Significance and Confidence

When to Call For

New Dice

A Presentation by

Dr. Peter Crickmore

CEI Inc.

Significance, Confidence Level and Confidence Interval

Confidence Interval for a Mean

Are Two Means Significantly Different?

Confidence Interval for a Variance

Are Two Variances Significantly Different?

Confidence Intervals for Regression Parameters

Significance?

- An occurrence is significant if it is unlikely to have arisen from random variation alone.

When a player in a dice game throws a 12 with two dice, that is not significant. If he throws thirty-seven 12s in a row, that is significant. What if he throws 12 three times? Is that significant? Should you call for new dice (assuming you were betting against him)? How confident would you be?

Definitions

- We also need to define confidence level and confidence interval and see how these relate to the concept of significance.

- The confidence level that we associate with a statement is the probability that the statement is true.

Definitions, continued

- Statisticians say that a result is significant if the confidence level that the result was unlikely to have arisen from random variation alone is 95% or greater.

Definitions, continued

- Alternatively, a result is significant if the confidence level that the result was likely to have arisen from random variation alone is 5% or less.

Aside, Back to Dice

- If we rely on random chance alone, the probability of rolling 12 with two dice three times in a row is 1/36 x 1/36 x 1/36 or about 21/1,000,000. This would be a very rare event (certainly occurring less than 5% of the time, on average), so that we would judge it to be significant.

Definitions, continued

- A confidence interval for a parameter (e.g. mean, variance, slope of a regression line) is an interval in which we have a particular confidence level that the true value of the parameter is to be found.
- The most common confidence intervals are those associated with a 95% confidence level (so we can talk about "significance")

???

These definitions become a lot clearer
after we look at some examples.

Example 1: Confidence Intervals When We Know Almost Everything

- A mill is turning out oriented strand board with a nominal thickness of ½ inch. Based on a large number of measurements of boards over several years of production, the variation in actual thickness of the boards produced follows a Gaussian distribution with a mean of 0.509 inch and a variance of 0.0015 $inch^2$.

Example 1: Confidence Intervals When We Know Almost Everything, continued

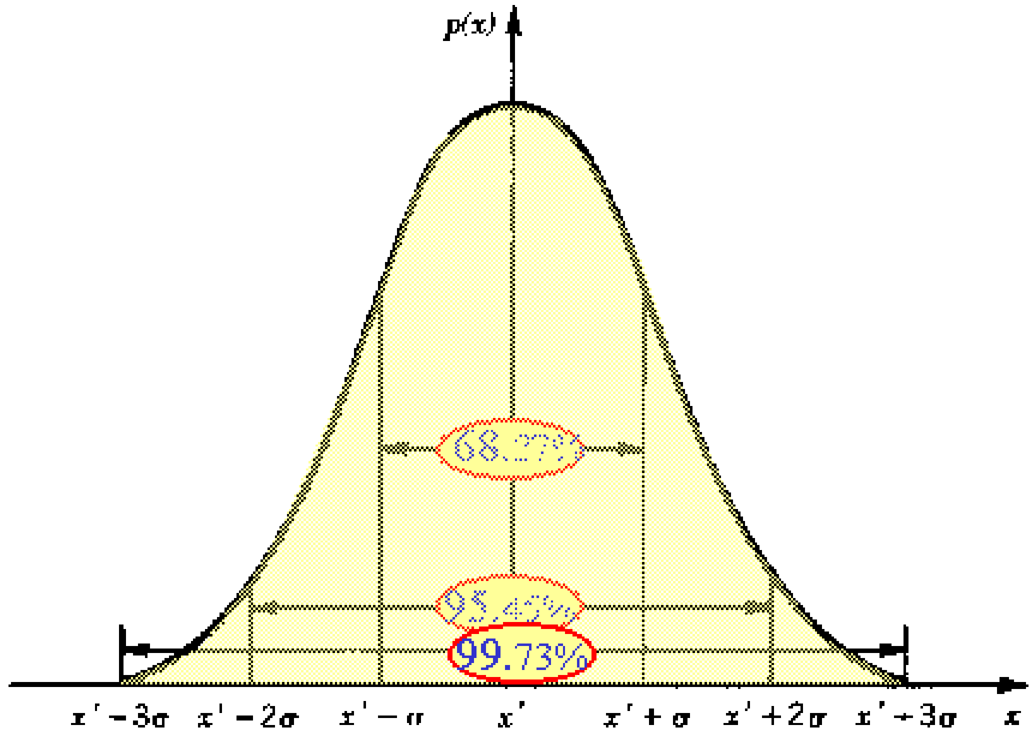
- If a board is selected at random from the mill's production, what is the 95% confidence interval for the measured thickness?

Gaussian?

Normal-Gaussian distribution

$$P(-z_1 \le \beta \le z_1) = \frac{1}{(2\pi)^{1/2}} \int_{-z_1}^{z_1} e^{-\beta^2/2}\, d\beta = 2\left[\frac{1}{(2\pi)^{1/2}} \int_0^{z_1} e^{-\beta^2/2}\, d\beta\right]$$

$\beta = (x - x')/\sigma$



Example 1: Confidence Intervals When We Know Almost Everything, continued

- A check with standard tables for the Gaussian distribution shows that indeed 95% of the area of the probability density function lies within ± 1.960 standard deviations from the mean. For our example, the standard deviation is 0.03873. Thus, the 95% confidence interval would be 0.509 ± (1.960) x (0.03873)

Example 1: Confidence Intervals When We Know Almost Everything, continued

- This works out to a confidence interval of

  {0.43309 , 0.58491}, or better as

  {0.43 , 0.58}

Example 1: Confidence Intervals When We Know Almost Everything, continued

- Note that there were a number of possibilities for the confidence interval. We chose the only one that was centred on the mean, because we felt that readings above the mean and below the mean were equally likely.

Why Example 1 is Unrealistic

- We assumed that we knew the mean, the variance, and the type of probability density function followed by our variable. Usually, we have only a few experimental readings, the mean is unknown, the variance is unknown and we have no real idea of the probability density function.

Confidence Interval for the Mean With Only a Few Data Points

Example 2: Confidence Interval for the Mean With only a Few Data Points

- Consider a pile of oil sand that has not been homogenized. Five spot samples are taken from the pile and submitted for O/W/S analysis. The following bitumen assays are determined: 11.3%, 8.4%, 10.9%, 12.1% and 11.6%. What is the 95% confidence interval for the mean?

Example 2, continued

- Using a calculator or a spreadsheet, we can calculate "best" estimates of the mean (μ) and variance (σ^2) of the bitumen concentration distribution:

$$x_{av} = (\Sigma\, x_i\,)/\, n \;=\; 10.86$$

$$s^2 = (\Sigma\,(x_i - x_{av})^2)\,/\,(n - 1) \;=\; 2.08$$

Example 2, continued

- Note that we use x_{av} rather than μ and s^2 rather than σ^2 to signify that these "best" estimates might not be very good.
- If we wanted to ignore the fact that we didn't know the variance and didn't know whether a Gaussian probability density function was appropriate, we could naively say that:

the 95% confidence interval for μ was $x_{av} \pm 1.960 \times s$. In numbers, that would be

$$\{8.04 , 13.68\}$$

But what if σ^2 is much larger than the estimate given by s^2? The 95% C.I. for μ would have to be much larger.

Introducing the t-statistic

- It can be shown that the $(1 - \alpha) \times 100$ % confidence interval for the mean (if we have n observations) is given by

$$\{ X_{av} \pm t_{(n-1), (1 - \alpha/2)} \times s \}$$

For a 95% confidence interval, this is

Introducing the t-statistic, continued

$$\{ X_{av} \pm t_{((n-1), (0.975))} \text{ x } s\}$$

Returning to our example, the 95% confidence interval for the mean bitumen assay of the oil sand in the pile is

$$\{ 10.86 \pm 2.776 \text{ x } 1.443 \} \text{ or}$$

$$\{ 6.85 , 14.87 \}$$

This is much wider than if σ^2 was equal to s^2.

Aside on nomenclature

The quantity (n – 1) used in looking up the appropriate t statistic is normally called the "number of degrees of freedom"

Aside on homogenization and compositing

To the oil sand researcher, the confidence interval for the mean bitumen assay may be too wide to be useful. A tighter bound on the mean (with the same number of samples analyzed) can be obtained in two ways:

1. If you don’t mind reducing the point-to-point variation, homogenize the pile.

2. If you want to preserve the variance in the pile, use composites instead of point samples.

For instance, if each of the five samples assayed were a composite of five smaller point samples, you would expect the variance (and the width of the confidence interval) to decrease by a factor of $\sqrt{5}$ (or 2.236)

Summarizing Confidence Intervals for the Mean With only a Few Data Points

Thus, if we only have a few observations and we do not know the true mean or variance, we can use the t-statistic to get the appropriate confidence interval for the mean.

Have we "finessed" anything here?

What we "finessed"

The t distribution is exact if and only if the true distribution describing the population from which samples are taken is Gaussian. In our example, we do not know this (it is probably a reflected gamma distribution, but the data in the example doesn't tell us this).

The Gaussian distribution is as good a guess as any with only a few samples and has the advantage that it does not assume that either positive or negative deviations from the mean are more likely. It would require 500 + samples to specify the underlying distribution with any degree of precision.

Biographical Note

In older texts, the t-distributions is often referred to as “Student’s t-distribution”. “Student” was the pseudonym of William S. Gosset who was a chemist in charge of quality control at the Guinness Breweries in Dublin. Guinness had a policy against publishing, so Gosset had to use a pseudonym for his early papers.

He also referred to the example of counting blood cells (rather than yeast cells which generated his actual data). Gosset's papers appeared in 1907 and 1908[1,2].

1 Student, "On the error of counting with a haemocytometer", Biometrika 5, 351 (1907).

2 Student, "The probable error of a mean", Biometrika 6, 1 (1908).

William S. Gosset

"Student"

CEI Inc.

General Problem of Comparing Two Means

General Problem of Comparing Two Means

- Often we are asked to compare means:
  - Is the mean bitumen content of pile A different from that of pile B?
  - Is the mean response of instrument A when running a standard the same as instrument B running the same standard?
  - Has the mean nitrogen assay of hydrotreated gas oil changed after a new hydrotreating catalyst was introduced?

General Problem of Comparing Two Means, cont'd

- We can change this new problem into the problem we have just solved.
- We say that μ_A is significantly different from μ_B if the 95% confidence interval for $(\mu_A - \mu_B)$ does not include zero.
- Why does this help?

General Problem of Comparing Two Means, cont'd

Assume that the underlying populations are both Gaussian. If property A is distributed with mean μ_A and variance σ^2_A, property B is distributed with mean μ_B and variance σ^2_B and both are Gaussian, then A – B is distributed with mean $\mu_A - \mu_B$, variance $\sigma^2_A + \sigma^2_B$ and is itself Gaussian.

That means we can use the t statistic to gain a confidence interval for the unknown mean of A – B. If the C.I. does not include zero, we say that the two means are significantly different.

Confidence Interval for the Difference of Two Means With Only a Few Data Points

Example 3: C. I. for the Difference of Two Means With Only a Few Data Points

Consider two piles of oil sand, one from an estuarine environment of deposition and one from a marine. Neither pile has been homogenized. Five spot samples are taken from the estuarine pile and analyzed for O/W/S; six spot samples from the marine pile are similarly treated. The bitumen assays from the estuarine pile are:

{10.9, 12.3, 13.1, 11.9, 10.6},

while the bitumen assays for the marine pile are:

{7.6, 14.1, 8.5, 12.9, 7.3, 13.3}.

Are the mean bitumen assays for the piles significantly different?

If we have n samples from population A that lead to $x_{av,A}$ and s^2_A as estimates of A's mean and variance and m samples from population B that lead to $x_{av,B}$ and s^2_B as estimates of B's mean and variance, it can be shown that a 100(1 - α) % confidence interval for $\mu_A - \mu_B$ is given by

$$\{(x_{av,A} - x_{av,B}) \pm b\, R^{1/2}\}$$

where b is the t statistic with n + m – 2 degrees of freedom at the 1 - α /2 level, and

$R = Q\,(n\,s^2_A + m\,s^2_B)/(n + m - 2)$ and

$Q = 1/n + 1/m$

For our example, a calculator gives $x_{av,e} = 11.76$ and $s^2_e = 1.048$; $x_{av,m} = 10.62$ and $s^2_m = 10.395$. We will set α to 0.05 to get a 95 % confidence interval. Since n is 5 and m is 6, the t statistic with 9 degrees of freedom at the 0.975 level is 2.262 (this is b). Continuing, Q is 0.36667, R is 2.754 and the 95 % confidence interval for $\mu_A - \mu_B$ is

{-2.614 , 4.894}

Thus we cannot say that the mean bitumen assays of the two piles are significantly different (since the C.I. includes zero.

Confidence Interval for the Variance With Only a Few Data Points

Example 4: Confidence Interval for the Variance With Only a Few Data Points

- Return to Example 2 and consider a single pile of oil sand that has not been homogenized. Recall that five spot samples were taken from the pile and submitted for O/W/S analysis. The following bitumen assays were found: 11.3%, 8.4%, 10.9%, 12.1% and 11.6%. What is the 95% confidence interval for the variance?

Example 4: C. I. for the Variance With Only a Few Data Points, cont'd

Recall, from Example 2, that

$$s^2 = (\Sigma (x_i - x_{av})^2) / (n - 1) = 2.08$$

Gosset showed that a (1 - α) x 100 % confidence interval for the variance is

$$\{ n s^2 / b , n s^2 / a \} \text{ , where}$$

a is $\chi^2_{((n-1), \alpha/2)}$ and b is $\chi^2_{((n-1), 1-(\alpha/2))}$.

Note: by tradition, χ^2 is read as "chi square", not "chi squared"

Example 4: C. I. for the Variance With Only a Few Data Points, cont'd

From the example, n is 5, n – 1 is 4, s^2 is 2.08 and α is 0.05. Turning to the χ^2 tables, a is 0.484 and b is 11.1. Thus, the 95% confidence interval for the variance is

$$\{ 4 \times 2.08 / 11.1 , 4 \times 2.08 / 0.484 \}, \text{ or}$$

$$\{ 0.750 , 17.19 \}.$$

Note that the 95% C.I. for the variance is very much broader than that for the mean.

Biographical Note

The χ^2 distribution was derived in 1900 for "goodness of fit" tests by Karl Pearson[1]. W.S. Gosset in his 1908 paper on the t distribution was the first to show that the χ^2 distribution gave the appropriate confidence interval for variances[2].

1 K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", Phil. Mag. 1, 157 (1900)

2 Student, "The probable error of a mean", Biometrika 6, 1 (1908).

Karl Pearson

General Problem of Comparing Two Variances

General Problem of Comparing Two Variances

- We say that two variances are significantly different if the confidence interval for their ratio does not include 1
- How do we determine a confidence interval for the ratio of two variances?

Confidence Interval for the Ratio of Two Variances With Only a Few Data Points

Example 5: C. I. for the Ratio of Two Variances With Only a Few Data Points

Return to Example 3 and consider two piles of oil sand, one from an estuarine environment of deposition and one from a marine. Neither pile has been homogenized. Five spot samples are taken from the estuarine pile and analyzed for O/W/S; six spot samples from the marine pile are similarly treated. The bitumen assays from the estuarine pile are {10.9, 12.3, 13.1, 11.9, 10.6}, while the bitumen assays for the marine pile are {7.6, 14.1, 8.5, 12.9, 7.3, 13.3}.

Example 5: C. I. for the Ratio of Two Variances With Only a Few Data Points

Are the variances in bitumen assays for the piles significantly different?

If we have n samples from population A that lead to s^2_A as an estimate of A's variance and m samples from population B that lead to s^2_B as an estimate of B's variance, it can be shown that a $100(1 - \alpha)$ % confidence interval for σ^2_B / σ^2_A is given by

$$\{R / F_{((m-1), (n-1))} , R F_{((n-1), (m-1))}\}$$

where $F_{(x,y)}$ is the F statistic with x and y degrees of freedom at the $1 - \alpha /2$ level and R is given by

$$R = (m\, s^2_B / (m-1)) / (n\, s^2_A / (n-1))$$

In our particular example, α is 0.05, n is 5, m is 6, s^2_E is 1.048 and s^2_M is 10.395. R is then calculated to be 9.522, $F_{(4,5)}$ is 7.39 and $F_{(5,4)}$ is 9.36. Thus our 95 % confidence interval for σ^2_M / σ^2_E is

{1.017 , 70.368}

Note that the confidence interval for the ratio of the variances does not include 1. We can then say that the variances of the bitumen assays for the estuarine and the marine oil sand are significantly different (at the 95 % level).

Biographical Note

The F-distribution was derived by Ronald A. Fisher for the express purpose of giving confidence intervals for the ratio of two variances and presented in 1924[1] to aid in the interpretation of agricultural tests. Fisher is usually considered to be the father of "analysis of variance" techniques, the concept of "likelihood", and the field of experimental design.

1 R.A. Fisher, "On a distribution yielding the error function of several well known statistics", Proceedings of the International Congress of Mathematics, Toronto 2, 805 (1924).

Ronald A. Fisher

Confidence Intervals for Other Parameters

Confidence Intervals for Other Parameters

We have studied confidence intervals for two parameters that represent populations -- the mean and the variance. Confidence intervals can also be specified for other quantities estimated from data; the most common of these are the best-fit parameters from a regression.

Confidence Intervals for Regression Parameters

Example: Fines vs Bitumen

(with apologies to F. W. Camp)

Consider the following six oil sand samples and their corresponding bitumen assays and –325 mesh fines determinations. Find the best straight line that will fit the data, determine 95 % confidence intervals for the regression coefficients, and a 95 % confidence interval for the fines level predicted for a 10 % bitumen oil sand.

% B	7.1	16.0	14.3	4.9	11.3	12.0
% F	22	5	10	32	15	17

There are many ways to find the least squares estimates of β_0 and β_1 for the relationship

$$\% \text{ Fines} = \beta_0 + \beta_1 \times (\% \text{ Bitumen}).$$

A spreadsheet or a calculator or first principles can be used; if we work from first principles, we end up with the two equations

$$N \beta_0 + \beta_1 \Sigma (\%B) = \Sigma(\%F)$$

$$\beta_0 \Sigma (\%B) + \beta_1 \Sigma (\%B)^2 = \Sigma (\%B)(\%F)$$

where N is the number of data pairs.

Using our data leads to

$$6\,\beta_0 + 65.8\,\beta_1 = 101.0, \quad \text{and}$$

$$65.8\,\beta_0 + 811.16\,\beta_1 = 912.5\ .$$

Solving gives $\beta_0 = 40.73$ and $\beta_1 = -2.18$ or

% fines = 40.73 - 2.18 x (% bitumen).

From the posing of the question (and from common sense), it would seem that the regression coefficients are not exact. What are the 95 % C.I.s for the two β parameters?

It can be shown that, for the relationship $y = \beta_0 + \beta_1 x$, the 100 $(1 - \alpha)$ % confidence interval for β_0 is given by

$$\{b_0 \pm s_E(b_0)\, t_{((N-2),\, 1-\alpha/2)}\}$$

where b_0 is the best estimate of β_0, $s_E(b_0)$ is given by

$$s^2_E(b_0) = D^2 \Sigma x^2 / (N \Sigma (x - x_{av})^2) ,$$

$D^2 = (\Sigma (y_p - y)^2) / (N - 2)$ and y_p is the predicted value of y.

Similarly, the 100 (1 - α) % confidence interval for β_1 is given by

$\{b_1 \pm s_E(b_1)\, t_{((N-2),\, 1-\alpha/2)}\}$

where b_1 is the best estimate of β_1, $s_E(b_1)$ is given by

$s^2_E(b_1) = D^2 / (\Sigma\, (x - x_{av})^2)$,

$D^2 = (\Sigma(y_p - y)^2) / (N - 2)$ and y_p is the predicted value of y.

For Our Example

For our example, $N = 6$, $b_0 = 40.73$, $b_1 = -2.18$, $\alpha = 0.05$, $x_{av} = 10.967$, $t_{4,0.975} = 2.776$, $D^2 = 5.41175$, and $s_E(b_0) = 2.858$. Putting these into the formula for the confidence interval for β_0 leads to being 95% confident that β_0 is in the interval

{32.8, 48.7}.

For Our Example, cont'd

Adding that $\Sigma x^2 = 811.16$, we are then 95% confident that β_1 is in the interval

{-2.86, -1.50}

Note that neither of these C.I.s include 0, so the inclusion of either of these terms is not in doubt.

For Our Example, cont'd

Our example asked us to find the 95% confidence interval for the percentage fines associated with a 10% bitumen oil sand. It is easy to get the best point estimate (y_p):

% fines = 40.73 - 2.18 x (10) = 18.93.

It can be shown that the 100 (1-α)% confidence interval for y at x_0 is given by

$$\{y_p \pm t_{N-2,\, 1-\alpha/2}\, D\, (1/N + (x_0 - x_{av})^2/(\Sigma(x-x_{av})^2))^{1/2}\}$$

For Our Example, cont'd

Substituting our values gives the 95 % C.I. for the % fines to be expected for a 10% bitumen oil sand (based on our data) as

{16.2 , 21.6}.

Note that because of the $(x_0 – x_{av})^2$ term, the confidence interval will be smallest at the "middle" of the data and increase with distance from this point.

What Have We Finessed?

In the above, we have made the following assumptions:

1. All of the “errors” follow a Gaussian distribution
2. The variance does not vary with the independent (or dependent) variable.

In order to remove these assumptions, we would have to have a large amount of data (500 points plus). With plant data, these may be possible, but for a research project, this is unlikely.

Other Issues

1. How to use the F-test to determine significance of a regression
2. How to sample in a "representative" manner, what size of sample to take,…
3. How to handle large amounts of data (Pearson's system of frequency curves, etc.
4. How to best gather data for regressions.
5. How to analyse "time series"
6. How to analyse fuzzy data