

10. Estimation

- A. Introduction
- B. Degrees of Freedom
- C. Characteristics of Estimators
- D. Confidence Intervals
 - 1. Introduction
 - 2. Confidence Interval for the Mean
 - 3. t distribution
 - 4. Confidence Interval for the Difference Between Means
 - 5. Confidence Interval for Pearson's Correlation
 - 6. Confidence Interval for a Proportion

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

The concept of degrees of freedom and its relationship to estimation is discussed in Section B. “Characteristics of Estimators” discusses two important concepts: bias and precision.

Point estimates are usually supplemented by interval estimates called confidence intervals. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$. The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%. The sections on confidence intervals show how to compute confidence intervals for a variety of parameters.

Introduction to Estimation

by David M. Lane

Prerequisites

- Chapter 3 Measures of Central Tendency
- Chapter 3: Variability

Learning Objectives

1. Define statistic
2. Define parameter
3. Define point estimate
4. Define interval estimate
5. Define margin of error

One of the major applications of statistics is estimating population parameters from sample statistics. For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a point estimate of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

Point estimates are usually supplemented by interval estimates called confidence intervals. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$. The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

In an experiment on memory for chess positions, the mean recall for tournament players was 63.8 and the mean for non-players was 33.1. Therefore a point estimate of the difference between population means is 30.7. The 95% confidence interval on the difference between means extends from 19.05 to 42.35. You will see how to compute this kind of interval in another section.

Degrees of Freedom

by David M. Lane

Prerequisites

- Chapter 3: Measures of Variability
- Chapter 10: Introduction to Estimation

Learning Objectives

1. Define degrees of freedom
2. Estimate the variance from a sample of 1 if the population mean is known
3. State why deviations from the sample mean are not independent
4. State the general formula for degrees of freedom in terms of the number of values and the number of estimated parameters
5. Calculate s^2

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean $(8-6)^2 = 4$ is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df. If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance, $(5-6)^2 = 1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on two independently and randomly selected Martians. The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

As you are probably thinking, it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean (μ) with the sample mean (M). The process of estimating the mean affects our degrees of freedom as shown below.

Returning to our problem of estimating the variance in Martian heights, let's assume we do not know the population mean and therefore we have to estimate it from the sample. We have sampled two Martians and found that their heights are 8 and 5. Therefore M , our estimate of the population mean, is

$$M = (8+5) / 2 = 6.5.$$

We can now compute two estimates of variance:

$$\text{Estimate 1} = (8-6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5-6.5)^2 = 2.25$$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of M . Since the first Martian's height of 8 influenced M , it also influenced Estimate 2. If the first height had been, for example, 10, then M would have been 7.5 and Estimate 2 would have been $(5-7.5)^2 = 6.25$ instead of 2.25. The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be $13-5 = 8$.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest (σ^2). Therefore, the estimate of variance has $2 - 1 = 1$ degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to $N - 1$ where N is the number of observations.

Recall from the section on variability that the formula for estimating the variance in a sample is:

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

The denominator of this formula is the degrees of freedom.

Characteristics of Estimators

by David M. Lane

Prerequisites

- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability
- Chapter 9: Introduction to Sampling Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 10: Introduction to Estimation
- Chapter 10: Degrees of Freedom

Learning Objectives

1. Define bias
2. Define sampling variability
3. Define expected value
4. Define relative efficiency

This section discusses two important characteristics of statistics used as point estimates of parameters: bias and sampling variability. Bias refers to whether an estimator tends to either over or underestimate the parameter. Sampling variability refers to how much the estimate varies from sample to sample.

Have you ever noticed that some bathroom scales give you very different weights each time you weigh yourself? With this in mind, let's compare two scales. Scale 1 is a very high-tech digital scale and gives essentially the same weight each time you weigh yourself; it varies by at most 0.02 pounds from weighing to weighing. Although this scale has the potential to be very accurate, it is calibrated incorrectly and, on average, overstates your weight by one pound. Scale 2 is a cheap scale and gives very different results from weighing to weighing. However, it is just as likely to underestimate as overestimate your weight. Sometimes it vastly overestimates it and sometimes it vastly underestimates it. However, the average of a large number of measurements would be your actual weight. Scale 1 is biased since, on average, its measurements are one pound higher than your actual weight. Scale 2, by contrast, gives unbiased estimates of your weight. However, Scale 2 is highly variable and its measurements are often very far from

your true weight. Scale 1, in spite of being biased, is fairly accurate. Its measurements are never more than 1.02 pounds from your actual weight.

We now turn to more formal definitions of variability and precision. However, the basic ideas are the same as in the bathroom scale example.

Bias

A statistic is biased if the long-term average value of the statistic is not the parameter it is estimating. More formally, a statistic is biased if the mean of the sampling distribution of the statistic is not equal to the parameter. The mean of the sampling distribution of a statistic is sometimes referred to as the expected value of the statistic.

As we saw in the section on the sampling distribution of the mean, the mean of the sampling distribution of the (sample) mean is the population mean (μ). Therefore the sample mean is an unbiased estimate of μ . Any given sample mean may underestimate or overestimate μ , but there is no systematic tendency for sample means to either under or overestimate μ .

In the section on variability, we saw that the formula for the variance in a population is

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

whereas the formula to estimate the variance from a sample is

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

Notice that the denominators of the formulas are different: N for the population and N-1 for the sample. If N is used in the formula for s^2 , then the estimates tend to be too low and therefore biased. The formula with N-1 in the denominator gives an unbiased estimate of the population variance. Note that N-1 is the degrees of freedom.

Sampling Variability

The sampling variability of a statistic refers to how much the statistic varies from sample to sample and is usually measured by its standard error ; the smaller the standard error, the less the sampling variability. For example, the standard error of

the mean is a measure of the sampling variability of the mean. Recall that the formula for the standard error of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

The larger the sample size (N), the smaller the standard error of the mean and therefore the lower the sampling variability.

Statistics differ in their sampling variability even with the same sample size. For example, for normal distributions, the standard error of the median is larger than the standard error of the mean. The smaller the standard error of a statistic, the more efficient the statistic. The relative efficiency of two statistics is typically defined as the ratio of their standard errors. However, it is sometimes defined as the ratio of their squared standard errors.

Confidence Intervals

by David M. Lane

A. Introduction

B. Confidence Interval for the Mean

C. t distribution

D. Confidence Interval for the Difference Between Means

E. Confidence Interval for Pearson's Correlation

F. Confidence Interval for a Proportion

These sections show how to compute confidence intervals for a variety of parameters.

Introduction to Confidence Intervals

by David M. Lane

Prerequisites

- Chapter 5: Introduction to Probability
- Chapter 10: Introduction to Estimation
- Chapter 10: Characteristics of Estimators

Learning Objectives

1. Define confidence interval
2. State why a confidence interval is not the probability the interval contains the parameter

Say you were interested in the mean weight of 10-year-old girls living in the United States. Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds. This sample mean of 90 is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean. For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure (presented in the next section) that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.

It is natural to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple. One problem is that the computation of a confidence interval does not take into account any other information you might have about the value of the population mean. For example, if numerous prior studies had all found sample means above 110, it would not make sense to conclude that there is a 0.95 probability that the population mean is between 72.85 and 107.15. What about situations in which there is no prior information about the value of the population mean? Even here the interpretation is complex. The problem is that there can be more than one procedure that produces intervals that contain the population parameter 95% of the time. Which procedure produces the “true” 95% confidence interval? Although the various methods are equal from a purely mathematical point of view, the standard method of computing confidence intervals has two desirable properties: each interval is symmetric about the point estimate and each interval is contiguous. Recall from the introductory section in the chapter on probability that, for some purposes, probability is best thought of as subjective. It is reasonable, although not required by the laws of probability, that one adopt a subjective probability of 0.95 that a 95% confidence interval, as typically computed, contains the parameter in question.

Confidence intervals can be computed for various parameters, not just the mean. For example, later in this chapter you will see how to compute a confidence interval for ρ , the population value of Pearson's r , based on sample data.

t Distribution

by David M. Lane

Prerequisites

- Chapter 7: Normal Distribution,
- Chapter 7: Areas Under Normal Distributions
- Chapter 10: Degrees of Freedom

Learning Objectives

1. State the difference between the shape of the t distribution and the normal distribution
2. State how the difference between the shape of the t distribution and normal distribution is affected by the degrees of freedom
3. Use a t table to find the value of t to use in a confidence interval
4. Use the t calculator to find the value of t to use in a confidence interval

In the introduction to normal distributions it was shown that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean. Therefore, if you randomly sampled a value from a normal distribution with a mean of 100, the probability it would be within 1.96σ of 100 is 0.95. Similarly, if you sample N values from the population, the probability that the sample mean (M) will be within $1.96 \sigma_M$ of 100 is 0.95.

Now consider the case in which you have a normal distribution but you do not know the standard deviation. You sample N values and compute the sample mean (M) and estimate the standard error of the mean (σ_M) with s_M . What is the probability that M will be within $1.96 s_M$ of the population mean (μ)? This is a difficult problem because there are two ways in which M could be more than $1.96 s_M$ from μ : (1) M could, by chance, be either very high or very low and (2) s_M could, by chance, be very low. Intuitively, it makes sense that the probability of being within 1.96 standard errors of the mean should be smaller than in the case when the standard deviation is known (and cannot be underestimated). But exactly how much smaller? Fortunately, the way to work out this type of problem was solved in the early 20th century by W. S. Gosset who determined the distribution of a mean divided by its estimate of the standard error. This distribution is called the Student's t distribution or sometimes just the t distribution. Gosset worked out the t

distribution and associated statistical tests while working for a brewery in Ireland. Because of a contractual agreement with the brewery, he published the article under the pseudonym “Student.” That is why the t test is called the “Student's t test.”

The t distribution is very similar to the normal distribution when the estimate of variance is based on many degrees of freedom, but has relatively more scores in its tails when there are fewer degrees of freedom. Figure 1 shows t distributions with 2, 4, and 10 degrees of freedom and the standard normal distribution. Notice that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t distribution is therefore leptokurtic. The t distribution approaches the normal distribution as the degrees of freedom increase.

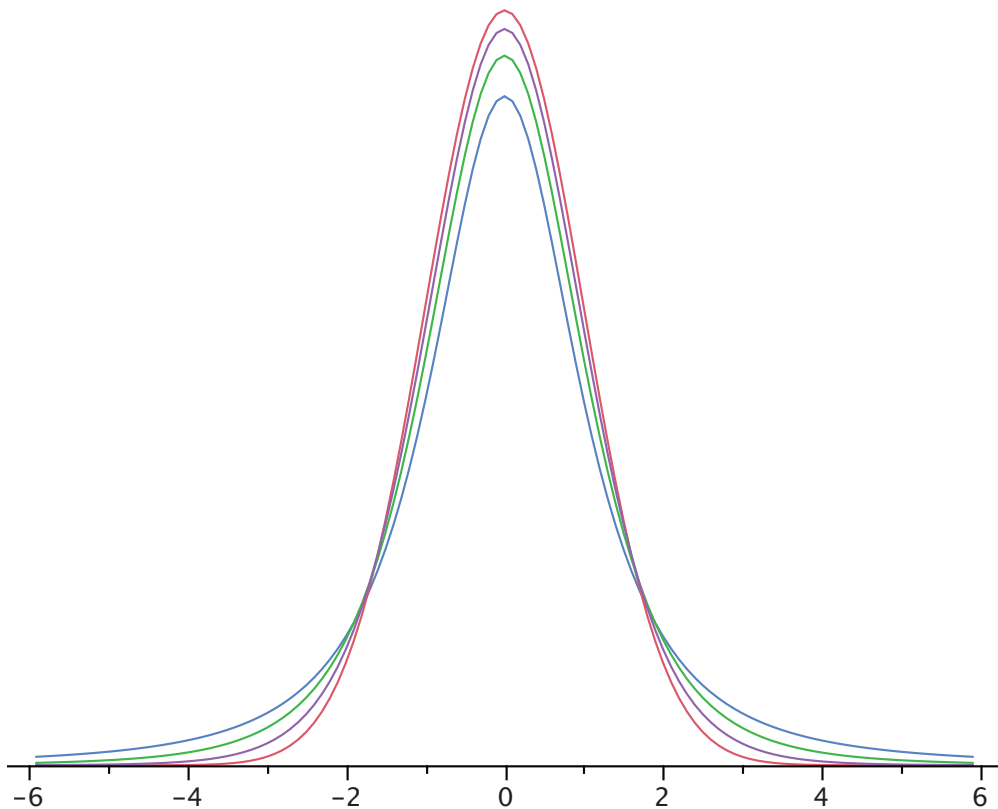


Figure 1. A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the highest peak is the 2 df distribution, the next highest is 4 df, the highest after that is 10 df, and the lowest is the standard normal distribution.

Since the t distribution is leptokurtic, the percentage of the distribution within 1.96 standard deviations of the mean is less than the 95% for the normal distribution. Table 1 shows the number of standard deviations from the mean required to contain 95% and 99% of the area of the t distribution for various degrees of freedom. These are the values of t that you use in a confidence interval. The corresponding values for the normal distribution are 1.96 and 2.58 respectively. Notice that with few degrees of freedom, the values of t are much higher than the corresponding values for a normal distribution and that the difference decreases as the degrees of freedom increase. The values in Table 1 can be obtained from the “Find t for a confidence interval” calculator.

Table 1. Abbreviated t table.

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Returning to the problem posed at the beginning of this section, suppose you sampled 9 values from a normal population and estimated the standard error of the mean (σ_M) with s_M . What is the probability that M would be within $1.96s_M$ of μ ? Since the sample size is 9, there are $N - 1 = 8$ df. From Table 1 you can see that with 8 df the probability is 0.95 that the mean will be within $2.306 s_M$ of μ . The probability that it will be within $1.96 s_M$ of μ is therefore lower than 0.95.

A “t distribution” calculator can be used to find that 0.086 of the area of a t distribution is more than 1.96 standard deviations from the mean, so the probability that M would be less than $1.96s_M$ from μ is $1 - 0.086 = 0.914$.

As expected, this probability is less than 0.95 that would have been obtained if σ_M had been known instead of estimated.

Confidence Interval for the Mean

by David M. Lane

Prerequisites

- Chapter 7: Areas Under Normal Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 10: Introduction to Estimation
- Chapter 10: Introduction to Confidence Intervals
- Chapter 10: t distribution

Learning Objectives

1. Use the inverse normal distribution calculator to find the value of z to use for a confidence interval
2. Compute a confidence interval on the mean when σ is known
3. Determine whether to use a t distribution or a normal distribution
4. Compute a confidence interval on the mean when σ is estimated

When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean, there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from the section on the sampling distribution of the mean that the mean of the sampling distribution is μ and the standard error of the mean is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of $36/3 = 12$. Note that the standard deviation of a sampling distribution is its standard error. Figure 1 shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48

to 113.52. These limits were computed by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

$$\begin{aligned}90 - (1.96)(12) &= 66.48 \\90 + (1.96)(12) &= 113.52\end{aligned}$$

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean; 12 is the standard error of the mean.

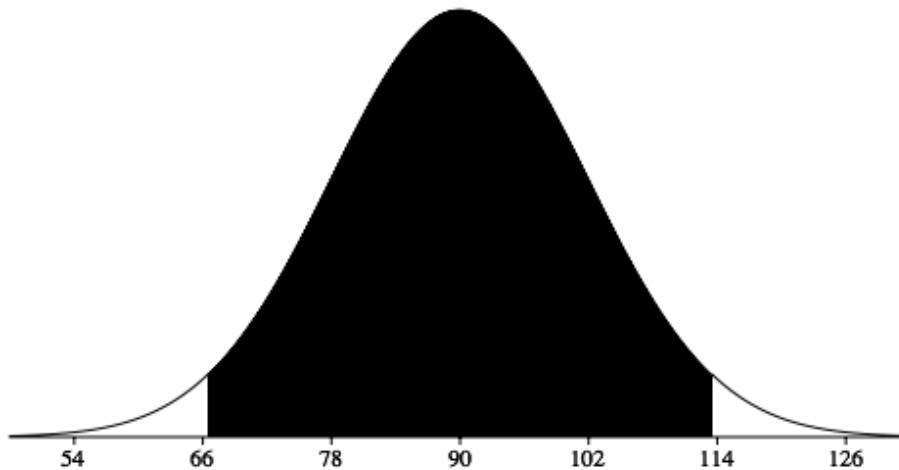


Figure 1. The sampling distribution of the mean for $N=9$. The middle 95% of the distribution is shaded.

Figure 1 shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the mean (M) from a sample, and create an interval ranging from $M - 23.52$ to $M + 23.52$, this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

$$\begin{aligned}\text{Lower limit} &= M - Z_{.95}\sigma_m \\ \text{Upper limit} &= M + Z_{.95}\sigma_m\end{aligned}$$

where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area and σ_M is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the standard deviation (σ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when σ is known is easier than when σ has to be estimated, and serves a pedagogical purpose. Later in this section we will show how to compute a confidence interval for the mean when σ has to be estimated.

Suppose the following five numbers were sampled from a normal distribution with a standard deviation of 2.5: 2, 3, 5, 6, and 9. To compute the 95% confidence interval, start by computing the mean and standard error:

$$M = (2 + 3 + 5 + 6 + 9) / 5 = 5.$$

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{2.5}{\sqrt{5}} = 1.118$$

$Z_{.95}$ can be found using the normal distribution calculator and specifying that the area is 0.95 and indicating that you want the area to be between the cutoff points. The value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been 2.58.

The confidence interval can then be computed as follows:

$$\begin{aligned} \text{Lower limit} &= 5 - (1.96)(1.118) = 2.81 \\ \text{Upper limit} &= 5 + (1.96)(1.118) = 7.19 \end{aligned}$$

You should use the t distribution rather than the normal distribution when the variance is not known and has to be estimated from sample data. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution is leptokurtic, which means it has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78

standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table 1. The first column, df, stands for degrees of freedom, and for confidence intervals on the mean, df is equal to N - 1, where N is the sample size.

Table 1. Abbreviated t table.

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the “inverse t distribution” calculator to find the t values to use in confidence intervals.

Assume that the following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9 and that the standard deviation is not known. The first steps are to compute the sample mean and variance:

$$M = 5$$

$$s^2 = 7.5$$

The next step is to estimate the standard error of the mean. If we knew the population variance, we could use the following formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Instead we compute an estimate of the standard error (s_M):

$$s_m = \frac{S}{\sqrt{N}} = 1.225$$

The next step is to find the value of t . As you can see from Table 1, the value for the 95% confidence interval for $df = N - 1 = 4$ is 2.776. The confidence interval is then computed just as it is with σ_M . The only differences are that s_M and t rather than σ_M and Z are used.

$$\begin{aligned}\text{Lower limit} &= 5 - (2.776)(1.225) = 1.60 \\ \text{Upper limit} &= 5 + (2.776)(1.225) = 8.40\end{aligned}$$

More generally, the formula for the 95% confidence interval on the mean is:

$$\begin{aligned}\text{Lower limit} &= M - (t_{CL})(s_M) \\ \text{Upper limit} &= M + (t_{CL})(s_M)\end{aligned}$$

where M is the sample mean, t_{CL} is the t for the confidence level desired (0.95 in the above example), and s_M is the estimated standard error of the mean.

We will finish with an analysis of the Stroop Data. Specifically, we will compute a confidence interval on the mean difference score. Recall that 47 subjects named the color of ink that words were written in. The names conflicted so that, for example, they would name the ink color of the word “blue” written in red ink. The correct response is to say “red” and ignore the fact that the word is “blue.” In a second condition, subjects named the ink color of colored rectangles.

Table 2. Response times in seconds for 10 subjects.

Naming Colored Rectangle	Interference	Difference
17	38	21
15	58	43
18	35	17
20	39	19
18	33	15
20	32	12
20	45	25
19	52	33
17	31	14
21	29	8

Table 2 shows the time difference between the interference and color-naming conditions for 10 of the 47 subjects. The mean time difference for all 47 subjects is 16.362 seconds and the standard deviation is 7.470 seconds. The standard error of the mean is 1.090. A t table shows the critical value of t for $47 - 1 = 46$ degrees of freedom is 2.013 (for a 95% confidence interval). Therefore the confidence interval is computed as follows:

$$\text{Lower limit} = 16.362 - (2.013)(1.090) = 14.17$$

$$\text{Upper limit} = 16.362 + (2.013)(1.090) = 18.56$$

Therefore, the interference effect (difference) for the whole population is likely to be between 14.17 and 18.56 seconds.

Difference between Means

by David M. Lane

Prerequisites

- Chapter 9: Sampling Distribution of Difference between Means
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Mean

Learning Objectives

1. State the assumptions for computing a confidence interval on the difference between means
2. Compute a confidence interval on the difference between means
3. Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. We take as an example the data from the “Animal Research” case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 1.

Table 1. Means and Variances in Animal Research study.

Condition	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is the difference in the population. The difference in sample means is used to estimate the difference in population means. The accuracy of the estimate is revealed by a confidence interval.

In order to construct a confidence interval, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.

3. Each value is sampled independently from each other value.

The consequences of violating these assumptions are discussed in Chapter 12. For now, suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference.

A confidence interval on the difference between means is computed using the following formula:

$$\text{Lower Limit} = M_1 - M_2 - (t_{CL})(S_{M_1-M_2})$$

$$\text{Upper Limit} = M_1 - M_2 + (t_{CL})(S_{M_1-M_2})$$

where $M_1 - M_2$ is the difference between sample means, t_{CL} is the t for the desired level of confidence, and $(S_{M_1-M_2})$ is the estimated standard error of the difference between sample means. The meanings of these terms will be made clearer as the calculations are demonstrated.

We continue to use the data from the “Animal Research” case study and will compute a confidence interval on the difference between the mean score of the females and the mean score of the males. For this calculation, we will assume that the variances in each of the two populations are equal.

The first step is to compute the estimate of the standard error of the difference between means $(S_{M_1-M_2})$. Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference in means in the population is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2 . Since we are assuming the population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2 . In this example,

$$MSE = (2.743 + 2.985)/2 = 2.864.$$

Note that MSE stands for “mean square error” and is the mean squared deviation of each score from its group’s mean.

Since n (the number of scores in each condition) is 17,

$$s_{M_1-M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805.$$

The next step is to find the t to use for the confidence interval (t_{CL}). To calculate t_{CL} , we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to $(n_1 - 1) + (n_2 - 1)$ where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use “ n ” to refer to the sample size of each group. Therefore, the degrees of freedom is $16 + 16 = 32$.

From either the above calculator or a t table, you can find that the t for a 95% confidence interval for 32 df is 2.037.

We now have all the components needed to compute the confidence interval. First, we know the difference between means:

$$M_1 - M_2 = 5.353 - 3.882 = 1.471$$

We know the standard error of the difference between means is

$$s_{M_1-M_2} = 0.5805$$

and that the t for the 95% confidence interval with 32 df is

$$t_{CL} = 2.037$$

Therefore, the 95% confidence interval is

$$\text{Lower Limit} = 1.471 - (2.037)(0.5805) = 0.29$$

$$\text{Upper Limit} = 1.471 + (2.037)(0.5805) = 2.65$$

We can write the confidence interval as:

$$0.29 \leq \mu_f - \mu_m \leq 2.65$$

where μ_f is the population mean for females and μ_m is the population mean for males. This analysis provides evidence that the mean for females is higher than the mean for males, and that the difference between means in the population is likely to be between 0.29 and 2.65.

Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 2.

Table 2. Example Data

Group 1	Group 2
3	5
4	6
5	7

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. For the data in Table 2, the reformatted data look as follows:

Table 3. Reformatted Data

G	Y
1	3
1	4
1	5
2	5
2	6
2	7

Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the sample with the larger sample size more than the sample with the smaller sample size. Computationally this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2$$

where M_1 is the mean for group 1 and M_2 is the mean for group 2. Consider the following small example:

Table 4. Example Data

Group 1	Group 2
3	2
4	4
5	

$$M_1 = 4 \text{ and } M_2 = 3.$$

$$SSE = \frac{(3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2}{4} =$$

Then, MSE is computed by:

$$MSE = SSE/df$$

where the degrees of freedom (df) is computed as before:

$$df = (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3.$$
$$MSE = SSE/df = 4/3 = 1.333.$$

The formula

$$s_{M_1-M_2} = \sqrt{\frac{2MSE}{n}}$$

is replaced by

$$s_{M_1-M_2} = \sqrt{\frac{2MSE}{n_h}}$$

where n_h is the harmonic mean of the sample sizes and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4$$

and

$$s_{M_1-M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054.$$

t_{CL} for 3 df and the 0.05 level = 3.182.

Therefore the 95% confidence interval is

$$\text{Lower Limit} = 1 - (3.182)(1.054) = -2.35$$

$$\text{Upper Limit} = 1 + (3.182)(1.054) = 4.35$$

We can write the confidence interval as:

$$-2.35 \leq \mu_1 - \mu_2 \leq 4.35$$

Correlation

by David M. Lane

Prerequisites

- Chapter 4: Values of the Pearson Correlation
- Chapter 9: Sampling Distribution of Pearson's r
- Chapter 10: Confidence Intervals

Learning Objectives

1. State why the z' transformation is necessary
2. Compute the standard error of z'
3. Compute a confidence interval on ρ

The computation of a confidence interval on the population value of Pearson's correlation (ρ) is complicated by the fact that the sampling distribution of r is not normally distributed. The solution lies with Fisher's z' transformation described in the section on the sampling distribution of Pearson's r. The steps in computing a confidence interval for ρ are:

1. Convert r to z'
2. Compute a confidence interval in terms of z'
3. Convert the confidence interval back to r.

Let's take the data from the case study Animal Research as an example. In this study, students were asked to rate the degree to which they thought animal research is wrong and the degree to which they thought it is necessary. As you might have expected, there was a negative relationship between these two variables: the more that students thought animal research is wrong, the less they thought it is necessary. The correlation based on 34 observations is -0.654. The problem is to compute a 95% confidence interval on ρ based on this r of -0.654.

The conversion of r to z' can be done using a [calculator](#). This calculator shows that the z' associated with an r of -0.654 is -0.78.

The sampling distribution of z' is approximately normally distributed and has a standard error of

$$\frac{1}{\sqrt{N-3}}$$

For this example, $N = 34$ and therefore the standard error is 0.180. The Z for a 95% confidence interval ($Z_{.95}$) is 1.96, as can be found using the normal distribution calculator (setting the shaded area to .95 and clicking on the “Between” button). The confidence interval is therefore computed as:

$$\begin{aligned}\text{Lower limit} &= -0.78 - (1.96)(0.18) = -1.13 \\ \text{Upper limit} &= -0.78 + (1.96)(0.18) = -0.43\end{aligned}$$

The final step is to convert the endpoints of the interval back to r using a table or the calculator. The r associated with a z' of -1.13 is -0.81 and the r associated with a z' of -0.43 is -0.40. Therefore, the population correlation (ρ) is likely to be between -0.81 and -0.40. The 95% confidence interval is:

$$-0.81 \leq \rho \leq -0.40$$

To calculate the 99% confidence interval, you use the Z for a 99% confidence interval of 2.58 as follows:

$$\begin{aligned}\text{Lower limit} &= -0.775 - (2.58)(0.18) = -1.24 \\ \text{Upper limit} &= -0.775 + (2.58)(0.18) = -0.32\end{aligned}$$

Converting back to r , the confidence interval is:

$$-0.84 \leq \rho \leq -0.31$$

Naturally, the 99% confidence interval is wider than the 95% confidence interval.

Proportion

by David M. Lane

Prerequisites

- Chapter 7: Introduction to the Normal Distribution
- Chapter 7: Normal Approximation to the Binomial
- Chapter 9: Sampling Distribution of the Mean
- Chapter 9: Sampling Distribution of a Proportion
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Mean

Learning Objectives

1. Estimate the population proportion from sample proportions
2. Apply the correction for continuity
3. Compute a confidence interval

A candidate in a two-person election commissions a poll to determine who is ahead. The pollster randomly chooses 500 registered voters and determines that 260 out of the 500 favor the candidate. In other words, 0.52 of the sample favors the candidate. Although this point estimate of the proportion is informative, it is important to also compute a confidence interval. The confidence interval is computed based on the mean and standard deviation of the sampling distribution of a proportion. The formulas for these two parameters are shown below:

$$\mu_p = \pi$$

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

Since we do not know the population parameter π , we use the sample proportion p as an estimate. The estimated standard error of p is therefore

$$s_p = \sqrt{\frac{p(1 - p)}{N}}$$

We start by taking our statistic (p) and creating an interval that ranges $(Z_{.95})(s_p)$ in both directions where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area. (See the section on the confidence interval for the mean). The value of $Z_{.95}$ is computed with the normal calculator and is equal to 1.96. We then make a slight adjustment to correct for the fact that the distribution is discrete rather than continuous.

s_p is calculated as shown below:

$$s_p = \sqrt{\frac{.52(1 - .52)}{500}} = 0.0223$$

To correct for the fact that we are approximating a discrete distribution with a continuous distribution (the normal distribution), we subtract $0.5/N$ from the lower limit and add $0.5/N$ to the upper limit of the interval. Therefore the confidence interval is

$$p \pm Z_{.95} \sqrt{\frac{p(1 - p)}{N}} \pm \frac{0.5}{N}$$

$$\text{Lower: } 0.52 - (1.96)(0.0223) - 0.001 = 0.475$$

$$\text{Upper: } 0.52 + (1.96)(0.0223) + 0.001 = 0.565$$

$$.475 \leq \pi \leq .565$$

Since the interval extends 0.045 in both directions, the margin of error is 0.045. In terms of percent, between 47.5% and 56.5% of the voters favor the candidate and the margin of error is 4.5%. Keep in mind that the margin of error of 4.5% is the margin of error for the percent favoring the candidate and not the margin of error for the difference between the percent favoring the candidate and the percent favoring the opponent. The margin of error for the difference is 9%, twice the margin of error for the individual percent. Keep this in mind when you hear reports in the media; the media often get this wrong.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 10: Proportions

In July of 2011, Gene Munster of Piper Jaffray reported the results of a survey in a note to clients. This research was reported throughout the media. Perhaps the fullest description was presented on the CNNMoney website (A service of CNN, Fortune, and Money) in an article entitled "Survey: iPhone retention 94% vs. Android 47%." The data were collected by asking people in food courts and baseball stadiums what their current phone was and what phone they planned to buy next. The data were collected in the summer of 2011. Below is a portion of the data:

Phone	Keep	Change	Proportion
iPhone	58	4	0.94
Android	17	19	0.47

What do you think?

The article contains the strong caution: "It's only a tiny sample, so large conclusions must not be drawn." This caution appears to be a welcome change from the overstating of findings typically found in the media. But has this report understated the importance of the study? Perhaps it is valid to draw some "large conclusions."?

The confidence interval on the proportion extends from 0.87 to 1.0 (some methods give the interval from 0.85 to 0.97). Even the lower bound indicates the vast majority of iPhone owners plan to buy another iPhone. A strong conclusion can be made even with this sample size.

Exercises

Prerequisites

- All material presented in the Estimation Chapter

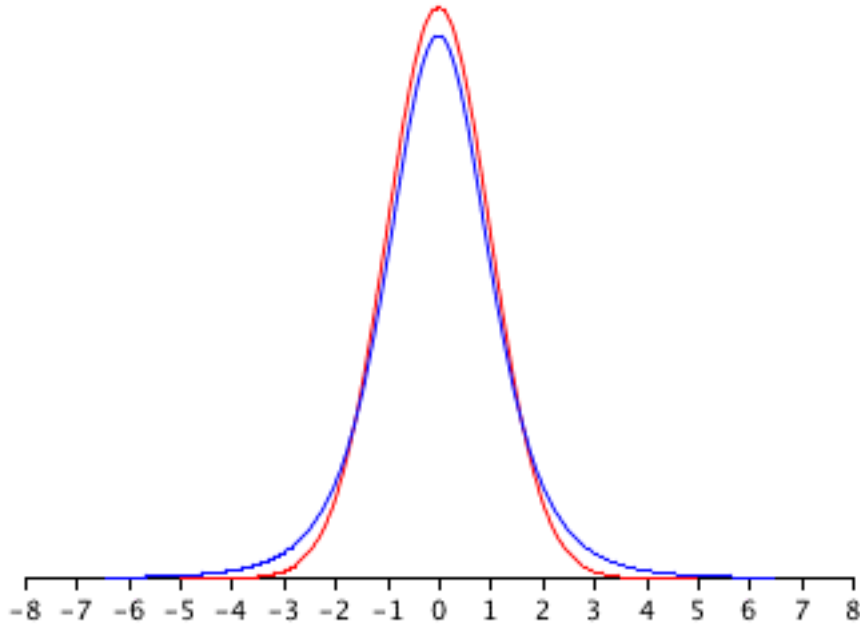
1. When would the mean grade in a class on a final exam be considered a statistic? When would it be considered a parameter?
2. Define bias in terms of expected value.
3. Is it possible for a statistic to be unbiased yet very imprecise? How about being very accurate but biased?
4. Why is a 99% confidence interval wider than a 95% confidence interval?
5. When you construct a 95% confidence interval, what are you 95% confident about?
6. What is the difference in the computation of a confidence interval between cases in which you know the population standard deviation and cases in which you have to estimate it?
7. Assume a researcher found that the correlation between a test he or she developed and job performance was 0.55 in a study of 28 employees. If correlations under .35 are considered unacceptable, would you have any reservations about using this test to screen job applicants?
8. What is the effect of sample size on the width of a confidence interval?
9. How does the t distribution compare with the normal distribution? How does this difference affect the size of confidence intervals constructed using z relative to those constructed using t? Does sample size make a difference?
10. The effectiveness of a blood-pressure drug is being investigated. How might an experimenter demonstrate that, on average, the reduction in systolic blood pressure is 20 or more?

11. A population is known to be normally distributed with a standard deviation of 2.8. (a) Compute the 95% confidence interval on the mean based on the following sample of nine: 8, 9, 10, 13, 14, 16, 17, 20, 21. (b) Now compute the 99% confidence interval using the same data.
12. A person claims to be able to predict the outcome of flipping a coin. This person is correct 16/25 times. Compute the 95% confidence interval on the proportion of times this person can predict coin flips correctly. What conclusion can you draw about this test of his ability to predict the future?
13. What does it mean that the variance (computed by dividing by N) is a biased statistic?
14. A confidence interval for the population mean computed from an N of 16 ranges from 12 to 28. A new sample of 36 observations is going to be taken. You can't know in advance exactly what the confidence interval will be because it depends on the random sample. Even so, you should have some idea of what it will be. Give your best estimation.
15. You take a sample of 22 from a population of test scores, and the mean of your sample is 60. (a) You know the standard deviation of the population is 10. What is the 99% confidence interval on the population mean. (b) Now assume that you do not know the population standard deviation, but the standard deviation in your sample is 10. What is the 99% confidence interval on the mean now?
16. You read about a survey in a newspaper and find that 70% of the 250 people sampled prefer Candidate A. You are surprised by this survey because you thought that more like 50% of the population preferred this candidate. Based on this sample, is 50% a possible population proportion? Compute the 95% confidence interval to be sure.
17. Heights for teenage boys and girls were calculated. The mean height for the sample of 12 boys was 174 cm and the variance was 62. For the sample of 12 girls, the mean was 166 cm and the variance was 65. Assuming equal variances and normal distributions in the population, (a) What is the 95% confidence interval on the difference between population means? (b) What is the 99%

confidence interval on the difference between population means? (c) Do you think it is very unlikely that the mean difference in the population is about 5? Why or why not?

18. You were interested in how long the average psychology major at your college studies per night, so you asked 10 psychology majors to tell you the amount they study. They told you the following times: 2, 1.5, 3, 2, 3.5, 1, 0.5, 3, 2, 4. (a) Find the 95% confidence interval on the population mean. (b) Find the 90% confidence interval on the population mean.
19. True/false: As the sample size gets larger, the probability that the confidence interval will contain the population mean gets higher.
20. True/false: You have a sample of 9 men and a sample of 8 women. The degrees of freedom for the t value in your confidence interval on the difference between means is 16.
21. True/false: Greek letters are used for statistics as opposed to parameters.
22. True/false: In order to construct a confidence interval on the difference between means, you need to assume that the populations have the same variance and are both normally distributed.

23. True/false: The red distribution represents the t distribution and the blue distribution represents the normal distribution.



Questions from Case Studies

Angry Moods (AM) case study

24. (AM) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the “Anger-Out” scores, compute a 99% confidence interval on the difference between gender means.
25. (AM) Calculate the 95% confidence interval for the difference between the mean Anger-In score for the athletes and non-athletes. What can you conclude?
26. (AM) Find the 95% confidence interval on the population correlation between the Anger- Out and Control-Out scores.

Flatulence (F) case study

27. (F) Compare men and women on the variable “perday.” Compute the 95% confidence interval on the difference between means.

28. (F) What is the 95% confidence interval of the mean time people wait before farting in front of a romantic partner.

Animal Research (AR) case study

29. (AR) What percentage of the women studied in this sample strongly agreed (gave a rating of 7) that using animals for research is wrong?

30. (AR) Use the proportion you computed in #29. Compute the 95% confidence interval on the population proportion of women who strongly agree that animal research is wrong.

31. (AR) Compute a 95% confidence interval on the difference between the gender means with respect to their beliefs that animal research is wrong.

ADHD Treatment (AT) case study

32. (AT) What is the correlation between the participants' correct number of responses after taking the placebo and their correct number of responses after taking 0.60 mg/kg of MPH? Compute the 95% confidence interval on the population correlation.

Weapons and Aggression (WA) case study

33. (WA) Recall that the hypothesis is that a person can name an aggressive word more quickly if it is preceded by a weapon word prime than if it is preceded by a neutral word prime. The first step in testing this hypothesis is to compute the difference between (a) the naming time of aggressive words when preceded by a neutral word prime and (b) the naming time of aggressive words when preceded by a weapon word prime separately for each of the 32 participants. That is, compute an d for each participant.

a. (WA) Would the hypothesis of this study be supported if the difference were positive or if it were negative?

b. What is the mean of this difference score?

c. What is the standard deviation of this difference score?

d. What is the 95% confidence interval of the mean difference score?

e. What does the confidence interval computed in (d) say about the hypothesis.

Diet and Health (DH) case study

34. (DH) Compute a 95% confidence interval on the proportion of people who are healthy on the AHA diet.

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

The following questions are from ARTIST (reproduced with permission)



35. Suppose that you take a random sample of 10,000 Americans and find that 1,111 are left-handed. You perform a test of significance to assess whether the sample data provide evidence that more than 10% of all Americans are left-handed, and you calculate a test statistic of 3.70 and a p-value of .0001. Furthermore, you calculate a 99% confidence interval for the proportion of left-handers in America to be (.103,.119). Consider the following statements: The sample provides strong evidence that more than 10% of all Americans are left-handed. The sample provides evidence that the proportion of left-handers in America is much larger than 10%. Which of these two statements is the more appropriate conclusion to draw? Explain your answer based on the results of the significance test and confidence interval.
36. A student wanted to study the ages of couples applying for marriage licenses in his county. He studied a sample of 94 marriage licenses and found that in 67 cases the husband was older than the wife. Do the sample data provide strong evidence that the husband is usually older than the wife among couples

applying for marriage licenses in that county? Explain briefly and justify your answer.

37. Imagine that there are 100 different researchers each studying the sleeping habits of college freshmen. Each researcher takes a random sample of size 50 from the same population of freshmen. Each researcher is trying to estimate the mean hours of sleep that freshmen get at night, and each one constructs a 95% confidence interval for the mean. Approximately how many of these 100 confidence intervals will NOT capture the true mean?
- a. None
 - b. 1 or 2
 - c. 3 to 7
 - d. about half
 - e. 95 to 100
 - f. other