a new discrete probability distribution by placing a probability of $1/n$ on each data point in our sample. This probability distribution is known as the *empirical distribution*. Instead of taking repeated samples from the population (since this is usually impractical), we instead take repeated samples from the empirical distribution which can be done with a computer. That is, we *re-sample* our data. For each of these re-sampled data sets, compute the statistic of interest. Have the computer do this 100 times or 1000 times or 10,000 times yielding a “distribution” of the re-sampled statistic. This distribution is known as the bootstrap distribution.

The question most people ask at this point is: “If I take a sample of size n from my data set of n observations, won't I simply reproduce my data set?” The answer is no if you *sample with replacement*. For the bootstrap, sampling is done with replacement. To sample with replacement, randomly choose one of your data points, record its value, put it back in with the other data points. Next, pick another data point at random and record its value and put it back (note – you may end up picking the same data point more than once and this is okay). Do this n times. This produces a bootstrap sample.

Here is an example of the bootstrap procedure for the sample mean. Let x_1, \dots, x_n denote our n observations in our data set.

1. Obtain sample n observations *with replacement* from the data; call them x_1^*, \dots, x_n^* .

2. Compute the mean of x_1^*, \dots, x_n^* and call it $\bar{x}_1^* = \sum_{i=1}^n x_i^*/n$.
3. Repeat steps 1 and 2 many times obtaining $\bar{x}_1^*, \bar{x}_2^*, \dots, \bar{x}_B^*$ where B is the number of bootstrap samples.

For large data sets, the total number of distinct bootstrap samples will be astronomically large. In practice, one simply obtains a large number B of randomly chosen bootstrap samples which the computer can do very easily. To determine the standard error of a statistic, a value of $B = 500$ or a 1000 may suffice. For hypothesis testing, larger bootstrap samples may be needed (say 10,000).

Although the bootstrap idea is fairly straightforward, when first introduced to it, students may find it difficult to understand why it works. We shall illustrate the bootstrap with the vole skull length data. First, we will consider the first 10 observations only for the sake of illustration. The table below gives the raw data (the first 10 observations from the data) in the first row. The second row is the first bootstrap sample, followed by the second bootstrap sample, etc. until the last bootstrap sample for $B = 500$ say.

Raw Data	2355	2305	2388	2370	2470	2535	2385	2445	2435	2330
Bootstrap 1	2355	2370	2445	2445	2305	2535	2385	2330	2330	2388
Bootstrap 2	2330	2470	2385	2435	2470	2535	2355	2535	2445	2305
⋮						⋮				
Bootstrap 500	2470	2445	2435	2370	2388	2470	2535	2388	2470	2370

1000 bootstrap samples were obtained from the full vole data and for each bootstrap sample, the mean and variance was computed. Figure 4 below shows histograms for the bootstrap distribution of the sample mean and the sample variance based on these 1000 bootstrap samples. Note that the histogram for the bootstrap distribution of the sample mean (left frame) is symmetric and mound-shaped, consistent with a normal distribution. The sample variance (right frame) has a distribution that is skewed to the right.

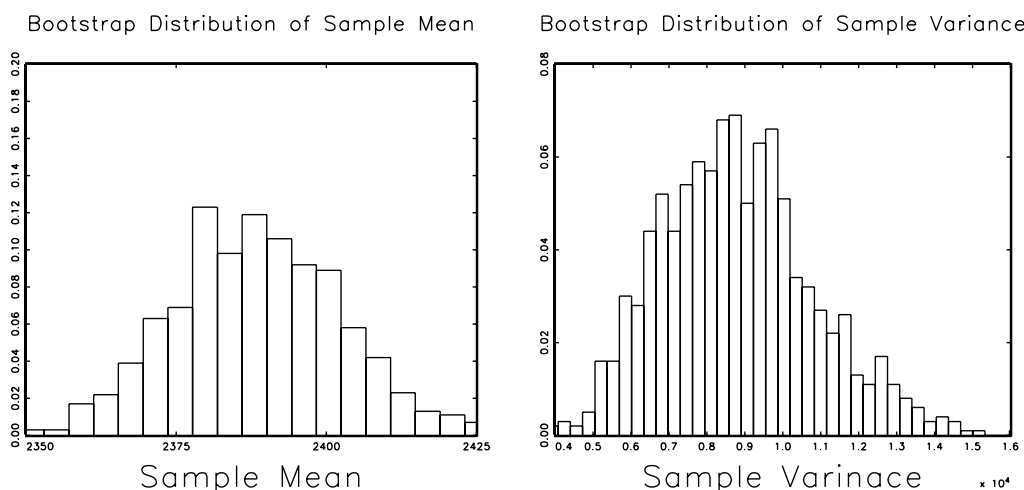


Figure 4 Bootstrap distribution for the sample mean and the sample variance for the skull length data of voles.

An excellent reference book for implementing the bootstrap in many different applications is *An Introduction to the Bootstrap* (1993) by Efron and Tibshirani. The bootstrap can be used for confidence intervals and hypothesis testing.

4 Confidence Intervals

If we use \bar{X} to estimate μ , all we have is a single *point estimator* for the parameter. A more informative approach to estimating a parameter is to provide an interval of plausible values for the parameter. The length of the interval will indicate the precision of the estimator and our level of confidence that the true parameter value will lie in the interval. These intervals are known as *confidence intervals*.

We can illustrate the idea of the confidence interval by using \bar{X} to estimate μ . Consider a random sample X_1, X_2, \dots, X_n from a normal $N(\mu, \sigma^2)$ distribution. Then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution. Recall from Chapter 4 that

$$P(-1.96 < Z < 1.96) = 0.95.$$

We can express this in terms of X as:

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$

Using some algebra, we can rewrite the inequality in this statement as:

$$P(\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}) = 0.95.$$

Thus, the *random interval* $\bar{X} \pm 1.96\sigma/\sqrt{n}$ contains μ with probability 0.95. Once we collect the data, we have a realized value for \bar{X} denoted by \bar{x} . We can then compute the following *95% confidence interval* for μ :

$$\bar{x} \pm 1.96\sigma/\sqrt{n}. \quad (1)$$

Because μ is some unknown fixed value, this interval either does or does not contain μ . In practice we will not know if μ is in the interval. However, what we do know is that if we were to repeat the experiment over and over and compute a confidence interval each time, then 95% of the confidence intervals will contain μ . In this sense, we say we are 95% confident that the interval contains μ . The confidence interval process is one that works 95% of the time.

In this illustration, we used a confidence level of 0.95. We can use other confidence levels as well. If we increase our confidence level, then we need to replace the 1.96 by a larger percentile of the standard normal distribution leading to a wider confidence interval. This makes sense - if we want to be more confident that the interval contains μ , we need to make the interval wider. If we want to be 99% confident, we need to replace the 1.96 by 2.575. If we make the interval narrower, we will have less confidence that it contains μ .

The t -Distribution. Everything we have discussed up to this point regarding confidence intervals cannot be utilized in practice because our confidence intervals, such as (1), require that we know the value of σ to plug into the formula. In real life applications, σ , like μ is an unknown parameter of the distribution. The solution to this problem is to replace σ by the sample standard deviation S . The standardized version of \bar{X} using S in place of σ is denoted by t :

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (2)$$

Replacing a fixed parameter σ by a statistic S that varies from sample to sample causes t to be more variable than Z . When we are sampling from a normal distribution, the t in (2) has a probability distribution known as the *Student's t -Distribution on $n - 1$ degrees of freedom*. In 1908, William Gossett, who was working for the Guinness Brewery in Ireland, was the first to determine the probability distribution for t . He published his work under the pseudonym "Student" and this is the name that has become attached to the distribution.

Note that the t -distribution depends on the sample size n . Because the n deviations $(X_i - \bar{X})$ always sum to zero, there are only $n - 1$ independent pieces of information in the deviations. The degrees of freedom for the t -distribution is the sample size minus one: $n - 1$.

Here are some facts about the t distribution:

1. The density functions for the t -distributions are symmetric and bell-shaped and centered at zero like the standard normal, but they have "heavier tails" than the standard normal.
2. As the degrees of freedom increase to infinity, the t -distribution becomes a standard normal distribution.

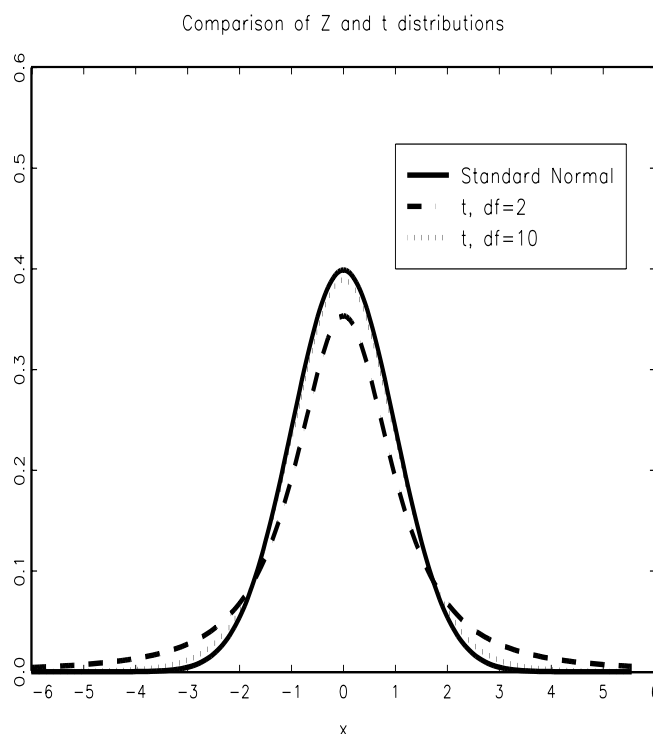


Figure 3: Density functions for the standard normal and t distributions on 2 and 10 degrees of freedom.

Figure 3 shows a plot of probability density functions for the t distributions on 2 and 10 degrees of freedom as well as the standard normal density. Note how the t -distributions have heavier tails than the standard normal. Also, the t -distribution on 10 degrees of freedom coincides more closely with the standard normal than the t -distribution on 2 degrees of freedom.

In order to compute confidence intervals using the t -distribution, we need to use percentiles of the t -distribution. These values can be computed in SAS.

Notation. The $100 \times p$ percentile of a t -distribution with ν degrees of freedom is denoted by:

$$t_{\nu,p}.$$

If X_1, \dots, X_n denotes a random sample from a normal population with mean μ , then $\nu = n - 1$ and

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,p}\right) = p.$$

t -distribution percentiles can be computed in SAS using the “`tin`” function. For example, in SAS if we run

```
t=tinv(.975,10);
```

produces a value $t_{10,0.975} = 2.22814$. The degrees of freedom in this example is 10, the second argument of the `tin` function. Note that this is larger than 1.96, the 97.5th percentile of the standard normal distribution.

In order to obtain a 95% confidence interval for μ , we want a range that covers 95% of the t probability. That will leave $1 - 0.95 = 0.05$ probability for the left and right tail. Because the distribution is symmetric, we divide this 0.05 equally to have 0.025 for each tail. Therefore, using the 0.975 percentile of the t -distribution, we can write

$$P(-t_{n-1,0.975} < \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,0.975}) = 0.95.$$

Once again, some simple algebra yields a *random* interval

$$\bar{X} \pm t_{n-1,0.975}(S/\sqrt{n})$$

that contains μ with probability 0.95. Once the data is collected, we have a fixed interval which is called a 95% confidence interval for μ :

$$\bar{x} \pm t_{n-1,0.975}(s/\sqrt{n}).$$

The interpretation of this interval is the same as before – we are 95% confident that μ lies in this interval. If we repeat the experiment over and over, 95% of the intervals will contain μ .

We can use other confidence levels besides 0.95. In general, choose a small probability value α . In the case of 95% confidence intervals, $\alpha = 0.05$.

Confidence Interval for the Mean μ of a Normal Distribution: A $100\% \times (1 - \alpha)$ confidence interval for μ is

$$\bar{x} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n}). \quad (3)$$

The assumptions required for this to be a valid confidence interval are (1) the n observations are obtained from a random sample and (2) the population is normal.

For large sample sizes, $n > 100$ say, the confidence intervals computed using the t -percentiles are almost the same as if standard normal percentiles were used.

If the underlying distribution deviates somewhat from being normal, this confidence procedure will still be approximately valid, especially if the sample size is large (due to the central limit theorem effect). If a plot of the data indicates that the distribution deviates strongly from being normal, then another confidence procedure should be used, such as the bootstrap, see Section 3.

Example. Let us estimate the mean skull length for the voles from the data presented earlier (and reproduced in the SAS program below) using 95% confidence. To get the statistics needed for the confidence interval, we can run PROC MEANS in SAS:

```

/*****
Vole skull length data from
    Airoidi, Flury and Salvioni (1996)
*****/
data voles;
input length;
datalines;
2355
2305
2388

```

```

2370
2470
2535
2385
2445
2435
2330
;
run;
proc means;
run;
data tpercent;
t=tinv(0.975,9);
proc print data=tpercent;
run;

```

The output from running this program is:

```

                    The MEANS Procedure

                    Analysis Variable : length

   N           Mean           Std Dev           Minimum           Maximum
-----
  10           2401.80         69.5713862         2305.00           2535.00
-----

                   Obs           t
                   1           2.26216

```

A 95% confidence interval with the mean skull length is

$$2401.80 \pm (2.26216)(69.57/\sqrt{10}) = 2401.80 \pm 49.767.$$

This yields an interval of

$$(2352.0, 2451.6).$$

Thus, with 95% confidence, we estimate that the average vole skull length lies between 2352 and 2451 ($mm \times 100$).

Caution on Interpretation of Confidence Intervals: Confidence intervals are very commonly misinterpreted, even by seasoned scientists. In the previous example, it is **incorrect** to say that there is a 95% probability that μ lies in the 95% confidence interval. The probability that μ lies in the interval is either zero or one because either μ lies in this fixed interval or it does not. We do not know if μ lies in the interval because we still do not know the true value of μ . The **correct** interpretation of the word

confidence is that the confidence interval procedure works 95% of the time. That is, if each person at our university were to go out and collect a random sample of voles and each student computed their own 95% confidence interval for μ , about 95% of these intervals will indeed contain the true value of μ . Note that the confidence intervals computed by all the students will differ from each other because they are each based on different random samples. This is what we mean by “95% confidence.”

If we want to be more confident, say 99% confident, then we need to make the interval wider by using the 99.5th percentile of the t -distribution on $n - 1 = 9$ degrees of freedom, which from SAS is $t_{9,0.995} = 3.24984$. If we are willing to be less confident, say 90% confidence, then the result is a narrower interval.

We can also make any confidence interval narrower by obtaining a larger sample size. Larger sample sizes will lead to smaller t -percentiles but more importantly, the plus/minus part of the interval has $1/\sqrt{n}$ factor. As n grows, this factor gets smaller leading to narrower intervals. Narrow intervals lead to more precision when estimating a parameter.

5 Sample Size Determination

When designing a study, a natural question to ask is: How many subjects? That is, how large a sample size should be collected. The more data we have, the more information we have and this leads to more precise estimates of μ (i.e. narrower confidence intervals). However, collecting data can be expensive and time consuming. When the goal is to estimate the mean of a population using a confidence interval we may specify a particular confidence level and a desired length of the interval. Suppose we want a confidence interval with a *half-width* d :

$$\bar{x} \pm d.$$

How large a sample size is required? Recalling our confidence interval formula:

$$\bar{x} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n}),$$

we can set $t_{n-1,1-\alpha/2}(s/\sqrt{n})$ equal to d and solve for the required sample size n :

$$n = \frac{t_{n-1,1-\alpha/2}^2(s^2)}{d^2}. \quad (4)$$

The problem with using this formula is that it requires that we know s . However, we are using this formula to determine the required sample size and therefore we would not have the data to estimate s . This formula also requires the t percentile which depends on the sample size as well. To get around these problems, one can plug a reasonable guess in for s either based on the knowledge of the experiment or estimates from previous experiments or surveys or pilot studies. Another alternative is to recall that the empirical rule says that about 99.7% of the observations will lie between $\pm 3s$. Thus, the range of observations is about 6 standard deviations. If we have some idea of the range of values we would obtain from a sample, simply divide it by 6 and use this as an estimate of s to get the sample size. As for the t percentile in (4), we can use an initial guess for n and compute the corresponding t percentile from SAS; plug this t percentile into (4) and see what sample size pops out. Now use this sample size to get a better guess for the t critical value. Repeat these steps a few times until it converges to a given sample size. Usually (4) will yield a non-integer value – in these cases, simply round the answer to the next largest integer.

There are numerous software programs for doing sample size computations. For example, the following web page will do sample size computations for a variety of statistical inference procedures:

<http://www.stat.uiowa.edu/~rlenth/Power/>

6 Confidence Intervals for Other Parameters

We have seen how to use a confidence interval to estimate the mean μ of a distribution. Confidence intervals are widely used for estimating other parameters as well. For example, suppose interest lies in estimating a parameter θ using a confidence interval. If $\hat{\theta}$ is a point estimator of θ and the sampling distribution of $\hat{\theta}$ is approximately normal (due to the central limit theorem effect), then an approximate $100 \times (1 - \alpha)$ confidence interval for θ will be of the form

$$\hat{\theta} \pm z_{1-\alpha/2} SE(\hat{\theta}),$$

where $SE(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$.

A Binomial Proportion p . A very common example is estimating a population proportion p from binomial (or approximate binomial) data. Recall that a binomial random variable X can be approximated by the normal distribution if the number of trials n is large enough. It follows that the sample proportion $\hat{p} = X/n$ will also have an approximate normal sampling distribution. The variance for X is npq and from our rules for linear combinations, the variance of \hat{p} is pq/n . Thus, an approximate $100 \times (1 - \alpha)$ confidence interval for p is

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}.$$

An exact confidence interval for p can be computed using the binomial distribution instead of the normal approximation.

Confidence Intervals for σ^2 .

Suppose X_1, \dots, X_n represents a random sample from a distribution with mean μ and variance σ^2 . Then one can show using a little algebra that the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (5)$$

is an unbiased estimator of the population variance σ^2 . That is

$$E[S^2] = \sigma^2.$$

Note: this explains why we divide by $n - 1$ in the formula (5) for S^2 instead of n . If we divided by n , then S^2 would be too small on average.

If we assume the random sample is from a normal distribution $N(\mu, \sigma^2)$, then the sampling distribution of

$$(n-1)S^2/\sigma^2$$

is called a *chi-squared* distribution on $n - 1$ degrees of freedom. The chi-squared distributions are skewed to the right, but as the degrees of freedom increases, the distribution becomes more and more normal (again due to the central limit theorem since S^2 is formed by averaging). The chi-square distribution plays a very important role in much of statistical inference, particularly when using maximum likelihood estimation.

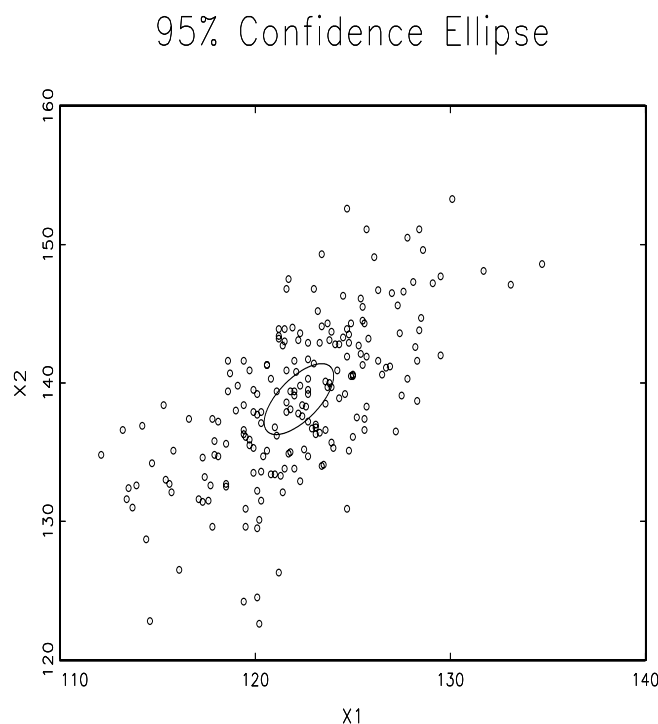


Figure 4: An example of what a 95% confidence ellipse looks like for bivariate data.

The chi-squared distribution comes about by squaring normal random variables. For instance, if Z_1, \dots, Z_k are independent $N(0, 1)$ random variables, then the sampling distribution of $Z_1^2 + \dots + Z_k^2$ is chi-square on k degrees of freedom.

If the sample is from a normal distribution, then a $(1 - \alpha) \times 100\%$ confidence interval for σ^2 is provided by

$$((n - 1)s^2 / \chi_{n-1, 1-\alpha/2}^2, (n - 1)s^2 / \chi_{n-1, \alpha/2}^2), \quad (6)$$

where $\chi_{n-1, 1-\alpha/2}^2$ and $\chi_{n-1, \alpha/2}^2$ are the $1 - \alpha/2$ and $\alpha/2$ percentiles of the chi-square distribution on $n - 1$ degrees of freedom. The confidence interval procedure based on the chi-square distribution is not very robust however. If the underlying distribution deviates from normality, then the stated confidence level may no longer be approximately correct.

Confidence Ellipses. In multivariate statistics where one is dealing with multiple correlated random variables, interest lies in estimating the means of each of the random variables. For example, if we are studying plants, we might measure X_1 , the leaf length, and X_2 , the leaf width. We could construct separate confidence intervals for the means μ_1 and μ_2 of these two variables. When considered jointly, these two confidence intervals will form a confidence rectangle formed by their Cartesian product. However, these two random variables are going to be correlated. If we take the correlation into consideration, narrower confidence regions can typically be obtained compared to the confidence rectangle. In particular, if the bivariate data comes from a bivariate normal distribution, then the confidence region will be in the shape of an ellipse in the X_1, X_2 coordinate system.

Figure 4 shows an example of what a 95% confidence ellipse looks like for a bivariate data example.

7 Maximum Likelihood Estimation

We have looked at how to estimate the mean μ and variance σ^2 using their sample counterparts. Often, data from a study will require complicated statistical models that are defined in terms of model parameters. One of the most popular methods of estimating parameters from a model is *maximum likelihood estimation*. Maximum likelihood estimation was pioneered by geneticist and statistician R. A. Fisher in the early 1900's. The concept behind maximum likelihood estimation is quite intuitive:

Given a set of data and a statistical model, what values of the model parameters makes the observed data most likely to have occurred?

For example, suppose we want to estimate the prevalence of autism in young boys. A large survey is conducted of 100,000 boys and 10,000 of the boys have autism. Let p denote the true proportion of boys with autism. Given the results of the survey, what value of p makes the observed data most likely. The answer is the maximum likelihood estimator (mle) of p , denoted \hat{p} . For instance, a value of $p = 0.90$ does not seem very likely. A much more likely value is the sample proportion $\hat{p} = 15,000/100,000 = 0.15$ which, as it turns out, is the mle of p .

For the autism example, here is how the mle is determined. For a single boy, let $x = 1$ if the boy has autism and let $x = 0$ if the boy does not have autism. Then the probability function for this boy can be written

$$p^x(1-p)^{1-x}.$$

Since we have a random sample of n boys, we will have values x_1, \dots, x_n and by independence, the probability function for the sample is

$$p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2} \dots p^{x_n}(1-p)^{1-x_n} = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

If we think of this now as a function of p , we call it the *likelihood function*. The goal then is to find the value of p that maximizes the likelihood function. The maximization problem is a simple calculus problem. The problem is simplified by taking the logarithm of the likelihood and maximizing that instead – this is typically done in maximum likelihood estimation.

The reason maximum likelihood estimation is so popular is that maximum likelihood estimators tend to perform better than other types of estimators in terms of precision. Also, the sampling distributions of mle's tend to be approximately normal due to the central limit theorem effect.

In the example above, it was fairly straightforward to compute the mle. However, for complicated models with many parameters, determining the mle's often requires multidimensional calculus differentiation and computational algorithms. The details of these will not be given here.

8 Problems

1. The average bodyfat percentage for adult men in the U.S.A. is $\mu = 19\%$ with standard deviation $\sigma = 7.8$. A university is concerned about the physical health of its students. The bodyfat percentage for a random sample of $n = 30$ male students was obtained and the average bodyfat percentage for these thirty students was found to be 21.3. If the average bodyfat percentage for all male students

at the university is equal to the national average of 19, then how likely is it that the average bodyfat percentage of 30 randomly selected students will take a value of 21.3 or greater?

2. Data on male turtles of unknown species was collected. The SAS program “maleturtle.sas” (see appendix) analyzes the data on the carapace length and width (in mm) from this sample. Use the SAS results to do the following parts.
 - a) Find a 99% confidence interval for the mean carapace length.
 - b) Find a 99% confidence interval for the mean carapace width.
 - c) Make a scatterplot of length versus width for the male turtle data and draw on your plot the rectangle formed by the Cartesian product of the two confidence intervals from (a) and (b). Suppose for the species of Painted Turtles, the mean length and width for the males are $\mu_L = 107$ and $\mu_W = 91.9$ respectively. Plot the point (91.9, 107) on your scatterplot. Note that the values $\mu_L = 107$ and $\mu_W = 91.9$ fall in their respective confidence intervals from (a) and (b). Based on your plot and your confidence intervals, does it seem reasonable that this sample of male turtles belong to the Painted Turtle species? Explain.
 - d) How would the widths of the confidence intervals you computed in parts (a) and (b) change if instead of 99% confidence intervals, we computed 95% confidence intervals instead? (Circle one)

Wider Narrower Stay the same width
 - e) The confidence interval procedure used here is based on the assumption that the data comes from an (approximate) normal distribution. Access this assumption and comment.
3. Using the data on body temperatures and pulse rates obtained from $n = 130$ health adults, which can be found in the Appendix of Chapter 2, do the following parts:
 - a) Find a 95% confidence interval for the average body temperature for healthy adults.
 - b) Re-do part (a) using 99% confidence.
 - c) How do the confidence intervals in parts (a) and (b) compare?
 - d) Does 98.6 appear to be a plausible value for the mean body temperature? Explain.
 - e) Compute a 90% confidence interval for the mean pulse rate.
 - f) (True or False) There is a 90% chance that the mean pulse rate for the population of healthy adults lies in the interval computed in part (e). Explain.
4. A study of a newly discovered species of *midge* (a small gnatlike insect) was undertaken. Based on a sample size of $n = 9$ midges, a 95% confidence interval for the average wing length μ was found to be (1.73, 1.89) (mm). Circle the correct statement below:
 - a) There is a probability of 0.95 that the true population mean lies between 1.73 and 1.89 mm.
 - b) In repeated sampling, the population mean μ will lie in the interval (1.73, 1.89) ninety five percent of the time.
 - c) In repeated sampling, approximately 95% of all confidence intervals for μ will contain the true population mean μ .
 - d) In repeated sampling, the sample mean \bar{x} will lie in 95% of the confidence intervals.

- e) \bar{x} lies in the interval (1.73, 1.89) with probability 0.95.
5. Use the results stated in Problem 4 to do this problem.
- What was the standard deviation for the midge wing lengths?
 - Using this standard deviation, estimate the required sample size needed to estimate the mean wing length using a 95% confidence interval so that the half-width is 0.1 mm.
 - Re-do part (b) using 99% confidence.
 - Re-do part (b) using a half-width of 0.05 mm.
5. The lengths of the leaves on $n = 125$ *Sagittaria lancifolia* plants from the Florida Everglades were measured. A confidence interval for the average leaf length is to be computed. Assuming the leaf lengths follow an approximate normal distribution, which confidence interval will be the *narrowest*?
- 99% CI
 - 95% CI
 - 90% CI
- d) All confidence intervals will have the same length since they are all computed with the same data set.
- e) There is not enough information to determine which interval is shortest.
6. The survival time of a fish exposed to a particular toxin follows a probability distribution that is skewed to the right. The survival times X_1, X_2, \dots, X_{100} , of a sample of $n = 100$ fish are recorded. Interest lies in estimating the average survival time μ . The sample mean \bar{X} is used to estimate μ . We know that \bar{X} will behave approximately like a normal random variable because: (circle one)
- Survival times are normally distributed.
 - Survival times have an exponential distribution.
 - \bar{X} is unbiased.
 - \bar{X} has a smaller variance.
 - Even though the fish are exposed to toxins, they will still be normal.
 - The Central Limit Theorem.

References

- Airoldi, J. P., Flury, B. and Salvioni, M., (1996), "Discrimination between two species of *Microtus* using both classified and unclassified observations. *Journal of Theoretical Biology*, **177**, p 247–262.
- Bradly Efron, Robert J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman & Hall.
- Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, **268**, 1578-1580.

9 Appendix: SAS Code and Data

Male Turtle Data

```
/******  
  Data on Male turtles (n=24)  
column 1 = Carapace length  
Column 2 = Carapace width  
*****/  
options ls=70;  
data turtles;  
input length width;  
datalines;  
  93    74  
  94    78  
  96    80  
101    84  
102    85  
103    81  
104    83  
106    83  
107    82  
112    89  
113    88  
114    86  
116    90  
117    90  
117    91  
119    93  
120    89  
120    93  
121    95  
125    93  
127    96  
128    95  
131    95  
135    106  
;  
run;  
proc ttest alpha = .01;  
run;  
proc plot;  
  plot length*width;  
  run;  
quit;
```