

WHAT IS SIGNIFICANT?

Some Elementary Notes on Confidence Intervals and Significance Tests

BU-156-M

S. R. Searle

Abstract

The underlying principles of confidence intervals and significance tests are discussed in terms of a simple hypothetical example. The discussion covers the sampling of a population, the distribution of sample means and the use of the normal distribution. Certain properties of the procedures are also considered briefly. Mathematical derivations are not given, the purpose of the notes being to emphasize the meaning of confidence intervals and significance tests rather than their mathematical niceties.

WHAT IS SIGNIFICANT?

Some Elementary Notes on Confidence Intervals and Significance Tests

BU-156-M

S. R. Searle

Introduction

Research reports often present findings in terms of statistical significance tests, using statements of the nature "the difference was significant, $F < 0.05$ ". Those who have had at least a first course in statistics are usually acquainted with this style of phrase, and it is well appreciated by those who have had additional training in statistics. But since many statistical procedures eventually have recourse to a significance test of one sort or another, it does no harm to review in detail the reasoning underlying confidence intervals and significance tests in general, using a simple hypothetical example as a framework for the discussion. This will cover the sampling of a population, the distribution of sample means, and the use of the normal distribution. Details of mathematical derivations will not be included, the purpose of these notes being to emphasize the meaning of confidence intervals and significance tests rather than their mathematical niceties.

The statistical analysis of almost any research data involves a situation of sampling from a population, be it the population of all cows in the U.S.A., the population of pigs in Iowa or the hypothetical population of beef cattle that may get fed the new grain ration now being tested on 8 or 10 animals at the local experiment station. Usually one or more characteristics of the population are unknown and we seek to learn something about them from information obtained from a sample of the population. For example, we may know the annual milk yields of 500 cows in Virginia, and from these want to make some sort of statement about the average yield

of all cows in Virginia. The average of the 500 cows is an established fact, but there does exist an average of the yields of all the cows in Virginia, and it is concerning this that we want to draw some kind of conclusion based on the average of the sample of 500 cows. Many factors are involved, but four obvious ones are (i) the way in which the 500 cows were chosen from all cows in Virginia, (ii) the magnitude of their average production, (iii) the extent to which cows vary one from another, in their milk yields and (iv) the reasoning by which these facts are assembled into some sort of statement about the mean of the population of cows. We shall not discuss these factors separately, but, so far as (i) and (iii) are concerned, will simply assume that the sample of cows consists of animals selected randomly from the population, and that we know the variability among the milk yields of all cows. Then, rather than develop mathematically the procedure of utilizing the average yield of the sample to make a statement about the population average, we shall just state the procedure and discuss it.

Simple example of a confidence interval

Suppose we want to learn something about the mean daily yield of milk of all cows in the U.S.A., from a random sample of n cows, knowing that the variance, the measure of variability among daily yields, is some value that will be denoted by σ^2 . Then if \bar{X} is the mean of the sample, we can find from many texts on statistics that a confidence interval for the population mean is

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

It might be that the mean of the sample would be reported as

$$\bar{X} \pm \sigma/\sqrt{n} \quad - - - (2)$$

where σ/\sqrt{n} is the standard error of the mean of a random sample of size n . But from the form of (2) we could readily visualize (1), knowing that the confidence level of that interval is, by definition, the probability that it includes the population mean; and we know that for large values of n this probability is 0.95; and (1) is referred to as the 95% confidence interval.

Now what does all this mean? Explanation could well be given entirely in the terms already defined, the sample mean \bar{X} , the population variance, σ^2 , and the sample size n , but in so doing much of the basic understanding of sampling would be lost. It is more illuminating to first consider a hypothetical situation of assuming that we know the distribution of daily milk yields throughout the cows of the U.S.A. The specific implications of using a sample can then be demonstrated and the application of the normal distribution discussed.

We will consider the somewhat trifling situation of using a sample of just 2 cows; i.e. $n = 2$. And if we take the value of σ^2 as 300, the confidence interval in equation (1) becomes

$$\bar{X} \pm 1.96\sqrt{300}/\sqrt{2} = \bar{X} \pm 24. \quad - - - (3)$$

Using this interval is now illustrated.

Hypothetical population

In practice, daily milk yields vary from perhaps 10 pounds up to more than 70 pounds, with a small proportion of cows giving as little as 10 pounds and a small proportion also giving as much as 70 pounds. Different proportions of the total cow population will have yields of 10, 11, 12, 13 ... 70 pounds of milk. Generally speaking we do not know these proportions, - nor do we need to to make use of confidence intervals, but by considering a situation in which we assume we do know them the meaning of confidence intervals can be easily illustrated. For simplicity's sake let us further suppose that all cows give either 10 pounds milk per day, 25 pounds per day or 40, 55 or 70 pounds per day, and that no other daily productions occur; i.e. every cow gives either 10, 25, 40, 55 or 70 pounds milk per day. Now suppose that 1/9th of all cows give 10 pounds daily, 2/9ths give 25 pounds, 3/9ths give 40 pounds, 2/9ths give 55 pounds and 1/9th give 70 pounds. Thus the distribution of daily yields in our hypothetical population is as shown in Table 1.

(Show Table 1)

Sampling

The confidence interval given in equation (3) is based upon a sample of 2 cows. Let us therefore consider the process of sampling our hypothetical population to obtain a random sample of 2 cows. Since each cow in the population has a daily yield which is one of only five distinct values (10, 25, 40, 55 or 70 pounds) there are 25 possible samples, because the first cow drawn for the sample will have one of the five possible yields and so will the second. Thus any sample of 2 cows that we select and use in our confidence interval will be one of the 25 possible samples. We therefore investigate the properties of these 25 possible samples. The probability attaching to each of them depends upon the proportions given in Table 1. For example, the probabilities that cows chosen at random have yields of 10 and 25 are $1/9$ and $2/9$ respectively, so that the probability of such a sample of 2 cows is $2/81$. Since a sample consisting of a 10 pound yield by the first cow chosen and a 25 pound yield by the second has the same mean as a sample which consists of a 25 pound yield by the first cow chosen and a 10 pound yield by the second, these two samples can be considered equivalent for our purposes. So can the samples 10, 40 and 40, 10, and so on. In this way the 25 possible samples can be grouped into 15, the means and probabilities of which are shown in Table 2.

(Show Table 2)

As an example, the probability of the second sample shown in Table 2 is

$$\begin{aligned} & \text{Pr}(\text{first cow's yield is 10 lb. and second cow's yield is 25 lb.}) \\ & + \text{Pr}(\text{second " " " 25 lb. and first " " " 10 lb.}) \\ & = (1/9)(2/9) + (2/9)(1/9) \\ & = 4/81, \end{aligned}$$

and the sample mean is $\frac{1}{2}(10 + 25) = 17\frac{1}{2}$.

Scrutiny of the last column of Table 2 indicates that some of the samples have the same mean; e.g. both (10, 40) and (25, 25) have a mean of 25. The probabilities of these samples are $6/81$ and $4/81$ respectively, so that the total probability of a sample having mean 25 is $10/81$. Table 3 shows the sample means and their total probabilities obtained in this manner.

(Show Table 3)

Conceptually there is a very large number of possible samples of 2 cows that could be drawn from all cows in the U.S.A. - the population of cows we are interested in. Table 3 represents the distribution of sample mean yields in this conceptual population of samples of 2 in the same way that Table 1 represents the distribution of individual cow yields, i.e. $1/81^{\text{th}}$ s of all possible samples of 2 cows have a mean yield of 10, $4/81^{\text{th}}$ s have a mean of $17\frac{1}{2}$ and so on. The mean of the distribution of sample means is 40, the same as the population mean, and the variance is 150, half the population variance, corresponding to the sample size of 2. This conceptual population of sample means is basic to an understanding of confidence intervals and significance tests.

Confidence Intervals

The estimation procedure being considered is the use of a random sample of 2 cows for constructing a confidence interval for the mean daily milk yield of the cows in the U.S.A., assuming a variance of σ^2 . The confidence interval set out in equation (3) is accordingly $\bar{X} \pm 24$.

In practice the mean daily yield of the population of cows is unknown; but if it is 40 (as in Table 1) the interval $\bar{X} - 24$ to $\bar{X} + 24$ contains the value 40 when \bar{X} is between 16 and 64. Furthermore, if the distribution of daily yields is that given in Table 1, the mean of a random sample of 2 cows will be one of the values shown in Table 3; and if it is any of these values except 10 or 70 it will be between 16 and 64; and from Table 3 the probability of a sample mean being other than 10 or 70 is

$4/81 + 10/81 + 16/81 + 19/81 + 16/81 + 10/81 + 4/81 = 79/81 = 0.975, \dots (4)$
corresponding to the possible sample means $17\frac{1}{2}, 25, \dots, 62\frac{1}{2}$. Consequently if \bar{X} is the mean yield of a random sample of 2 cows drawn from the population specified in Table 1, the probability that the interval $\bar{X} \pm 24$ will include the population mean, is 0.975.

One may ask how can $\bar{X} \pm 24$, derived from $\bar{X} \pm 1.96\sigma/\sqrt{n}$, be called a confidence interval with confidence level 0.95 when the probability has been calculated as 0.975? And since obtaining this probability required knowledge of the population mean and distribution, how can this idea of a

confidence interval be used in real life situations where such knowledge is usually lacking? The second question will be answered first, by noting that the mean of the conceptual population of sample means (Table 3) is identical to the population mean. This is no complete answer, because, since the probability 0.975 was obtained by adding and subtracting 2σ to the mean and adding up the probabilities attaching to the possible sample means that fell within this range, it would seem that the mean must be known. At this point use is made of the application of the normal distribution to sample means, for it provides facilities for this adding up of probabilities, without needing to know the mean. It also provides reasons for the 0.95 confidence level and for defining the confidence interval in terms of the expression $1.9\sigma/\sqrt{n}$.

Large samples and the normal distribution

Table 3 shows means and probabilities for samples of size 2; Table 4 shows the same thing for samples of size 4, derived similarly to Table 3 except with $n = 4$ rather than $n = 2$. [The probability of a random sample of size n having a mean value m is the coefficient of t^m (where m exists) in the expansion of $(t^{10/n} + 2t^{25/n} + 3t^{40/n} + 2t^{55/n} + t^{70/n})^n/9^n$.] Figure 1 shows the probabilities in these two tables plotted against possible values for the sample mean. The probabilities for the

(Show Figure 1)

7 different means from samples of 4 are shown as dotted lines, their sum being unity. The probabilities for the 9 different means from samples of 2 are greater than the corresponding probabilities in samples of 4, the differences being shown as solid lines. Hence in this representation the curve for the samples of 2 is above that of the samples of 4. But both curves are similar in shape to the familiar bell-shaped curve of the normal distribution, with their peaks at the mean, the curve for $n = 4$ showing greater similarity than does that for $n = 2$. As n increases, the similarity increases and the vertical lines representing probabilities (whose sum is unity) get more numerous and closer together, until for large values of n the probability curve becomes almost indistinguishable from that of the normal distribution, where the area under the curve represents the total

probability of 1. Hence we come to the general result concerning the application of the normal distribution to means, that the distribution of the means of random samples of size n is approximately a normal distribution with mean identical to the mean of the population from which the random samples are supposedly drawn; the mean shall be denoted by μ . The degree of approximation is greater for large sample size than for small, and the variance is the variance of the sample mean, σ^2/n . Figure 2, which will be discussed shortly, shows the shape of the probability curve of a normal distribution.

(Show Figure 2)

Now we know that in a population having a normal distribution, 95% of the population values lie within 1.96 standard deviation units of the mean. Therefore, since the distribution of our random sample means is approximately normal with mean μ and variance σ^2/n , 95% of the conceptual population of random samples means lie in the range $\mu \pm 1.96\sigma/\sqrt{n}$. And this is exactly the range within which probabilities were added in the example of Table 3, for with $\sigma^2 = 300$, $n = 2$, and $\mu = 40$, $\mu \pm 1.96\sigma/\sqrt{n} = 40 \pm 24$ is the interval 16 to 64, and we found that the probability of a sample mean lying between 16 and 64 was 0.975. That this is not equal to 0.95, but only approximately so, is a consequence of the distribution of random sample means being only approximately normal, an approximation that has greater accuracy for samples of more than 2 observations. (It shall be demonstrated shortly that for $n = 4$ the probability is closer to 0.95).

It has now been shown for random samples of size n that the probability of a sample mean \bar{X} lying in the interval $\mu \pm 1.96\sigma/\sqrt{n}$ is approximately 0.95. How then, is the confidence interval $\bar{X} \pm 1.96\sigma/\sqrt{n}$ derived from this? By considering implications of the above probability statement.

Thus we can write

$$\begin{aligned} 0.95 &\div \Pr(\bar{X} \text{ lies in the interval } \mu \pm 1.96\sigma/\sqrt{n}) \\ &= \Pr(\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}) \\ &= \Pr(\mu \leq \bar{X} + 1.96\sigma/\sqrt{n} \text{ and } \bar{X} - 1.96\sigma/\sqrt{n} \leq \mu) \\ &= \Pr(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}) \\ &= \Pr(\text{the interval } \bar{X} \pm 1.96\sigma/\sqrt{n} \text{ includes } \mu). \end{aligned} \quad \text{---(5)}$$

The interval $\bar{X} \pm 1.96\sigma/\sqrt{n}$ is called the confidence interval, and by definition the probability 0.95 is called the confidence level.

Although the last of the above expressions could be written as $\Pr(\mu \text{ lies in the interval } \bar{X} \pm 1.96\sigma/\sqrt{n})$ it would be read as "the probability that the interval $\bar{X} \pm 1.96\sigma/\sqrt{n}$ includes μ " and not as "the probability that μ lies inside the interval", because μ is a fixed (albeit unknown) constant and not a random variable, and therefore can have no probability statements made about it. The probability applies to the interval, not to μ . Since \bar{X} is the mean of a random sample, it is a random variable from the conceptual population of random sample means. Hence $\bar{X} \pm 1.96\sigma/\sqrt{n}$ is a random interval, and it is to this that the probability statement applies, the probability that this interval includes the population mean. This probability is called the confidence level of the confidence interval, in this example a value of 0.95. It is, in effect, a measure of the "confidence" with which we believe the interval includes the population mean, in the sense that if a large number of random samples of size n are drawn from the conceptual population of such samples, and the interval $\bar{X} \pm 1.96\sigma/\sqrt{n}$ calculated for each, then for approximately 95% of the samples drawn, the calculated interval will include the unknown population mean.

Properties of confidence intervals

(1) Large samples. The example discussed was for $n = 2$. Let us now consider $n = 4$. Then $\mu \pm 1.96\sigma/\sqrt{n} = 40 \pm 17$ is the interval 23 to 57, within which the sample mean must lie for $\bar{X} \pm 17$ to include the mean of 40. From Table 4 the probability of this is

$$(266 + 504 + \dots + 504 + 266)/6561 = 0.951,$$

which is appreciably closer to 0.95 than the 0.975 obtained for $n = 2$. And for $n = 5, 6, 7 \dots$ the probability gets progressively nearer to 0.95, corresponding to the distribution of sample means becoming closer to the normal distribution as n increases.

Notice also that the width of the interval itself gets smaller as n increases; e.g. for $n = 2$ the width was 48 but for $n = 4$ it is 34. Hence for large n the confidence interval is narrower than for small n , although the confidence level is the same. Here we have a clear indication of the value of using large samples, they lead to shorter confidence intervals, which of course are of more practical use than wide ones; e.g. if the variance of 300 is correct, the width of a confidence interval based on 75 cows is 3.92 pounds, compared to 48 pounds using 2 cows. As information

on the daily milk yield of the whole population of cows, the former is considerably more useful than the latter.

(2) Population distributions. The example based on the population distribution shown in Table 1, demonstrated in Figure 1 that means of random samples of size 2 and 4 taken from this population have distributions that are approximately normal. To all intents and purposes this property of random sample means arises no matter what the population distribution is, i.e. for practically any population, the distribution of the conceptual population of means of random samples of size n, is approximately normal with mean equal to the population mean and variance σ^2/n . The degree of the approximation depends largely on n, and to some extent on the population distribution, especially when n is small - and for large n the approximation becomes very close. Hence when dealing with means of reasonably-sized samples we need have no concern for the population distribution in setting up confidence intervals for the population mean.

(3) Confidence levels other than 0.95. There is nothing sacrosanct about the confidence level 0.95. As we have seen, it is based on the property of the assumed normal distribution of the conceptual population of sample means that 95% of the sample means of size n lie between $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$. This is so because of the probability integral

$$\int_{\mu - 1.96\sigma/\sqrt{n}}^{\mu + 1.96\sigma/\sqrt{n}} \frac{\exp \frac{-(x-\mu)^2}{2\sigma^2/n}}{\sigma\sqrt{2\pi/n}} dx = 0.95.$$

Equally well we could apply the result

$$\int_{\mu - 2.58\sigma/\sqrt{n}}^{\mu + 2.58\sigma/\sqrt{n}} \frac{\exp \frac{-(x-\mu)^2}{2\sigma^2/n}}{\sigma\sqrt{2\pi/n}} dx = 0.99$$

and have $\bar{X} \pm 2.58\sigma/\sqrt{n}$ as a 99% confidence interval for the mean, the values 2.58 and 0.99 being obtainable from tables of the standardized normal distribution in the usual way. e.g.

$$(2\pi)^{-\frac{1}{2}} \int_{-1.96}^{+1.96} e^{-\frac{1}{2}x^2} dx = 0.95.$$

Likewise $\bar{X} \pm 1.65\sigma/\sqrt{n}$ is a 90% confidence interval and $\bar{X} \pm 1.28\sigma/\sqrt{n}$ is an 80% confidence interval. In general, for given n , the length of the confidence interval increases if the confidence level is increased, and, as we have seen, for a given confidence level the length decreases for increases in n .

The probability areas of the normal distribution on which the above intervals are deduced by the arguments of equation (5), are illustrated in Figure 2. Note that this does not illustrate confidence intervals themselves.

(Show Figure 2)

but simply probability areas of the assumed normal distribution of the conceptual population of sample means from which the confidence intervals are derived; the intervals cannot be represented in this form for reasons arising out of the discussion following equation (5). Nevertheless, the illustration is useful in giving pictorial representation of the distribution of sample means.

The choice of the confidence level to be used in a particular situation, and accordingly of the confidence interval, is the researcher's, and depends upon the data and the conclusions to be made. If a high degree of assurance is desired, that the interval include the mean, then a confidence level of 0.95 or 0.99 may be chosen, knowing that this involves a wider interval than would be used for a confidence level of 0.90 or 0.80 with the same sample size. And vice versa.

(4) Non-symmetric confidence intervals. The confidence intervals discussed so far have all been of the form $\bar{X} \pm$ a constant and hence are symmetric about the sample mean. They have been derived from probability areas of the normal distribution that are symmetric about the population mean, as illustrated in Figure 2. But just as $\bar{X} \pm 1.96\sigma/\sqrt{n}$ is a 95% confidence interval derived from the standardized normal integral which can be expressed as

$$\begin{aligned}
 0.95 &= (2\pi)^{-\frac{1}{2}} \int_{-1.96}^{+1.96} e^{-\frac{1}{2}x^2} dx \\
 &= 0.475 + 0.475 \\
 &= (2\pi)^{-\frac{1}{2}} \int_{-1.96}^0 e^{-\frac{1}{2}x^2} dx + (2\pi)^{-\frac{1}{2}} \int_0^{1.96} e^{-\frac{1}{2}x^2} dx,
 \end{aligned}$$

so can we also write, for example,

$$0.95 = 0.495 + 0.455 \quad \text{---(6)}$$

$$= (2\pi)^{-\frac{1}{2}} \int_{-2.58}^0 e^{-\frac{1}{2}x^2} dx + (2\pi)^{-\frac{1}{2}} \int_0^{1.70} e^{-\frac{1}{2}x^2} dx$$

so that $\bar{X} - 1.70\sigma/\sqrt{n}$ to $\bar{X} + 2.58\sigma/\sqrt{n}$ is also a 95% confidence interval, one that is not symmetric about \bar{X} . Since there is an infinite number of ways of breaking up 0.95 in the manner of equation (6) there is no unique 95% confidence interval. There is of course only one symmetric confidence interval for any given confidence level.

Hypothesis testing

Confidence intervals are a method of estimation. But instead of estimating the population mean we are sometimes more interested in answering a question like "could the population mean be 43?" This is equivalent to hypothesizing that the mean is 43 and posing the question "are the sample data consistent with such a hypothesis?" More specifically we could ask "to what extent is the sample consistent with the hypothesis that the population mean is 43?" In other instances we may know the mean of one population, have sample data that we believe come from another population, and wish to answer the question "do the populations have the same means?", or "to what extent is the sample consistent with the hypothesis that the populations have the same means?" Statistical procedures available in these and similar situations are the methods of hypothesis testing, a technique that occupies much space in the literature of statistics. Only a brief discussion will be given here, outlining basic concepts. The example of the daily milk yield of dairy cows will be continued, assuming σ^2 known equal to 300. The question to be considered is that just mentioned "could the population mean be 43?"

Suppose that a random sample of 75 cows has yielded a mean of $\bar{X} = 50$ lb.

milk. Now consider the normal distribution given in Figure 2. It represents the approximate distribution of the conceptual population of means of random samples of size n . Since this distribution is a good approximation to the exact distribution of sample means we will consider it as being the exact distribution. Then, from the probability areas indicated in Figure 2, we conclude that for 95% of all possible random samples of size n , the sample mean will be in the range $\mu \pm 1.96\sigma/\sqrt{n}$ whatever the population mean, μ , may be. Now consider the hypothesis that the population mean is 43 lb. milk. If it is true, and we consequently suppose μ is 43, then the probability that the mean of a random sample of n cows lies in the range $43 \pm 1.96\sigma/\sqrt{n}$ is 0.95, i.e.

$$0.95 = \Pr(\bar{X} \text{ lies in the interval } 43 \pm 1.96\sigma/\sqrt{n}).$$

Therefore

$$\begin{aligned} 0.05 &= 1 - 0.95 \\ &= \Pr(\bar{X} \text{ lies outside the interval } 43 \pm 1.96\sigma/\sqrt{n}) \\ &= \Pr(\bar{X} < 43 - 1.96\sigma/\sqrt{n}) + \Pr(\bar{X} > 43 + 1.96\sigma/\sqrt{n}) \\ &= \Pr(\bar{X} - 43 < -1.96\sigma/\sqrt{n}) + \Pr(\bar{X} - 43 > 1.96\sigma/\sqrt{n}) \end{aligned}$$

This can be expressed as

$$\Pr(|\bar{X} - 43| > 1.96\sigma/\sqrt{n}) = 0.05 \quad \text{---(7)}$$

This means that, when (i) assuming as true the hypothesis that 43 is the value of the unknown population mean, and (ii) drawing a single random sample of size n from all possible random samples of size n that exist (conceptually at any rate), the probability is 0.05 that $|\bar{X} - 43|/\sigma/\sqrt{n}$ exceeds 1.96.

Now consider the sample of 75 cows that have a mean of 50. In equation (7) the value of

$$1.96\sigma/\sqrt{n} \text{ is } 1.96\sqrt{300/75} = 3.92$$

and that of

$$|\bar{X} - 43| \text{ is } 7$$

which exceeds 3.92; i.e. the event

$$|\bar{X} - 43| > 1.96\sigma/\sqrt{n} \quad \text{---(8)}$$

has occurred. And from equation (7) we know that this is an event for which the probability of occurrence from among all possible samples of size 75 is 0.05. By this we mean that if the population mean is truly 43 and we draw

a large number of random samples of 75 cows, for 5% of them the event (8) will occur. If now, as human beings with subjective feelings, we contend that any event which has a chance of 1 in 20 (or less) of occurring is an unlikely event, we can say that in our sample a most unlikely event has actually happened. Now on the average, in fact by definition, most unlikely events do not happen very often - and we have had one occur first time. Thus there may be something wrong with our argument. It could be that the 75 cows are not a random sample; but if this were so there would be little or nothing that could be done with the information they provide, so we may as well accept them as a random sample. It could be that the approximation of normality for the distribution of means is not very good - but we have seen that it is quite good for $n = 4$ and gets better as n increases. Finally, it could be that the assumption about the hypothesis being true, namely that the population mean is 43, is wrong. And this is just what we conclude, that this is a false assumption. We speak of rejecting this hypothesis that μ equals 43 and declare that it is not 43 but something different.

Notice that the point on which hangs the decision not to accept the hypothesis as true is our personal feelings towards the occurrence of an event for which the probability of occurrence is only 1 in 20. On the average, in one time in 20, the mean of a random sample of 75 cows chosen from a population of cows having mean 43, will differ from 43 by more than $1.96\sqrt{300}/\sqrt{75}$; but since we feel that a chance of 1 in 20 is a slim one, the actual occurrence of an event that has this probability of happening if the hypothesis is true, makes us feel more inclined to say the hypothesis is not true rather than believe both that it is true and that our sample is one of those that have a slim chance (1 in 20) of happening. The probability level at which we decide to do this, the 1 in 20, is called the significance level, and the conclusion of not accepting the hypothesis is summarized by saying "the population mean is significantly different from 43, at the 5% (significance) level".

Suppose that the mean of the sample of 75 cows had been 45 lb. milk. Such an occurrence is not included in the event (8) and accordingly we would accept the hypothesis that μ is 43 lb., saying that "the population mean is not significantly different from 43, at the 5% level". Notice that this

does not mean that μ is equal to 43 with absolute certainty, but just that the data of the sample are consistent at the 5% significance level, with the population having a mean of 43.

Properties of significance tests

Very briefly we will list some of the properties of significance tests as they relate to our example.

(1) A general notation. The example dealt with testing the hypothesis that the population mean is 43. We might call this the hypothesis H and write $H: \mu = 43$. In general we will want to test the hypothesis $H_0: \mu = \mu_0$ where μ_0 is some pre-assigned numerical value. Equation (7) then becomes

$$\Pr(|\bar{X} - \mu_0| > 1.96\sigma/\sqrt{n}) = 0.05 \quad \text{---(9)}$$

i.e. $\Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.96\right) = 0.05$

or $\Pr(\bar{X} < \mu_0 - 1.96\sigma/\sqrt{n}) + \Pr(\bar{X} > \mu_0 + 1.96\sigma/\sqrt{n}) = 0.05$.

The range of values outside the interval $\mu_0 \pm 1.96\sigma/\sqrt{n}$ is known as the rejection region, or critical region; when \bar{X} lies in this region the hypothesis H_0 is rejected, at the 5% significance level.

(2) Significance levels other than 5%. There is nothing sacrosanct about the significance level 5% any more than there is about the confidence coefficient 95%. Comparable to the confidence intervals discussed earlier $\mu_0 \pm 2.58\sigma/\sqrt{n}$ are the limits of the 1% rejection region of the hypothesis $H_0: \mu = \mu_0$, and $\mu_0 \pm 1.65\sigma/\sqrt{n}$ are the limits of the 10% rejection region. The particular circumstances involved in any practical case determine the choice of significance level just as they do the choice of a confidence coefficient. Different degrees of significance are sometimes given as labels to different significance levels; e.g. 10% may be called significant, 5% very significant, and 1% highly significant.

(3) Alternative hypotheses. The hypothesis $H_0: \mu = \mu_0$ is called the null hypothesis, the word null arising from a commonly-used style of hypothesis, that of testing if some parameter is zero. Indeed H_0 can be put in this form as $H_0: \mu - \mu_0 = 0$.

Let us consider the effect of rejecting the null hypothesis H_0 .

Acknowledging the existence of a mean at all implies, very obviously, that μ does have some value. Therefore in rejecting the hypothesis H_0 we are automatically accepting an alternative one, that the value of μ is something other than μ_0 . This alternative hypothesis can be written as $H_1: \mu \neq \mu_0$. When \bar{X} is in the critical region, i.e. beyond the limits $\mu_0 \pm 1.96\sigma/\sqrt{n}$, we reject H_0 and accept the alternative H_1 .

(4) One-tailed tests. An alternative to H_0 that is often more useful than $H_1: \mu \neq \mu_0$ is $H_1: \mu > \mu_0$, for frequently we are interested in considering the hypothesis that μ is not just different from, but greater than some pre-chosen value, μ_0 . For example, in experimental work where a new treatment is being tried out we would seek to test if the mean yield from the new treatment exceeds that of an older treatment, i.e. is $\mu > \mu_0$? The test is developed as follows.

Consider first that the null hypothesis $H_0: \mu = \mu_0$ is true. Then since

$$\Pr(\bar{X} \text{ lies outside } \mu_0 \pm 1.96\sigma/\sqrt{n}) = 0.05$$

we have $\Pr(\bar{X} > \mu_0 + 1.96\sigma/\sqrt{n}) = 0.025$.

To emphasize that this probability statement is made on the assumption that H_0 is true we write it as

$$\Pr(\bar{X} > \mu_0 + 1.96\sigma/\sqrt{n} | H_0: \mu = \mu_0) = 0.025 .$$

Changing the null hypothesis to $H_0: \mu \leq \mu_0$ therefore gives

$$\Pr(\bar{X} > \mu_0 + 1.96\sigma/\sqrt{n} | H_0: \mu \leq \mu_0) \leq 0.025 ,$$

for if μ is less than μ_0 the probability that X exceeds $\mu_0 + 1.96\sigma/\sqrt{n}$ will be less than 0.025. Hence if an observed \bar{X} exceeds $\mu_0 + 1.96\sigma/\sqrt{n}$ we reject the null hypothesis $H_0: \mu \leq \mu_0$ and accept the alternative that $\mu > \mu_0$, at a significance level that is no greater than 0.025. Such a test would lead to the conclusion " μ is significantly greater than μ_0 ($P \leq 0.025$)."

The test just described is a one-tailed test because it involves only the upper "tail" of the normal probability distribution. The corresponding test for accepting the hypothesis $\mu < \mu_0$ is when $\bar{X} < \mu_0 - 1.96\sigma/\sqrt{n}$, involving the lower "tail" of the distribution. Notice that the significance level of these tests is $2\frac{1}{2}\%$, involving separately the limits used in the two-tailed 5% test. Similarly the limits of the 5% one-tailed tests are those of the two-tailed 10% tests, i.e. accept $\mu < \mu_0$ when $\bar{X} < \mu_0 - 1.65\sigma/\sqrt{n}$ and accept $\mu > \mu_0$ when $\bar{X} > \mu_0 + 1.65\sigma/\sqrt{n}$.

(5) Two types of error. If the null hypothesis is true, a significance test based on the 5% significance level will, on average (i.e. over a large number of samples of the same size) reject the null hypothesis 5% of the time. This rejection constitutes an error of judgment, known as a Type I error, or rejection error, i.e. wrongly rejecting the null hypothesis when it is true. Another type of error occurs if we accept the null hypothesis when in fact it is false, i.e. when the alternative hypothesis is true. This is called a Type II error, or acceptance error, for wrongly accepting the null hypothesis when the alternative is true. These two errors are shown in the following table.

Types of Error in Significance Tests

Null hypothesis	Assertion about null hypothesis derived from significance test	
	True	False
True	No Error	Type I Error
False	Type II Error	No Error

Probabilities of the two types of error are shown below for the null hypothesis $H_0: \mu = \mu_0$ tested against one specific alternative, $H_1: \mu = \mu_1 \neq \mu_0$, from the general class of alternatives $\mu \neq \mu_0$;

$$\begin{aligned} \Pr(\text{Type I Error}) &= \Pr(\text{rejecting } H_0 \text{ when it is true}) \\ &= \Pr(\bar{X} \text{ outside } \mu_0 \pm 1.96\sigma/\sqrt{n} | H_0) \\ &= 0.05 \end{aligned}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu_0 - 1.96\sigma/\sqrt{n}}^{\mu_0 + 1.96\sigma/\sqrt{n}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} dx.$$

$$\begin{aligned} \Pr(\text{Type II Error}) &= \Pr(\text{accepting } H_0 \text{ when } H_0 \text{ is false}) \\ &= \Pr(\text{accepting } H_0 \text{ when } H_1 \text{ is true}) \\ &= 1 - \Pr(\text{rejecting } H_0 \text{ when } H_1 \text{ is true}) \\ &= 1 - \Pr(\bar{X} \text{ outside } \mu_0 \pm 1.96\sigma/\sqrt{n} | H_1) \end{aligned}$$

$$= \frac{1}{2\pi\sqrt{\sigma}} \int_{\mu_0 - 1.96\sigma/\sqrt{n}}^{\mu_0 + 1.96\sigma/\sqrt{n}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx.$$

In general we should like both these probabilities as small as possible in a significance test, in fact to design a significance test so that this is the case. Unfortunately this generally cannot be achieved for when the probability of making one error is decreased that of the other is increased. One suggestion has been made that for a given value of the probability of a Type I error the critical region should be chosen to minimize the probability of a Type II error.

(6) Power of a test. The power of a significance test is

$$\begin{aligned} \text{Power of test} &= 1 - \text{Pr}(\text{Type II error}) \\ &= \text{Pr}(\text{rejecting } H_0 \text{ when } H_1 \text{ is true}), \\ &= \text{Pr}(\text{rejecting } H_0 \text{ when } H_0 \text{ is false}), \end{aligned}$$

i.e. the power of a test is the probability of rejecting the null hypothesis when it is false and should be rejected.

Figure 3 shows a representation of the normal distributions of sample means under the two hypotheses $H_0: \mu = \mu_0$ and $H_1: \mu = \mu_1$. Points P and Q

(Show Figure 3)

represent the limits of the critical region $\mu_0 \pm 1.96\sigma/\sqrt{n}$; when \bar{X} lies beyond these limits H_0 is rejected. The area beyond these limits under the normal distribution having mean μ_0 , i.e. $(-\infty)AP$ plus $BQ(+\infty)$, is the probability of a Type I error, with value 0.05 in this case. The area between the same limits but under the normal distribution having mean μ_1 , namely PA^*B^*Q represents the probability of a Type II error, the probability of accepting H_0 when H_1 is true. The sum of the areas $(-\infty)A^*P$ plus $B^*Q(+\infty)$ is the power of the test, the probability of rejecting H_0 when H_1 is true.

Conclusion

Although the errors and the power of a significance test occupy much space in statistical literature, they are treated only briefly here to be in keeping with the general tenor of these notes, namely a discussion of underlying principles rather than mathematical complexities. For the same reason, in the example used throughout - a simple one of means - the variance has been assumed known. In practice it is often unknown and confidence intervals and significance tests must then be based on an estimated variance such as $s^2 = (\sum X^2 - n\bar{X}^2)/(n - 1)$. The hypothesis

$H_0: \mu = \mu_0$ is then considered by testing $(\bar{X} - \mu_0) \sqrt{n}/s$ against the t-distribution, using its values for the various probability levels. The principles of confidence intervals and significance tests remain the same, just as they do for other statistics and distributions, for example those of the analysis of variance procedure. There may be small changes in some instances but the basic concepts are unaltered.

Summary

The underlying principles of confidence intervals and significance tests are discussed in terms of a simple hypothetical example. The discussion covers the sampling of a population, the distribution of sample means and the use of the normal distribution. Certain properties of the procedures are also considered briefly. Mathematical deviations are not given, the purpose of the notes being to emphasize the meaning of confidence intervals and significance tests rather than their mathematical niceties.

Table 1

Hypothetical population of daily milk yields

<u>Daily yield</u> <u>(lb. milk)</u>	<u>Proportion</u> <u>of cows</u>
10	1/9
25	2/9
40	3/9
55	2/9
70	1/9
	<hr/>
<u>Total</u>	<u>9/9 = 1</u>

$$\text{Mean} = (1/9)10 + (2/9)25 + (3/9)40 + (2/9)55 + (1/9)70 = 40$$

$$\begin{aligned} \sigma^2 &= (1/9)(10-40)^2 + (2/9)(25-40)^2 + (3/9)(40-40)^2 + (2/9)(55-40)^2 \\ &\quad + (1/9)(70-40)^2 = 300 \end{aligned}$$

Table 2

Unordered samples of 2 obtainable from hypothetical population

<u>Sample of yields</u>	<u>Probability of obtaining sample</u>	<u>Mean yield of sample</u>
10, 10	1/81	10
10, 25	4/81	17 $\frac{1}{3}$
10, 40	6/81	25
10, 55	4/81	32 $\frac{1}{2}$
10, 70	2/81	40
25, 25	4/81	25
25, 40	12/81	32 $\frac{1}{2}$
25, 55	8/81	40
25, 70	4/81	47 $\frac{1}{2}$
40, 40	9/81	40
40, 55	12/81	47 $\frac{1}{2}$
40, 70	6/81	55
55, 55	4/81	55
55, 70	4/81	62 $\frac{1}{2}$
70, 70	1/81	70
<u>Total</u>	<u>81/81 = 1</u>	

Table 3

Sample means and associated probabilities, of samples of size 2 obtainable from hypothetical population of Table 1

<u>Sample mean</u>	<u>Probability</u>
10	1/81
17½	4/81
25	10/81
32½	16/81
40	19/81
47½	16/81
55	10/81
62½	4/81
70	1/81
<u>Total</u>	<u>81/81 = 1</u>

$$\text{Mean} = (10 + 4 \cdot 17\frac{1}{2} + 10 \cdot 25 + 16 \cdot 32\frac{1}{2} + 19 \cdot 40 + 16 \cdot 47\frac{1}{2} + 10 \cdot 55 + 4 \cdot 62\frac{1}{2} + 70) / 81 = 40$$

$$\text{Variance} = [2(30)^2 + 2(4)(22\frac{1}{2})^2 + 2(10)(15)^2 + 2(16)(7\frac{1}{2})^2] / 81 = 150$$

Table 4

Sample means and associated probabilities of samples of size 4 obtainable from hypothetical population of Table 1

<u>Sample average</u>	<u>Probability</u>
10	1/6561
13.75	8/6561
17.5	36/6561
21.25	112/6561
25	266/6561
28.75	504/6561
32.5	784/6561
36.25	1016/6561
40	1107/6561
43.75	1016/6561
47.5	784/6561
51.25	504/6561
55	266/6561
58.75	112/6561
62.5	36/6561
66.25	8/6561
70	1/6561
<u>Total</u>	<u>6561/6561 = 1</u>

Figure 1. Distribution of means of random samples.

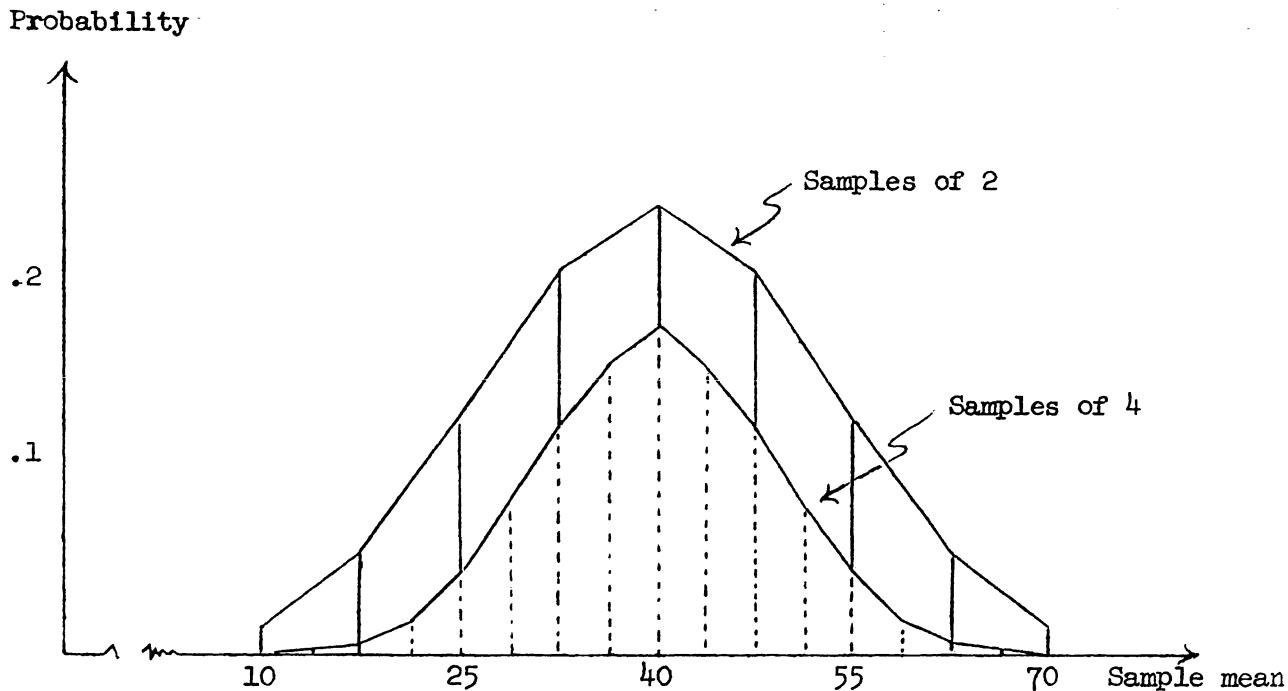


Figure 2. Probability areas in the normal distribution of mean μ and variance $s^2 = \sigma^2/n$.

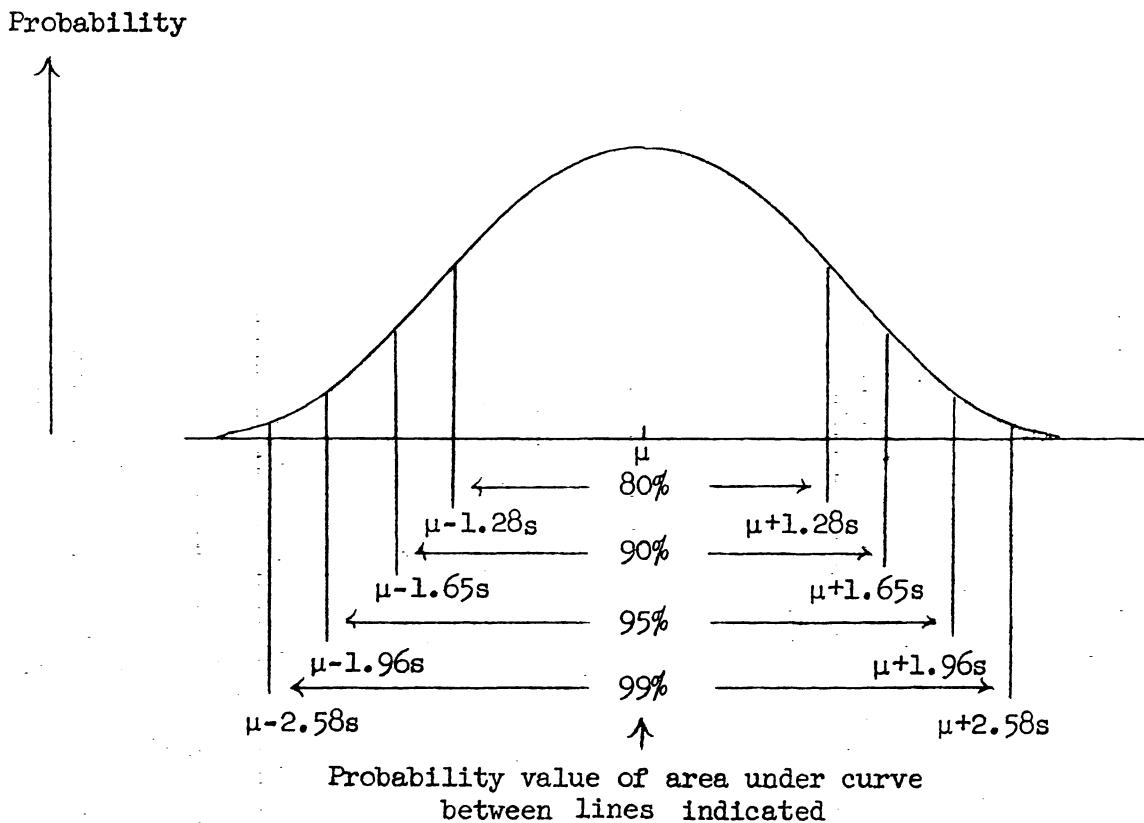


Figure 3. Normal distributions with variance σ^2/n , having means μ_0 and μ_1 corresponding to the null hypothesis $H_0: \mu = \mu_0$ and the alternative $H_1: \mu = \mu_1$.

