DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

Klaus Röder AECOM

www.klaus-roeder.com

Sampling and improving quality for testing food and food products

Workshop Baguio Dec. 01.-03. 2015

29.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Workshop schedule

Dec.01: 8:30-10:00 1st Morning session: Opening, Lectures
10:00-10:30 Coffee break
10:30-12:30 2nd Morning session: Lectures/group work participants
12:00-13:30 Lunch break
13:30-15:00: 1st Afternoon session: Lectures/group work participants
15:00-15:30 Coffee break
15:30-17:30 2nd Afternoon session: Day's synopsis by lecturer/group work participants

Dec.02: as above

Dec.03:
8:30-10:00 1st Morning session; 10:00-10:30 Coffee break
10:30-12:30: 2nd Morning session: Workshop's synopsis by lecturer and group assessment by participants

29.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

Page 2

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Workshop program Day 2 Morning

Day 2/ 1. and 2.Morning Session:
Sampling and statistical testing
The null hypothesis
Parametric tests and non-parametric tests
Type I errors and type II errors
Standard Error
Example calculations of sample size
Introduction to statistical regression
The least square solution

Group Work participants:

Group Work: Which statistical testing method applies to practices of the work of NPAL and other organizations and how and which method could or should applied for the work of NPAL and other organizations - Formulate of Test Hypothesis and distinguish between the two possible errors, when induct from the sample on the population (if possible for the 4 different types of sampling points)
Applying exercises on Regression Theory, if this falls into the area of work of the group participants

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 3

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Sampling and statistical testing

Research is conducted in order to determine the acceptability (or otherwise) of hypotheses. Having set up a hypothesis, we collect data which should yield direct information on the acceptability of that hypothesis. This empirical data requires to be organised in such a fashion as to make it meaningful. To this end, we organise it into frequency distributions and calculate averages, measures of spread or percentages. But often, these statistics on their own mean very little. The data we collect often requires to be compared and when comparisons have to be made, we must take into account the fact that our data is collected from a sample of the population and is subject to sampling and other errors. Today's subject is concerned initially with the statistical testing of sample data. One assumption which is made is that the survey results are based on random probability samples.

A Hypothesis Testing Experiment: The Lady tasting tea

The experiment provided the Lady with 8 randomly ordered cups of tea – 4 prepared by first adding milk, 4 prepared by first adding the tea. "The lady" claimed to be able to tell whether the tea or the milk was added first to a cup. She was to select 4 cups correctly.
The null hypothesis was that the Lady had no such ability.
The lady correctly identified every cup, which would be considered a statistically significant result

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 4

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

The null hypothesis (1)

The first step in evaluating sample results is to set up a null hypothesis (Ho). The null hypothesis is a hypothesis of no differences. We formulate it for the express purpose of rejecting it. It is formulated before we collect the data . For example, we may wish to know whether a particular promotional campaign has succeeded in increasing awareness amongst housewives of a certain brand of biscuit. Before the campaign we have a certain measure of awareness, say x%. After the campaign we obtain another measure of the awareness, say y%. The null hypothesis in this case would be that "there is no difference between the proportions aware of the brand, before and after the campaign",

Since we are dealing with sample results, we would expect some differences; and we must try and establish whether these differences are real (i.e. statistically significant) or whether they are due to random error or chance.

If the null hypothesis is rejected, then the alternative hypothesis may be accepted. The alternative hypothesis (H1) is a statement relating to the researchers' original hypothesis. Thus, in the above example, the alternative hypothesis could either be:

a. H1: There is a difference between the proportions of housewives aware of the brand, before and after the campaign,

or

b. H1: There is an increase in the proportion of housewives aware of the brand, after the promotional campaign.

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 5

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

The null hypothesis (2)

Note that these are clearly two different and distinct hypotheses. Case (a) does not indicate the direction of change and requires a two-tailed test. Case (b), on the other hand, indicates the predicted direction of the difference and a one-tailed test is called for.

"two Tailed vs singleTailed'(->Normal distribution to come later in the workshop)

Source http://dianayu24.blogspot.com/2011/04/two-tailed-vs-single-tailed.html

Some brief distinction between them:

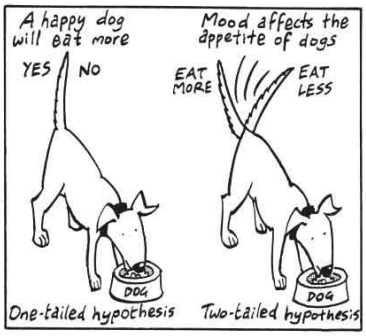
one-Tailed: "Defined-Prediction"

1. Researchers expect that experiment procedure has an influence to clear direction (can be defined).
2. Significance test with one-tailed test must be chosen.

two-Tailed: "Non-defined Prediction“

e.g.

H0: There is a positive influence between intellectual rate of kids and nutrition intake while in period of preschool. Is it two Tailed or singleTailed ?



29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 6

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Parametric tests and non-parametric tests

The next step is that of choosing the appropriate statistical test. There are basically two types of statistical test, parametric and non-parametric. Parametric tests are those which make assumptions about the nature of the population from which the scores were drawn (i.e. population values are "parameters", e.g. means and standard deviations). If we assume, for example, that the distribution of the sample means is normal, then we require to use a parametric test. Non-parametric tests do not require this type of assumption.

The example of the "The Lady tasting tea"

There are 70 possible combinations of choosing 4 cups correctly/ incorrectly. If we have set a significance level of 5% or 0.05, then the probability of choosing correct combinations is 1/70 = 0.014 or 1.4% . So we reject H_0 .Meaning ? 1-or 2-tailed?

When the term "statistical significance" is used , it simply means that enough data have been collected to establish that a difference does exist. In other words, statistical significance is a technical term with a far different meaning than ordinary significance. Unfortunately but understandably, many people tend to confuse statistical significance with ordinary significance.

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Type I errors and type II errors

The choice of significance level affects the ratio of correct and incorrect conclusions which will be drawn. Given a significance level there are four alternatives to consider:

Correct Conclusion	Incorrect Conclusion	Error Type
Accept a correct hypothesis	Reject a correct hypothesis	Type I
Reject an incorrect hypothesis	Accept an incorrect hypothesis	Type II

Obviously some sort of compromise is required. This depends on the relative importance of the two types of error. If it is more important to avoid rejecting a true hypothesis (type I error) a high confidence coefficient (low value of x) will be used. If it is more important to avoid accepting a false hypothesis, a low confidence coefficient may be used. An analogy with the legal profession may help to clarify the matter. Under our system of law, a man is presumed innocent of murder until proved otherwise. Now, if a jury convicts a man when he is, in fact, innocent, a type I error will have been made: the jury has rejected the null hypothesis of innocence although it is actually true. If the jury absolves the man, when he is, in fact, guilty, a type II error will have been made: the jury has accepted the null hypothesis of innocence when the man is really guilty. Most people will agree that in this case, a type I error, convicting an innocent man, is the more serious.

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Conclusion on Sampling and Testing ; Standard Error

No mention is made in these notes of considerations of costs of incorrect decisions. Statistical significance is not always the only criterion for basing action. Economic considerations of alternative actions is often just as important.

These, therefore, are the basic steps in the statistical testing procedure. The majority of tests are likely to be parametric tests. Researchers will obtain a result, say a difference between two means, calculate the standard error of the difference and then ask "How far away from the zero difference hypothesis is the difference we have found from our samples?" . To enable researchers to answer this question, they convert their actual difference into "standard errors"

The standard error (SE) is the standard deviation of the sampling distribution of a statistic, most commonly of the mean.

The standard error of the mean (SEM) is the standard deviation of the sample-mean's estimate of a population mean. SEM is usually estimated by the sample standard deviation divided by the square root of the sample size

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

! No confusion: Later we will also use SME as a simple denomination of the sample mean !

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 9

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Example calculations of sample size

Suppose a researcher wishes to measure a population with respect to the percentage of persons owning a maize shelter. He/she may have a rough idea of the likely percentage, and wishes the sample to be accurate to within 5% points and to be 95% confident of this accuracy.

Remembering the formula from the very beginning and being interested in the error term

$$T = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$$

We calculate $5\% = \frac{2\sqrt{(30)(70)}}{n}$

Then eliminating the square root by squaring both sides $25\% = \frac{4[(30)(70)]}{n}$

The simple maths solving for n will give a rounded sample of about 340 $n = \frac{4(30)(70)}{25} = \frac{8400}{25} = 336$

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 10

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Summary and Outlook at Regression Theory

Two major principles underlie all sample design: the desire to avoid bias in the selection procedure and to achieve the maximum precision for a given outlay of resources. Sampling bias arises when selection is consciously or unconsciously influenced by human choice, the sampling frame inadequately covers the target population or some sections of the population cannot be found or refuse to co-operate.

Random, or probability sampling, gives each member of the target population a known and equal probability of selection. Systematic sampling is a modification of random sampling. To arrive at a systematic sample we simply calculate the desired sampling fraction and take every nth case. We have seen several other sampling methods. We have not yet come across the important concepts of "confidence intervals" and significance test. This will come later

The test procedures at NPAL use regression and comparison of results to a reference regression to measure the contamination of samples and therefore that of the population. Regression theory relies on statistics principles which are identical or closely related to those of sample theory. Some of techniques to discover errors and / or predictions like "confidence intervals" and significance test we explain with regression theory in view. So this subject should be as useful for the non-analyst as for the chemist / analyst using the test procedures

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 11

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Introduction to Statistical Regression (1)

In the previous examples of statistical inference, we estimated the mean of a single population and we compared two population means. Finally, we might compare a number of population means. Now we ask whether we could improve the analysis if we are able to rank the populations numerically rather than in unordered categories. We can use the samples and tests to show how wheat yield depends on several different kinds of inputs (like irrigation or fertilizer). If we wish to consider how yield depends on several different amounts of fertilizer, we define fertilizer application on a numerical scale. If we plot the yield Y that follows from various fertilizer applications, a scatter plot similar lo the following figure might be observed. From this scatter plot, it seems clear that fertilizer does affect yield. Moreover, it should be possible to describe how by an equation relating Y to X. Estimating an equation is, equivalent geometrically to fitting a curve through this plot

In a study of how wheat yield depends on fertilizer, funds are available for only seven experimental observations. So the experimenter sets X at seven different values, taking only one observation Y in each case, as shown in the next slide. If you would graph these points, and roughly fit a line by eye you would come up with the next figure. Of course it is not done by hand, but by the "Trend Line" function of an EXCEL graph

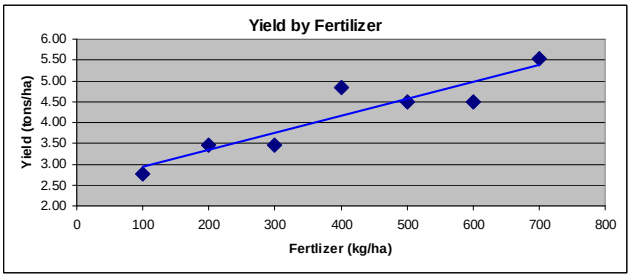
29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 12

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Introduction to Statistical Regression (2)

This is called the statistical "regression" of Y on X (with the data table net to it).



X Fertilizer (kg/ha)	Y Yield (tons/ha)
100	2.768
200	3.460
300	3.460
400	4.844
500	4.498
600	4.498
700	5.536

As a simple mathematical model, it will be useful as a brief and precise' description, or as a means of predicting the yield Y for a given amount of fertilizer X. Since yield depends on fertilizer, yield is called the "dependent variable" or "response variable" Y. Since fertilizer application is not depending on yield, but instead is determined independently by the experiment, we refer to it as an "independent variable" or "factor," or "regressor" X .

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 13

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Introduction to Statistical Regression (3)

Fitting a line

It is time to ask, more precisely, "What is a good fit?" The answer surely is, "A fit that makes the total error small." One typical error (deviation) would be the vertical distance from the observed Y, to the fitted value Ŷi on the line, that is, (Yi - Ŷi).
We note that this error is positive when the observed Yi is above the line and negative when the observed Yi is below the line.
1. As our first tentative criterion, consider a fitted line that minimizes the sum of all these errors:

Unfortunately, this works badly. The problem is one of sign; in both cases, positive errors just offset negative errors, leaving their sum equal to zero.

Reminds us of ? Correctly : Mean Absolute Deviation (MAD)

As the best way to overcome the sign problem, we choose lo minimize the sum of the squares of the errors:
This is the famous "least squares" criterion; one of its justifications: Squaring overcomes the sign problem by making all error positive and it forces the line to be as close to the points as possible.

$$\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 14

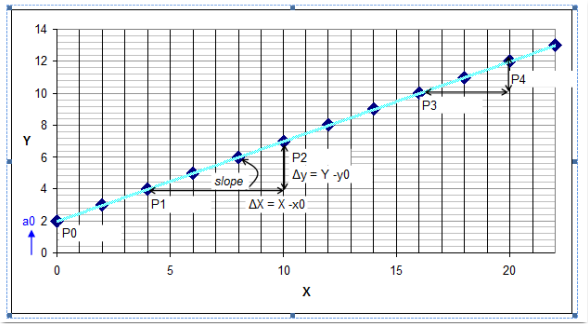
DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Introduction to Statistical Regression (4)

Lines and Planes; Elementary Geometry

The definitive characteristic of a straight line is that it continues forever in the same constant direction. We make this idea precise. In moving from one point P1 to another point P2, we denote the horizontal distance by ΔX (where Δ means change, or difference), and the vertical distance by ΔY. Then the slope30 is defined as:
slope = ΔY/ ΔX



The characteristic of a straight line is that this slope remains the same everywhere: slope = ΔY/ ΔX = b (constant). It is now very easy to derive the equation of a line, if we know its slope b and any one point on the line. Suppose that the one point we know its P0 , the Y-intercept; since its coordinates, are P0 (0,a0). In moving to any other point P(X,Y)on the line, we may write: slope = ΔY/ ΔX =b= (Y-a0)/(X-0).

29.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

Page 15

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

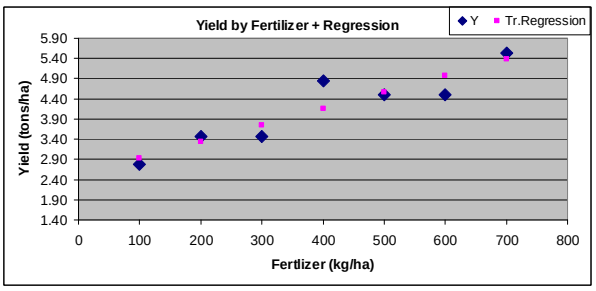
AECOM

Introduction to Statistical Regression (5)

After transformation we arrive at the equation of a line Y=a0+bX, where a0 is the intercept and b the slope.

The least squares solution

The scatter of observed X and Y values from the Table were graphed in the Figure
Our objective is to fit a line: Ŷ - a0 + bX The geometry of lines .



Without mathematical calculation: If we could apply the least square calculation, we obtain the regression equation as Ŷ = 2.521 + 0.0041*X. Not surprisingly the scatterplot of the Regression values will be the trend line if the translated regression values would be connected (Literature[1])

29.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

Page 16

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Day2: First Group Exercise:

Questions:

1. **What does a Test Hypothesis mean in the context of sampling of NPAL and other organizations. please name one?**
2. **Distinguish between the two possible errors, when to induct from the sample on the population apply for the work of NPAL and other organizations**

Exercise:

- **Please use the file "ExD02_1_Sample_Size", to calculate the size of a sample for a given probability derived from guesswork or other information (use 1% level). Or use your calculator to do a similar exercise**

Or alternatively

- **Please use file EX_D02_Source.xls. This file has been copied from the laboratory of NPAL (Metro-Manila) . You can use the file ExD02_1_RegressionEx.xlsx where these data are displayed in a line graph. Insert the sample data into the table and comment result**

Presentation:

1. **Please allow one group member to present the results, verbally, on flip chart or via computer (everything is ok)**

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 17

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Workshop program Day 2 Afternoon

Day 2/ Afternoon Session:

The Normal Distribution
The Central Limit Theorem
The Distribution of expected Mean from a Normal Population
The Distribution of expected Mean from a Non-normal Population
Confidence Intervals and t-Test
Hypothesis Testing
Hypothesis Testing Using Confidence Intervals

More on Regression theory:

Simplifying Assumptions
The Nature of the Error Term
Confidence Intervals
Example of Interval estimates
Dangers of extrapolation
Statistical Risk
Risk of an Invalid Model

Group Work participants:

Calculation of Probabilities Distributions characteristics of samples of NPAL and other organizations. Which sampling method applies to practices of the work of NPAL and other organizations and how and which method could be applied for the work of NPAL and other organizations - (if possible for the 4 different types of sampling points)

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 18

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

The Normal Distribution

For many random variables, the probability distribution is a specific bell-shaped curve, called the normal curve, or Gaussian curve. It is the most useful probability distribution in statistics. For example, errors made in measuring physical and economic phenomena often are distributed normally. In addition, many other probability distributions often can be approximated by the normal curve.

Standard Normal Distribution

A random variable Z is called standard normal if its probability distribution is:

The symbols "π" and "e" denote important mathematical constants, approximately 3.14 and 2.72 respectively.

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}$$

The normal distribution is remarkably useful because of the **central limit theorem**. In its most general form, under some conditions, it states that averages of random variables independently drawn from independent distributions converge in distribution to the normal, that is, become normally distributed when the number of random variables is sufficiently large.

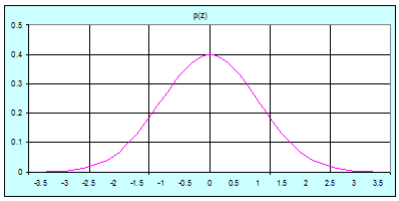
DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

The Central Limit Theorem and its implications

The assumption of the central limit theorem is extremely important because it justifies the test methods and confidence limits we will apply. The central limit theorem is not only remarkable, but very practical as well. For it completely specifics the distribution of the mean in large samples, and is therefore the key to large-sample statistical inference. In fact, in most cases when the sample size n reaches about 10 or 20, the distribution of the sample mean is already is practically normal. It can even be shown for a non-normal population; that how the distribution of the sample mean changes shape as sample size n increases. The sample mean becomes approximately normally distributed as n grows, no matter what the parent population is.

We draw the normal curve of the standard normal distribution in the following Figure to reach a maximum at z = 0. As we move to the left or right of 0, z increases in the negative exponent; therefore p(z) decreases, approaching zero in both tails. This curve also is symmetric: since z appears only in squared form, -z generates the same probability as +z.



DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

The Implications for Sampling: Confidence intervals

Since the population "gives birth" to the sample, we shall speak of the population distribution as the parent distribution. The distribution of the sample X is then called a derived distribution or a sampling distribution..

We concluded so far that $\bar{X}$ was a good estimator of μ for populations that are approximately normal. The specific sample mean, that we happen to observe is almost certainly a bit high or a bit low. Accordingly, if we want to be reasonably confident, that our inference is correct, we cannot claim that μ is precisely equal in the observed sample mean. Instead, we must construct an interval estimate or confidence interval of the form: μ = $\bar{X}$ ± a sampling error.

For convenience we will call the sample mean SME from now on

The crucial question is: How wide must this allowance for sampling error be? The answer, of course, will depend on how much the sample mean fluctuates .
First we must decide how confident we wish to be that our interval estimate is right—that it does indeed bracket μ. It is common to choose 95% confidence; in other words, we will use a technique that will give us, in the long run, a correct interval 19 times out of 20.

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 21

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

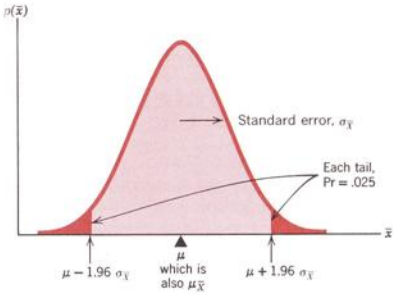
AECOM

Confidence intervals and t-Test

The Normal distribution of the sample mean around the fixed but unknown parameter μ. 95% of the probability (is contained within 1.96 % of the total area –the standard error of the sample)

The confidence interval P (μ - 1.96 σSME < SME < μ + 1.96 σSME) = 95%

Since σ is unknown, the statistician who wishes to evaluate the confidence interval (95%) must use some estimator of σ. The most obvious candidate is the sample standard deviation s (note that s, along with SME, always can be calculated from the sample data). Substituting s into the standard formula we estimate the 95% confidence interval using the generalized formula



$$T = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$$ The „initial“

Formulas for confidence intervals

The general $$\mu = \bar{X} \pm z_{.025}\frac{s}{\sqrt{n}}$$

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 22

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Confidence intervals and t-Test (2)

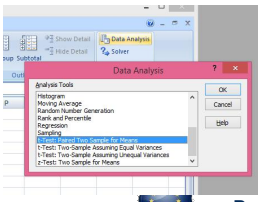
Provided that his sample is large (50 or 100), depending on the precision required), this will be an accurate enough approximation. But if the sample size is small, this substitution introduces an appreciable source of error. Therefore, if the statistician wishes to remain 95% confident, his interval estimate must be broadened. How much?
Recall that SME has a normal distribution; when σ was known, we formed the standardized normal variable, which is the transformed general formula for confidence intervals

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

By analogy we introduce "Student's t" variable

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

An important practical question is: When do we use the t distribution and when do we use the normal? If σ is known, the normal distribution is appropriate; if σ is unknown, then the t distribution is appropriate: The t-test you can conduct with EXCEL's Analytical functions form the "Data" Tab.



DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Hypothesis Testing Using Confidence Intervals

In general, any hypothesis that lies outside the confidence interval may be judged implausible or rejected. On the other hand, any hypothesis that lies within the confidence interval may be judged plausible, or acceptable. This means a confidence interval may be regarded as just the set of acceptable hypotheses.

Example:
At a large American university in, the male and female professors were sampled independently, yielding the following annual salaries (in ten-thousands of dollars, rounded):

Men (X1)	Women (X2)
20	9
14	12
17	8
14	10
15	16
12	
11	
19	
16	
22	

A husband claims that there is no difference between the salary means that is, if we denote the difference as H, he claims that: H = 0, his wife, however, claims that the difference is as large as seven thousand dollars: H=7

The short-cut calculation of the 95% confidence interval is used, and with the t-value for 95% = 2.16 :The following formula is the 95% confidence interval for two means in independent samples when population variances are equal and unknown.
So it translates to the Hypothesis: H = (SME1 - SME2) ± t.025 * SE √(1/n1 +1 /n2)
= 5.0 ± 2.16(1.87) this means= 5.0 ± ~4.0
Thus, with 95% confidence, H is estimated to be between 1 and 9. Thus the claim A = 0 seems implausible, because it falls outside this confidence interval

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Continuation of Statistical Regression (1)

Simplifying assumptions

Consider again the fertilizer-yield example in the previous chapter. Suppose that the experiment could be repeated many times at a fixed level of fertilizer x. Even though fertilizer application is fixed from experiment to experiment, we would not observe exactly the same yield each time. Instead, there would be statistical fluctuation of the Y values, clustered about a central value. There obviously would be great problems in analyzing populations peculiar and unique in their distributions and comparing them.

To keep the problem manageable, therefore, we make several assumptions about the regularity of the populations. We assume that:

1. The probability distributions p(Yi/xi) have the same variance σ2 for all xi.
2. The means E (Yi) lie on a straight line, known as the true (population) regression line: The population parameter α and β specify the line; they are to be estimated from sample information.
3. The random variables Yi are statistically independent.(i.e. is Y2. is "unaffected" by Y1)

It is useful to describe the deviation of Yi from its expected value or disturbance term ei so that the model alternatively may be written as: Obviously α and β correspond to intercept a and slope b from our sample

$$Y_i = \alpha + \beta x_i + e_i$$

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Regression: Confidence intervals and hypothesis tests for β

The Gauss-Markov Theorem (abbreviated)

The major justification for using the least squares method to estimate a linear regression is the following:

Gauss-Markov Theorem

The least squares estimator b has minimum variance (is most efficient)estimator of β, .and similarly a is the minimum variance estimator of α.

This theorem is important because it requires no assumption about the shape of the distribution of the error term. No proof will be given here, please refer to the Literature [1],[2],[3]

The distribution of b

Now we ask about the shape of the distribution of b. Let us add (for the first time) the strong assumption that the Yi are normal. Since b is a linear combination of the Yi, it follows that b also will be normal. But even without assuming that the Yi are normal, we know that, as sample size increases, the distribution of b usually will approach normality. This can be justified by a generalized form of the central limit theorem. Our objective is to develop a clear intuitive picture of how this estimator varies from sample to sample.

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Confidence intervals for a and b (and so for α and β)

We can derive the 95% confidence interval for b easily, arriving at a result familiar from previous Interval estimation

$$\beta = b \pm t_{.025} s_b$$

Using a similar argument for the intercept, we could easily derive

$$\alpha = a \pm t_{.025} s / \sqrt{n}$$

This is quite an amount of formulas at a time, but the estimates of predictions for regression functions and confidence intervals is elementary for understanding and applying the regression approach in statistics. We omit the formula to find the interval estimate for Yo,(you can refer to Literature [1]) but we will retain, that the combination of the above formulas (together with combinations of sampled values of X) will allow to calculate confidence intervals for the regression values Yi

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 27

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Example of Interval estimates for a Regression

In the previous section, we considered the broad aspects of the model namely, the position of the whole line, remember there were several assumed populations (determined by α and β). In this section, we shall consider two narrower problems:

(a) For a given value xo, what is the interval that will predict the corresponding mean value of Yo For example, in our fertilizer problem, we may want an interval estimate of the mean yield resulting from the application of 550 kg of fertilizer.

(b) What is the interval that will predict a single observed value of Yo (referred to as the prediction interval for an individual Yo). Again using our fertilizer example, what would we predict a single yield to be from an application of 550 kg of fertilizer? This individual value clearly is less predictable than the mean value' in (a). We now consider both in detail but before that we state:

$$E(Y_0) = E(a) + x_0 E(b)$$

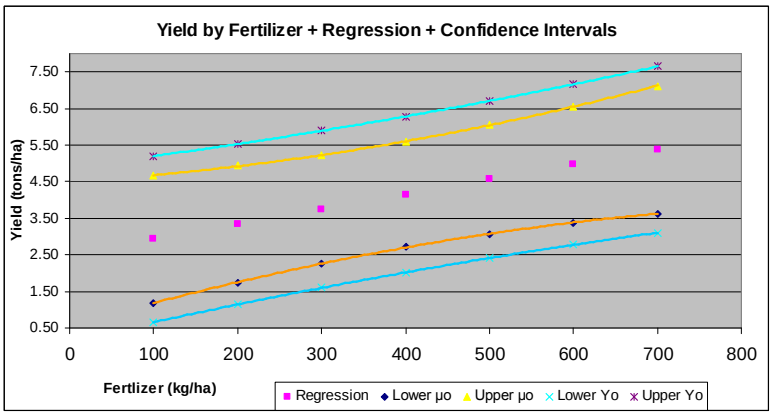
29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 28

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Example of Interval estimates calculated

We will not go into all the formulas, for this you can refer to Literature [1],[2].Instead will present a calculated example to understand the character of confidence intervals in regression theory



29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 29

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Example of Interval estimates calculated cont'd

The above mentioned and calculated example [1] was for one observation xo = 550. The relationship of prediction and confidence intervals is shown in the graph above. The combined sources of error in a confidence intervals for the mean are shown in yellow, the wider, blue bands gives the prediction intervals for individual Y observations. Note how both bands expand as xo moves farther away from its central value (the mean) this reflects the fact that xo2 appears in both variances. This example gives reason for some general remarks about predictions based on statistical regression.

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 30

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Dangers of extrapolation

We emphasize that xo may be any value of x. If xo lies among the observed values x1 . . . xn, the process is called interpolation. If x0 is out beyond the observed values xl . . . xn, then the process is called extrapolation. The techniques developed in previous section may be used for extrapolation, but only with great caution, as we shall see.
There is no sharp division between safe interpolation and dangerous extrapolation. Rather, there is continually increasing danger of misinterpretation as x0 gets further and further from its central value.
This is true for the exercises of pesticides where observed values lie within the reference regression

Statistical Risk

We emphasized in the previous section that prediction intervals get larger as x0 moves away from the centre. This is true, even if all the assumptions underlying our mathematical model hold exactly.

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Dangers of extrapolation (2)

Risk of Invalid Model

In practice, we must recognize that a mathematical model is never absolutely correct. Rather, it is a useful approximation. In particular, we cannot take seriously the assumption that the population means are strung out in an exactly straight line. If we consider the fertilizer example, it is likely that the true relation increases initially, but then bends down eventually as a "burning point" is approached, and the crop is overdosed. In the region of interest, from 0 to 700 kg, the relation is practically a straight line, and no great harm is done in assuming the linear model. However, if the linear model is extrapolated far beyond this region of experimentation, the result becomes meaningless. In such cases, a nonlinear model should be considered (Literature [1], [2])

Statistical Risk

We emphasized in the previous section that prediction intervals get larger as x0 moves away from the centre. This is true, even if all the assumptions underlying our mathematical model hold exactly.

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Concluding observations

Two points warrant emphasis. First, most of the theory of this section and, in particular, the Gauss-Markov justification of least squares requires no assumption of normality of the error term. The one exception occurs when the normality assumption was required only for small sample estimation and this because of a quite general principle that small sample estimation requires a normally distributed parent population to strictly validate the t distribution. But even here, t is often a reasonably good approximation in non-normal populations.

Second, we have assumed that the independent variable x has taken on a given set of fixed values (for example, fertilizer application was set at certain specified levels). But in many cases, x cannot be controlled in this way. For example we are examining the effect of rainfall, we must recognize that x (rainfall) is a random variable that is completely outside our control. The surprising thing is that most the findings of this section remains valid whether x is fixed or a random variable, provided that we assume that:
σ2 (and α and β) are independent of x, and the error term e is statistically independent of x
This greatly generalizes the application of the regression model
(Literature [1], [2])

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 33

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Day2: Second Group Exercise (1):

Exercise:

- Group Work on what Confidence Intervals and t-Test could signify in the context of sampling of NPAL.
- Use File ExD02_t-Test1.xlsx and test calculation of an assumption of equal means as explained in the presentation
- Please take a typical sample of your working area if you can. We have samples from the test laboratory . And apply the exercise if you use EXCEL, but also simple calculation would be possible

Or alternatively

- Please use file EX_D02_Source2.xls. This file has been copied from the laboratory of NPAL (Metro-Manila) . Try any of the data from the various Folders and apply to the Regression exercise in file ExD02_1_RegressionEx.xlsx where these data are displayed in a line graph. Insert the sample data into the table and comment again the results.

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 34

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Day2: Second Group Exercise (2):

Questions:

Group Work on which test would apply for confidence intervals nd t-test in your working area. What are your observations on sample size and normality assumption of your sampling methods - (if possible for the 4 different types of sampling points). Second assessment of NPAL and other staff members: What are our needs? What do we want to improve?

Or alternatively

- What did you find out about Regression theory, that you did not know? Where can there be any help in analyzing your samples and comparing them to reference regressions? How about confidence intervals ? What would they mean for acceptance / rejection of samples ?

Presentation:

Please allow one group member to present the results, verbally, on flip chart or via computer (everything is ok)

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 35

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Literature

[1] Klaus Röder : Handbook Introduction to Statistics: http://www.klaus-roeder.com/Ordner/PDFs/Projects/13WoD/13WoD_Handbook_WoD_and_Statistics_130208.pdf

[2] Introductory Statistics, 5th Edition 5th Edition, by Thomas H. Wonnacott (Author), Ronald J. Wonnacott (Author); ISBN-13: 978-0471615187

[3] Crawford, I. M. (1990), Marketing Research, Centre and Network for Agricultural Marketing Training in Eastern and Southern Africa, Harare, pp 36-48.

[4] FAO Sampling recommendations e.g.
http://www.fao.org/docrep/012/i1379e/i1379e05.pdf
http://www.fao.org/docrep/w3241e/w3241e08.htm

29.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 36

Thanks for your patience and cooperation