



# **MISSION REPORT**

## **Activity 5.4.b**

Upgraded NPAL's Chemist and Inspectors Knowledge on development of statistical design on sampling used for Field and Market Monitoring for fruits and vegetable.

DCI-ASIE/2012/023-158

Date of Report: 08 December 2015  
Reporting Period: 24 November to 04 December 2015  
Author of Report: Klaus Röder, Short-Term Expert



*With Crown Agents, DIN and GFA*

**Disclaimer:** The views expressed in this report do not necessarily reflect the views of the European Commission

## TABLE OF CONTENTS

<b>1.</b>	<b>ACRONYMS.....</b>	<b>4</b>
<b>2.</b>	<b>EXECUTIVE SUMMARY.....</b>	<b>5</b>
<b>3.</b>	<b>BACKGROUND AND CONTEXT.....</b>	<b>7</b>
	Introduction .....	7
	Presentation of the mission .....	7
	Expected results and target outputs .....	8
	Organisation of the report.....	8
<b>4.</b>	<b>PURPOSE OF THE MISSION .....</b>	<b>9</b>
<b>5.</b>	<b>DESCRIPTION OF THE MISSION ACTIVITIES .....</b>	<b>10</b>
<b>6.</b>	<b>IMPACT, FINDINGS AND RECOMMENDATIONS .....</b>	<b>12</b>
<b>7.</b>	<b>ANNEXES.....</b>	<b>16</b>
	Annex 1 - TORs .....	16
	Annex 2 – (i.e., seminar/training/workshop materials : ppt. presentation, etc.) .....	20

## 1. ACRONYMS

AECOM	AECOM International Development Europe
BAPTC	Benguet Agropinoy Trading Centre
BPI	Bureau of Plant Industry, Philippines
DA	Department of Agriculture
DTI	Department of Trade and Industry, Philippines
EC	Commission of the European Communities
IPPC	International Plant Protection Convention
ISPM	International Standard Phytosanitary Measures (provided by the IPPC)
NAIA	Ninoy Aquino International Airport (Manila)
NPAL	National Pesticide Analytical Laboratory
PC	Phytosanitary Certificate
PFA	Pest Free Area
PRA	Pest Risk Analysis
RP	Republic of the Philippines
SPAL	Satellite Pesticide Analytical Laboratory
SPS	Sanitary and Phytosanitary
STE	Short-Term Expert
TRTA	EU-Philippines Trade Related Technical Assistance Project
WTO	World Trade Organisation

## 2. EXECUTIVE SUMMARY

The purpose of the workshop is to provide BPI technicians and analysts and other technicians of the Philippines Department of Agriculture (DA), instruction on how to apply statistical sampling methodologies in their work. This will ensure that sampling methodologies used in the country are consistent.

One workshop was provided in this TRTA Activity, in Baguio City. The workshop had 30 delegates representing mainly the different NPAL and SPAL staff members together with experts from the contaminant laboratory of the Department of Agriculture and members of the Plant Quarantine Services - Department of Agriculture to control the sea ports in the Philippines in view of possible occurrence pesticides in imported fruit and vegetable.

The delegates of this workshop were provided with background information on the importance of their work and how using correct sampling methodologies can both protect public health and facilitate export trade.

The workshop delegates were introduced to statistical reasoning, the overarching and maybe the most important subject, to understand the difference and the relationship between a sample and the population.

The workshop had four sections:

1. An introduction to statistical basics and denominations
2. An introduction to statistic sampling and to various sampling strategies
3. An outlook on testing on samples together with the concepts of statistical error and confidence intervals
4. A brief outline of regression theory

Due to the restricted time frame of two and a half days these subjects could not be dealt with in theoretical depth but with sufficient relation to practical obligations of the delegates.

The lectures have been supported with about 80 PowerPoint slides in two presentations and further EXCEL files prepared for Exercises and samples of real data from NPAL and AMAS.

All of the lectures were followed by practical group work. The participant had formed 6 groups: NPAL and the 4 SPALs (Baguio, Cagayan de Oro, Cebu and Davao) and another group by the contaminants section. The delegates of the Philippine Plant Quarantine Services joined SPAL-Davao for their group work

The group work (4 times during the WS) consisted of two tasks each: Discussion in groups of the importance and relevance of the subjects treated in the lectures and a practical exercise in EXCEL or done with the calculating functions of a Smartphone related with practical background of the Laboratories to the mentioned subjects. (e.g.: ->). Exercises have been done mostly on several Laptops with EXCEL but also Smartphones were used.

At the end of the two lecture days a spokesperson of each group presented the results of the reasoning and the exercises. These presentations were done verbally or

### Questions:

What Mean, Mode and Median would signify in the context of sampling of NPAL and other organizations.

Why and how are these concepts important and apply for the work of NPAL and other organizations

### Exercise:

Please use the file "ExD01\_1\_Mango Yields Region IX.xlsx", a list of farmers from Region IX with supposed mango yields

Please calculate Mean, Variance and Standard Deviation

Try also Kurtosis and Skewness

What would you say about the data. Are these yields ?

**All Exercises based on data collected at NPAL or the Department of Agriculture**

via computer using the presentation device which was also used for the lectures.

The lecturer presented the delegates with two questionnaires, one at the beginning, and one at the end. The reason was to test for improvement of statistical knowledge through the Workshop. (Section 6 : Impact, Findings and Recommendations)

Through the workshop discussions, a few areas were identified (Section 6: Impact, Findings and Recommendations) that require follow-up by the organizers. The main issues are:

- Adaptation of the sampling plans by using complete sample frames for sampling fields and markets, if possible. .
- Application of tests for verification and assessment of samples after sample results.
- Harmonization of sampling schemes through the regions and for all the SPALs, other organizations to be included if feasible

The scope of work of the delegates was rather inhomogeneous: some sample in the fields, other markets, most are responsible for analysing samples in the analytic laboratories established for this purpose. So a simple on the job training: how to sample in an explicit environment would not respond to the necessities of a majority of participants not immediately confronted with this type of work. However, as the TOR stated: The workshop was intended to “provide the workshop participants with the statistical tools for designing a sampling plan for monitoring fruits and vegetables”.

It was thus intended to make participants aware of the principal basics of statistical sampling design, execution and testing. It was felt that a practical workshop for specific groups of delegates: e.g. sampling units on markets, packing houses or ports would benefit some participants in more practical, on-the-job aspects in addition to this workshop.

The results of the statistical assessment showed an increase in knowledge of participants due to the workshop; some practical issues would need some intensified training as could also be proved by the statistical assessment.

### 3. BACKGROUND AND CONTEXT

The National Pesticide Analytical Laboratory Quezon City (NPAL) conducts sampling and analysis of fruits and vegetable products at the 4 different types of sampling points: fields, markets, packing facilities and export facilities (ports), (the latter are tested by PQS) to detect contamination of these products. The detection of contamination by the 5 laboratories of NPAL and SPALs then prevents further use of these products either for consumption or export. So far the sampling of the mentioned products has been done without much of a statistical design and this should be improved to respect scientific design criteria and internationally acclaimed methodology.

#### INTRODUCTION

This report covers the period from 24th of November until 6th of December of return date of the STE from Manila.

As no background info for this mission was given to the STE beforehand, all information on practical work of the work of NPAL and other institutions on sampling and further activities had to be acquired after arrival at Manila.

The support by the NPAL staff was of extraordinary help for the preparation but it was clear even before the mission, that a sampling workshop for practitioners can only be useful if examples and material can be adapted to the reality of the participants. All of practical material for the workshop had to be prepared just before the workshop. However, practical data sets, if not abundant, but existent were handed over by NPAL staff and the Department of Agriculture and were used in examples in the workshop.

#### PRESENTATION OF THE MISSION

The number of days assigned to the expert was 10 working days, with the assignment commencing on the 24<sup>th</sup> of November 2015 and ending on the 6<sup>th</sup> of December 2015. The mission comprised preparations of the workshop, elaboration of workshop material and preparations, in liaison with the beneficiary organization (BPI), and delivery of one workshop in Baguio.

The Short-Term Expert has been engaged for the period and dates indicated on the following Timetable for the activities given by the author:

Activities on days	Tasks	Nov./Dec. 2015 WD per activity
Nov. 24	Briefing at BOI (Board of Investment) with Florian Albuero (Teamleader TRTA3), Key Expert Ian Watson, personnel from NPAL and FDA to get overview over program scope and purpose. Preparing exercises and evaluation and assessment documents	1
Nov. 25	Discussion with beneficiaries/mentoring. Meeting the workshop participants and data producers at NPAL- Manila. Meeting at NPAL and then proceeding with visits to the supermarkets and trading post in Manila. Specially important the discussion with data producers at NPAL to learn about documentary and possible computer based exercises at Workshop	4
Nov. 26-27	Visiting markets in Region 4 and farm and packing houses in Region 3. This will be conducted to see the actual conditions in the places where NPAL collect their samples which will give an idea of the acceptable sampling approach that you will be taught later.	

Activities on days	Tasks	Nov./Dec. 2015 WD per activity
Nov. 28-29	Compiling the final WS documents and fixing final programme schedule. Preparing exercises and evaluation and assessment documents	
Nov. 30	Travelling to Baguio and finalizing Workshop documents, photocopying and assuring final preparations on site	
Dec. 01-03	Workshop and Assessment	3
Dec. 04	Debriefing, Report writing. Analyzing WS assessment and results of exercises delivered by participants.	1
	Analyzing and numerically assessing (statistical analysis) of WS results and extended post WS questionnaire. Analysis of participant's WS results and contributions. In depth content analysis of recommendations for continuation and design of further program considering sampling in the NPAL context.	2
	Total WD	10

### EXPECTED RESULTS AND TARGET OUTPUTS

The TOR state: "For the purpose of the present activity it is expected the STE will provide the workshop participants with the statistical tools for designing a sampling plan for monitoring fruits and vegetables ". For this a number of sample designs should be dealt with in the workshop:

- *Simple random sampling,*
- *Stratified random sampling,*
- *Systematic random sampling,*
- *Cluster random sampling,*

All this was dealt with in the workshop, however the subject of statistical sampling was extended to sampling comparison by tests and statistical characteristics of regression as this touched more of the area of work of analyst, representing the majority of participants. It should be stated that NPAL and SPALS use a mixture of Quota sampling and Random sampling and Area sampling, also introduced in the workshop, seems to be a more appropriate method for many of the applications of the delegates.

### ORGANISATION OF THE REPORT

The outline of the Activity 5.4b report is as follows:

Executive Summary  
Background and Context  
    Introduction  
    Presentation of the mission  
    Expected results and target outputs  
Purpose of the Mission  
Description of the Mission Activities  
Findings and Recommendations  
Annexes

#### **4. PURPOSE OF THE MISSION**

The overall objective of the TRTA project is to contribute to the Philippines' integration into the international and regional trading and investment system, thereby strengthening economic development, inclusive of growth and poverty reduction.

The specific purpose of the Activity 5.4b is to Upgrade NPAL's Chemist and Inspectors Knowledge on development of statistical design on sampling used for Field and Market Monitoring for fruits and vegetable and so to enhance the capacity of the BPI workshop participants in sampling and by this improving quality for testing food and food products.

## 5. DESCRIPTION OF THE MISSION ACTIVITIES

### Time list of Activities during the mission

The Time list is in comprehensive tabular form. Detailed lecture subjects and exercise contents can be found under course documents activities are described in the ANNEX 2 - WORKSHOP MATERIALS

Day	Location	Activity
Nov. 24	Metro Manila	Briefing on Project and Preparation of Workshop at BOI with Ian Watson, Marilyn M. Pagayunan - FDA Phil., Albina M. Mendoza PDA Phil., German T. Yatco, BPI-Passo, Florian A. Alburo Team Leader Technical Assistance Team TRT3 and Gerry Jatco - Analyst at NPAL.
Nov. 25	Metro Manila	Visits to the markets in Metro-Manila : <ul style="list-style-type: none"> <li>• Supermarket in SM Mall Quezon City</li> <li>• Balintawak Wholesale Market</li> <li>• Pure Gold Supermarket Balintawak</li> <li>• Nepa Q Retail Market Quezon City</li> </ul>
Nov. 26	Metro Manila, Central Luzon	Visits outside Metro-Manila : <ul style="list-style-type: none"> <li>• Tarlac in Central Luzon , Visit of 2 markets, 1 Packing factory and 1 Okra field:</li> <li>• Tarlac vegetable and retail market in Tarlac</li> <li>• Jelfarm Fresh Produce Enterprise in San Manuel</li> <li>• Okra Field delivering regularly to Fresh Produce Enterprise in San Pascual (nearby San Manuel)</li> <li>• Pulilan vegetable and retail market in Pulilan</li> </ul>
Nov. 27	Metro Manila	Briefing on Analysis of samples at laboratories at NPAL on the technical explanation of the procedures how work is done at the analytical laboratories and working situation at the laboratories at NPAL (Metro-Manila). Visiting the extraction (physical preparation of analytical specimen in liquid form) and the technical calculation laboratories (where the further separated sub-samples are finally analyzed on specialized machines and test results are generated as computer outputs). Computer generated result are then subsumed in EXCEL tables and graphs. The explanation of how the samples are analyzed and technical procedures to detect and eventually reject a contaminated sample was explained. In the afternoon discussion on the quality assumptions and characteristics of the sampling at the ports of imported vegetables and fruits. This sampling is not done by NPAL but by the Quarantine Department. In the late afternoon a visit to the Department of Agriculture and the Agribusiness and Marketing Assistance Service (AMAS), receiving an incomplete list of the farming universe in the Philippines and the regions.
Nov. 28/ 29.	Metro Manila	Preparing Workshop material, ,handouts, exercises
Nov. 30	Metro Manila / Baguio	Travelling form Manila to Baguio
Dec.01	Baguio	Workshop : Day 1

Day	Location	Activity
		Opening of Workshop Lectures and Group Work With presence of Renea Cruz-Tan Knowledge Management Officer of TRTA3 and assistant
Dec.02	Baguio	Workshop : Day 2 Lectures and Group Work
Dec.03	Baguio / Metro Manila	Workshop : Day 3 Synopsis, Assessment and Closing Travelling form Baguio to Manila
Dec.04	Metro Manila	Debriefing on Project and Workshop at NPAL with Ian Watson, Maria Araceli S.Escobar Chief International Affairs Division DA, Florian A. Alburo Team Leader Technical Assistance Team TRT3, , and Maria Lourdes de Mato, Division Chief NPAL and assistants.
Dec.05	Metro Manila	Preparing report and recommendation for pursuit of sampling strategy and training at NPAL

## **6. IMPACT, FINDINGS AND RECOMMENDATIONS**

### **Findings about NPAL and sampling procedures**

NPAL owns an impressive laboratory for processing and further testing of plant samples. Visits to various sampling points showed a knowledgeable approach to sampling and procedures. However, the selection of markets and fields, likewise of packing facilities and probably also to containers arriving for import seemed to be guided by administrative necessities instead by scientific statistical sampling methods. Likewise is the work in the laboratories guided by experience and on the job training, the theoretical knowledge of the used regression technique for sample processing, testing and comparison seem to be limited. NPAL was, however, very transparent and frank to disclose the used methodology and techniques. The same is true for the SPAL laboratory of Baguio where the local head of SPAL Joy S. Calaunan was extremely helpful to inform about their work and procedures.

### **Baguio Workshop Findings**

The Standard operation procedures of random sampling of The National Pesticide Analytical Laboratory Quezon City (NPAL) to conduct sampling and analysis of fruits and vegetable products at the 4 different types of sampling points: fields, markets, packing facilities and export facilities (ports), (the latter are tested by PQS) to detect contamination of these products has been seen to be of good quality. These sampling methods are claimed to be of pure random characteristics but at least in the area of markets and fields they are not. Big markets as fields are administratively attributed and no sampling frame for stratification and strata selection is applied. This is a combination of quota sampling with random sampling at the last stage. The information of sampling techniques of PQS could only be derived from explanation of the present junior technicians of PQS.

For the sampling of packing facilities and export facilities (ports), only the following two recommendations will apply because the statistical conditions for sampling are easier and less prone to sampling errors

Recommendations:

- Adaptation of the sampling plans by using complete sample frames for sampling fields and markets, if possible. Feasibility study on costs and limits of these activities (regional characteristics might not allow sampling according to sampling frames – mountainous, islands, remote access). Access to sampling frames and the Philippine Statistics Authority should be contacted for this purpose.

Test procedures of samples (is the test acceptable- are we confident that two tests show the same results) are inadequately developed and left to estimates

- Application of tests for verification and assessment of samples after sample results due to standard operating procedure lead by NPAL

Different regions in the Philippines have very different characteristics. Different plants are grown and access facilities to markets and fields. However, sampling characteristics in many cases (like using sampling plans or random stratification) do not differ between regions. Regular communication between the SPALs and NPAL should allow for better harmonization of this

- Harmonization of sampling schemes through the regions and for all the SPALs, other organizations to be included if feasible. Distant learning facilities should be considered.
- Application of on the job training courses for sample collecting technicians to apply true sampling methodology in the working area, This presupposes sample frames for the corresponding area. It further will requires a feasibility analysis and comparison of cost related restrictions due to enlarged sampling through the use of sampling frames.

## General Findings of Learning Progress

The above mentioned questionnaire (pre- and post) has been assessed initially, More should be as done as it shows (from the questionnaire) that **elementary concepts of statistics, induction from sample to the population, error concept and testing of samples, confidence intervals in sample estimates need still to be improved**. This was possible after a more elaborate evaluation of the two questionnaires.

The questionnaires were quickly to be completed using only tick marks. The second questionnaire "Questions to Upgraded NPAL(Post-Course)\_Sol\_KR151130.docx" together with correct solutions marked in color, the questions 1, 3-9 had been asked before in the first questionnaire (all documents mentioned are to be found on ANNEX 2 - WORKSHOP MATERIALS

A brief interpretation of the results in Box 6-1:

The maximum points in Questionnaire 1(pre) was 16, 26 in the second. The two means of the relative frequencies were compared and a t-test should indicate if the Hypothesis ( $H_0$ ) that no change of knowledge can be detected between before and after the course. This hypothesis can be refused with a probability of more than 99.999 %, meaning there has been an improvement of knowledge (increase of the mean). The appropriate p-value has been marked. The results were calculated with a EXCEL Add-on (STATIXL), which was also briefly introduced to the delegates.

## Specific Findings of Learning Progress

A brief interpretation of the results in Box 6-2:

As seen from the Questionnaire in ANNEX 2 - WORKSHOP MATERIALS, Question 8 in Questionnaire1 (pre), Question 8 corresponds to Question 9 in Questionnaire2 (post). "Confidence intervals are used to "It is easy to compare the results from both tests: The percentage of correct answers was quite high but has not increased and stayed the same. Despite the fact that this subject was treated extensively during the course. The assumption of a lack of understanding the basics of statistics is further enhanced by the results of question10: "Which of the following examples are randomized samples?" in Questionnaire2 (post). Only 41% gave a correct answer. A rather positive outcome is the result of question14 in Questionnaire2 (post): "If we want to interpret statistical results, which of the following remarks are correct ? " has a correct answer rate of 86%, which may allow to summarize briefly."

The general knowledge of statistics among the participants was rather high and in general knowledge has been increased, but the awareness of basic theoretical statistical concepts should be improved. A majority of participants is aware, that "statistical results should improve the reasoning for more profound research in order to improve Pesticide control in general"

Sample 1 (before)			Sample 2 ( after)		
8 Questions(max =16)			16 14 Questions (max=26)		
Sample values			Sample values		
Frequency	rel.Frequency Sampl1		Frequency	rel.Frequency Sampl2	
1	7	0,44	10	0,38	
2	6	0,38	15	0,58	
3	5	0,31	11	0,42	
4	7	0,44	13	0,50	
5	8	0,50	14	0,54	
6	6	0,38	16	0,62	

..Shortened to save space..

27	5	0,31	9	0,33	
28	6	0,38	12	0,46	
29	6	0,38	13	0,50	
Descriptive Statistics for: Input Range = Sheet1!\$C\$5:\$C\$34			Descriptive Statistics for: Input Range = Sheet1!\$E\$5:\$E\$34		
rel.Frequency Sampl1			rel.Frequency Sampl2		
Mean	0,373		Mean	0,500	
Std Error	0,014		Std Error	0,017	
Std Dev.	0,076		Std Dev.	0,091	
Variance	0,006		Variance	0,008	
Coeff. Var.	20,276		Coeff. Var.	18,157	
Range	0,250		Range	0,385	
Count	29,000		Count	29,000	
Skewness	-0,320		Skewness	0,560	
Kurtosis	-0,814		Kurtosis	0,020	
Test Result for Datasets:					
Set 1 Range = Sheet1!\$C\$5:\$C\$34					
Set 2 Range = Sheet1!\$E\$5:\$E\$34					
Descriptive Statistics					
Variable	Mean	Std Dev.	Std Err	Lower 95% CL	Upper 95% CL N
rel.Frequency	0,373	0,076	0,014	0,344	0,402 29
rel.Frequency	0,500	0,091	0,017	0,465	0,535 29
Two-tailed t-Test					
Ho: Diff	Mean Diff.	SE Diff.	T	DF	P
0,000	-0,127	0,022	-5,796	56,000	0,000
Test for Equality of Variances					
Variable	Variance	F	DF 1	DF 2	P
rel.Frequency	0,006	1,442	28	28	0,169
rel.Frequency	0,008				

Box : 6-1

	Sample 1 Q8	Sample 2 Q9	Sample 2 Q10	Sample 2 Q14
1	0	1	0	1
2	1	0	0	1
3	1	1	0	0
4	1	1	1	1
5	0	1	0	1

..Shortened to save space..

25	1	1	0	1
26	1	1	0	1
27	1	1	0	1
28	1	1	0	1
29	1	0	0	1
	76%	76%	41%	86%

Box. : 6-2

ANNEX I - TORS



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



**TERMS OF REFERENCE**

**Activity 5.4.b Upgraded NPAL's Chemist and Inspectors  
Knowledge on development of statistical design on sampling  
used for Field and Market Monitoring for fruits and vegetable.**

DCI-ASIE/2012/023-158

Implementing Agency: Bureau of Plant Industry (BPI)/Department of  
Agriculture (DA)

Type of Expert: Short-Term Expert

Title of Expert: Short Term Expert on IPPC Standards

**General Information**

The Trade Related Technical Assistance Project 3 (TRTA 3) is a development cooperation project by and between the Republic of the Philippines (RP) and the Commission of the European Communities (EC). The implementation arrangements in this cooperation, including the role of the Department of Trade and Industry (DTI) as requesting authority and coordinating agency for the project, are prescribed under Financing Agreement DCI-ASIE/2012/023-158.

The overall objective is to contribute to the Philippines' integration into the international and regional trading and investment system, thereby strengthening economic development, inclusive growth and poverty reduction. The specific objective is to enhance the capacity of selected government agencies and private sector organisations to facilitate further integration into the international and regional trading and investment system.

It has six components, as follows:

1. Component 1: Trade Policy Development – to strengthen the capacity of the Government of the Philippines (GoP) to successfully implement the Philippine International Trade Strategy through a series of capacity-building initiatives including developing research and analytical tools; undertaking research (involving academe, government and industry); enhancing research networks and integrating industrial policy research with trade policy research;
2. Component 2: Competition Policy Development – to enhance capacities to effectively implement competition policy through studies on competition in selected sectors, provision of advice on drafting implementing legislation on competition policy, inter-agency coordination for competition procedures, among others;
3. Component 3: National Quality Infrastructure – to support consultations with public and private sector, including consumers, on content of a National Quality Infrastructure strategy and policies;
4. Component 4: SPS Conformity – to enhance capacities to develop and implement sanitary and phytosanitary (SPS) management and control

systems in-line with international standards and of producers to comply with these standards;

5. Component 5: Trade Facilitation – to strengthen capacities to comply with the Revised Kyoto Convention and to upgrade National Single Window and Warehousing operations in the Bureau of Customs;
6. Component 6: Rapid Response Facility – to provide a rapid response to unforeseen technical assistance needs from the public sector and business associations.

The project has an operational implementation period of 45 months – from 3 December 2012 to 3 September 2016 – and an estimated cost of EUR 8,975,000, of which EUR 8,000,000 and EUR 975,000 are to be contributed by EC and RP, respectively. It is to be implemented through “decentralized management” under which DTI will be the contracting and paying authority for certain expenditures including for training, seminars, workshops and operating costs.

For technical assistance, EC remains as the contracting and paying authority and for which ACE has been contracted as Service Provider. The expected period of execution of the EC-ACE service contract is 45 months and shall not in any case go beyond 3 September 2016.

## **Justification and Objective**

### **Overall objective**

The overall objective of the project is to contribute to the Philippines’ integration into the international and regional trading and investment system, thereby strengthening economic development, inclusive of growth and poverty reduction.

### **Specific objectives**

The specific objective for Component 4: SPS Conformity is to enhance capacities to develop and implement sanitary and phytosanitary (SPS) management and control systems in-line with international standards and procedures to comply with these standards.

### **Justification**

Sample survey methodology is used to gather information about a "population" by selecting and measuring a "sample" from that population. The sample is a fraction of the population studied. The population could consist of sampling units, which could be people, animals, plants, foods, etc. In the present activity the samples are fruits and vegetables from different extraction points.

Furthermore, the sample should be chosen on a statistically viable basis, such that each sampling unit has a measurable probability of being selected. Characteristics of the sampled items are measured and used to extrapolate to and represent the population. Standardized procedures are used to collect the information. Because sampling units have varying characteristics, scientific sample designs are used to reduce the risk of a distorted view of the population. The size of the sample depends on the purpose of the study, the variability among the sampling units, the desired precision, and the sample design. The sample designs includes the following type of sampling:

- *Simple random sampling*, which gives all units in the population an equal chance of being selected and randomly selects a subset of these units;
- *Stratified random sampling*, which divides the population into homogeneous subgroups and then takes a simple random sample from each subgroup;

- *Systematic random sampling*, which numbers the units in the population, selects a unit at random, and then systematically samples every  $n^{\text{th}}$  unit;
- *Cluster random sampling*, which divides the population into clusters, randomly selects a subset of these clusters, and measures all units in the selected clusters; and
- *Multistage sampling*, which combines any of these four sampling methods.

For the purpose of the present activity it is expected the STE will provide the workshop participants with the statistical tools for designing a sampling plan for monitoring fruits and vegetables

### **Scope of Work**

The activity is focused to the BPI / NPAL chemists and official inspectors. The development of the activity will be centred in a workshop on the development of statistical design on the collection of fruits and vegetables in the market and field for BPI-NPAL and processed products in market shelves as to improve the current monitoring system.

#### **Workshop**

The workshop will be pointed to the professional and official staff of BPI and FDA and their related agencies.

The initial programme is considering 30 total participants a three (3) days long course.

The course content will be in agreement with FAO Sampling for marketing research statistical procedures

### **Expected results**

- Enhanced criteria on the selection of vendors/farm where samples will be collected. This will be provided through final evaluation and case studies though is the main target of the workshop. Other system to achieve the output will be indicated in advance.

### **Expert Profile**

#### **Qualifications and skills**

- A degree in statistics or relevant background.
- Experience in market sampling is exclusively required.
- Excellent command of the English language verbally and in writing (knowledge of the Filipino language would be an advantage).
- Proven ability to work and interact with high level officials especially in a multi-disciplinary and multi-cultural setting
- Proficient in Microsoft Office (Word, Excel, PowerPoint) and relevant internet and email software

#### **General and specific**

- Demonstrable minimum 10 years' experience in market sampling.
- Experience in the region is an advantage.
- Demonstrable experience in development EU projects or with international donors would be an advantage

### **Timing, Logistics and other Arrangements**

The selected expert shall be engaged for the period and dates indicated on the following Timetable for the activities:

Activities	Tasks and sub activities	Four months 2015			Total WD per activity
		1	2	3	
Activity 5.4b Upgraded NPAL's Chemist and Inspectors Knowledge on development of statistical design on sampling used for Field and Market Monitoring for fruits and vegetable.					
	Briefing		.5		.5
	Discussion with beneficiaries/mentoring		5		5
	Workshop & preparation		4		4
	Debriefing		.5		.5
Total WD			10		10

Should travel outside Metro Manila be required, the cost shall be charged against the budget for Incidental Expenditures of the EC-ACE contract. The per diem or daily subsistence allowance covering food, lodging and local transport shall be in accordance with the approved and published rate in the EC website. **The expert shall bear the cost of travel within Metro Manila to be incurred in the performance of the duties and responsibilities spelled out in this Terms of Reference.**

A briefing meeting shall be convened by DTI on the first day of the engagement to discuss and ensure common understanding of the objective, tasks, expected outputs, and working arrangements set forth in this Terms of Reference. DTI shall notify the EU Delegation, and other concerned parties on the date, time and venue of the briefing meeting.

At the briefing meeting, the detailed schedule of activities and performance of the tasks as indicated in 3 above shall be discussed and preferably agreed upon and shall be prepared and submitted to the Team Leader within 5 workdays after the debriefing.

On or before the last day of the engagement, the STE shall present the substantive findings and recommendations at a debriefing meeting to be convened and notified by DTI. All concerned parties shall have the opportunity to react to the findings and recommendations to be reflected in the expert's Report at that meeting.

In accordance with ACE standard operating procedure, the expert will accomplish and sign a monthly timesheet for approval and signature of KE1/Team Leader and a responsible official of DTI, and to be noted by the TRTA 3 Imprest Administrator, reflecting the actual number of workdays rendered, briefly indicating the place where, and the task for which, these were rendered.

The following documents are attached to this ToR for guidance of the STE :

1. Guidelines for Preparation and Review of TRTA 3 Mission Report
2. Template for TRTA 3 Mission Report
3. Template for Timesheet
4. Guidelines for Funding of Training, Workshop and Seminar Activities
5. Guidelines for Preparation of Design of Training, Workshop and Seminar

**ANNEX 2 – (I.E., SEMINAR/TRAINING/WORKSHOP MATERIALS : PPT. PRESENTATION, ETC.)**

The Workshop PowerPoint Presentations and Exercises and test results have all been distributed to participants and are all be available from this website:

**The Workshop Programme, The Questionnaire (post) with correct answers (red), Handouts (Day1 and Day2)**

**The Workshop Programme**

**Workshop**

**Upgraded NPAL's Chemist and Inspectors Knowledge and development of statistical design on sampling used for Field and Market Monitoring for fruits and vegetable**

**Sampling methods for food and food products and how to establish a sampling plan for improving quality through standardization and testing, Regression theory applied for the needs of related testing**

**Dec.01: Subject - Introduction to sampling and (the necessary) basic statistical knowledge**

8:00-10:00 1st Morning session:

Opening of course

Introduction by lecturer

Questionnaire Pre (and Post) Course Knowledge (Questions to Upgraded NPAL.docx)

Lectures:

Deduction and Induction

Sampling—why and how?

Descriptive Statistics for Samples

Discrete Example

Continuous Example

10:00-10:30 Coffee break

10:30-12:00: 2nd Morning session:

Lectures:

Centre of a Distribution

Introduction to and comparison of Mean, Median, and Mode

Spread of a Distribution

Group work by participants:

Group Work on what Mean, Mode and Median would signify in the context of sampling of NPAL and other organizations.

Why and how are these concepts important and apply for the work of NPAL and other organizations

12:00-13:00 Lunch break;

13:00-15:00: 1st Afternoon session:

Lectures:

Probability

Introduction to Probability  
Concept of Probability  
Elementary Properties of Probability  
Probability Distributions

15:00-15:30 Coffee break;

15:30-17:30 2st Afternoon session: Day's synopsis by lecturer and lecture:

Lectures:

Random sampling  
Systematic sampling  
Stratified samples  
Sample sizes within strata  
Quota sampling  
Cluster and multistage sampling  
Area sampling

Group work by 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)

## **Dec.02: Subject - Advanced (and useful) statistical knowledge for sampling**

8:30-10:00 1st Morning session:

Lectures:

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

10:00-10:30 Coffee break

10:30-12:00: 2nd Morning session:

Lectures:

To continue from 1<sup>st</sup> Morning session

Group work by participants:

Group Work on 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 to 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

12:00-13:00 Lunch break;

13:00-15:00: 1st Afternoon session:

Lectures:

- 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

15:00-15:30 Coffee break;

15:30-17:30 2st Afternoon session: Lecture: to continue from morning lectures, Day's synopsis by lecturer:

Group work by participants:

Calculation of Probabilities t-tests, confidence intervals, sample size Distributions characteristics of samples of NPAL and other organizations. Which calculation and sampling method applies to practices of the work of NPAL and other organizations and how and which method . Exercises on Regression Analysis for Agriculturalists, Chemists, Analysts prepared but not mandatory (if possible for the 4 different types of sampling points)

Second brief assessment of NPAL and other staff members: What are our needs? What do we want to improve?

### **Dec.03: Final Group works, exercises an Assessment**

8:00-10:00: 1st Morning session:

Lectures:

- Course's synopsis by lecturer and lecture:
- In case of Omissions further questions and clarifications

Group work by participants:

Group Work:. Which method could and should be applied for the work of NPAL and other organizations- (if possible for the 4 different types of sampling points)


If required continuation of the application exercises on Sapling, Testing and Regression Theory, (if this falls into the area of work of the group participants).

10:00-10:30 Coffee break


10:30-12:30: 2nd Morning session:

- If necessary continuation of previous group works by participants:
- Group Evaluation of Workshop and
- Presentation of Course Results by groups
- Opinion based Evaluation questionnaire by participants and
- Questionnaire (Pre and) Post Course Knowledge (Questions to Upgraded NPAL.docx)
- Closing of course

## The Questionnaire:



DEPARTMENT OF TRADE AND INDUSTRY  
PHILIPPINES



TRADE RELATED TECHNICAL ASSISTANCE PROJECT 3  
A project supported by the European Union

407 Boulevard de la Woluwe (CCK Bldg.) 1200 Brussels, Belgium  
Tel: +32 (0) 2 745 17 17 / Fax: +32 (0) 2 745 17 17 / Email: dti@dti.gov.ph


**Upgraded NPA's Chemist and Inspectors Knowledge and development of statistical design on sampling used for Field and Market Monitoring for fruits and vegetable.**

**From 04 To 03 of December in Baguio City, Philippines**


**Some Final Questions about Sampling Statistics (please tick one answer or many, none if you think none of the proposed answers do apply or if you don't know.) Think of these questions as of a self-assessment not as a test (results will be kept strictly anonymous)**

- In Sampling the essential thinking is that we are arguing from a sample proportion to a population proportion. How do we continue reasoning?
  - a. Induction ☒
  - b. Deduction ☐
  - c. Regression ☐
- What is statistical inference?
  - a. Drawing conclusions by applying ideas of logic/statistics to observations or hypotheses ☒
  - b. Free association ☐
  - c. For lacking verbal argumentation, we choose statistical formulas ☐
  - d. Another way to impress the public ☐
  - e. Associative inference deals with associations, relationships, etc., but not with causal connections between variables ☒
- Random sampling method is used to:
  - a. avoid going through a complicated selection procedure ☐
  - b. give each member of the population an equal chance of being included in the sample ☒
  - c. select items (often called Units) from a population where the probability of choosing a specific item is the proportion of those items in the population ☒

Page 1



DEPARTMENT OF TRADE AND INDUSTRY  
PHILIPPINES





TRADE RELATED TECHNICAL ASSISTANCE PROJECT 3  
A project supported by the European Union

407 Boulevard de la Woluwe (CCK Bldg.) 1200 Brussels, Belgium  
Tel: +32 (0) 2 745 17 17 / Fax: +32 (0) 2 745 17 17 / Email: dti@dti.gov.ph

- Stratified sampling method is used to:
  - a. may reduce the sample size considerably ☒
  - b. to avoid under-sampling ☐
  - c. divide the members of the population into homogeneous subgroups before sampling ☒
- How many strata are used best in stratified sampling?
  - a. As many as can be justified economically ☒
  - b. About 5 to 15 ☒
  - c. As many as are proportional to the size of the sample size divided by the universe ☐
- What is a probability distribution of a sample / universe?
  - a. A number to indicate how many items are distributed in the sample ☐
  - b. A graphical expression of the sample with in the universe ☐
  - c. A probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey ☒
- Is it useful to use non-probability sampling techniques. Why/why not?
  - a. Because statistical reasoning is based on assumptions? ☒
  - b. To prove the validity of these assumptions ☒
  - c. The data collected 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 errors ☒
- Did you ever hear of hypothesis and of errors in testing and of which?
  - a. Hypothesis ☒
  - b. Hypothesis ☒
  - c. Type I error ☒
  - d. Type II error ☒
- Confidence intervals used to:
  - a. increase the confidence of samples ☐
  - b. to avoid random sampling ☐
  - c. assign a numerical range with a probability. We must concede the possibility that we are wrong. ☒

Page 1

**TRADE RELATED TECHNICAL ASSISTANCE PROJECT 3**  
A project supported by the European Union  
AP Board of Investors (DTI) Bldg. 9th Flr. 20 Pineda Avenue, Alab City 1400 Philippines  
Po. Box: 4462 73 4001 +662 976 0000 to 0007 • Fax No. +638 875 8712

10. Which of the following examples are randomized samples?

- ☒ a. Rainfall data collected by the Meteorological Institute
- ☐ b. Reports of people on Lancetiles
- ☐ c. Production occurrence measuring reports of hospital cases

11. Which of the following statements are true?

- ☒ a. A t-test can make inference about the difference of two means
- ☒ b. A t-test is only statistical hypothesis test in which the test statistic follows a Student's t distribution
- ☐ c. The Student's t distribution is independent of the sample size

12. What does the statistical regression do?

- ☐ a. Regression explains the cause of one event by observing others
- ☒ b. In the regression equation expresses mathematically the relationship between two or more regression
- ☐ c. With regression functions we can predict the future
- ☒ d. Regression functions allow predictions under certain statistical procedures

13. Which of the following statements are correct in statistics?

- ☒ a. Confidence intervals relate the quality of predictions with observed distribution of the sample
- ☐ b. It is better not to publish confidence intervals because nobody understands them anyway
- ☒ c. Confidence intervals are always related to a confidence level (probability)

14. If we want to interpret statistical results, which of the following remarks are correct?

- ☐ a. We say that once an official statistics published, it must be correct, even if we don't know it when the calculation of this statistic
- ☒ b. The interpretation of statistical results should improve the reasoning for more profound research to improve medical control in general
- ☐ c. The current use of the interpretation of statistical results is sufficient for Public health control in general

08.12.2015
Page 3

## Handouts

### Day1

#### Slide 1




**Sampling and improving quality for testing food and food products**

Workshop Baguio Dec. 01.-03. 2015

08.12.15
EU-Philippines  
Trade Related Technical Assistance Project 3

### Day2




**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:** *Maybe and if no other requests* occur 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

08.12.15
EU-Philippines  
Trade Related Technical Assistance Project 3
Page 1

## Day1

### Slide 2

**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 - presentation participant's group

**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 and presentation by participants

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 2

## Day2

**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 be applied for the work of NPAL and other organizations - Formulate of Test Hypothesis and distinguish between the two possible errors, when induced 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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 3

### Slide 3

**Workshop program Day 1 Morning**

Day 1 / 1. Morning Session:  
Deduction and Induction  
Sampling—why and how?  
Discrete Example  
Continuous Example

Day 1 / 2. Morning Session:  
Descriptive Statistics for Samples  
Centre of a Distribution  
Introduction to and comparison of Mean, Median, and Mode

Group Work participants:  
What Mean, Mode and Median does signify in the context of sampling of NPAL and other organizations.  
Why and how are these concepts important and apply for the work of NPAL and other organizations

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 3

**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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 4

### Slide 4

**Deduction and Induction (A First Example of a Sample)**

Before every presidential election, the pollsters try to pick the winner; specifically, canvassing all voters would be an impossible task. The only alternative, pollsters survey a sample of a few thousand in the hope that the sample proportion will constitute a good estimate of the total population proportion. This is a typical example of **statistical inference** or **statistical induction**: the (voting) characteristics of an unknown population are inferred from the (voting) characteristics of an observed sample

if the sampling is done fairly, randomly and adequately, we can have hopes that the sample proportion will be close to the population proportion. This will allow us to estimate the unknown population T proportion from the observed sample proportion with 95% confidence :

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

T being the population, P the sample proportion and n the sample size

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 4

**The null hypothesis (1)**

The first step in evaluating sample results is to set up a null hypothesis ( $H_0$ ). 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 ( $H_1$ ) is a statement relating to the researchers' original hypothesis. Thus, in the above example, the alternative hypothesis could either be:

a.  $H_1$ : There is a difference between the proportions of housewives aware of the brand, before and after the campaign,

or

b.  $H_1$ : There is an increase in the proportion of housewives aware of the brand, after the promotional campaign.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 5

### Slide 5

**Deduction and Induction (1)**

Example: Just before a presidential election, a poll of 2,000 voters shows 760 for Candidate A and 1,240 for Candidate B. Calculate the 95% confidence interval for the population proportion T that will vote for Candidate A, that is with 95% confidence, the proportion for candidate A among the whole population of voters will be between 38% and 48%.

The estimate is not made with certainty; we are only 95% confident. We must concede the possibility that we are wrong—simply because we were unlucky enough to draw a misleading sample. For example, if less than half the population is in fact supporters of candidate A it is still possible, though unlikely, for us to run into a string of supporters of candidate A in our sample.

As sample size n increases, we note that the error allowance in decreases.

Suppose that we feel that 95% confidence is not good enough, and that instead we want to be 99% sure of our conclusion. If the additional resources for further sampling are not available, then we can increase our confidence only by making a less precise statement. For 99% confidence the formula must have the coefficient 1.96 enlarged to 2.58; this yields the 99% confidence interval

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 4

**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 single Tailed (~Normal distribution to come later in the workshop)"  
Source: <http://dianay24.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 single Tailed?**

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 4

## Day1

### Slide 6

**Deduction and Induction (2)**

Deduction involves arguing from the general to the specific—i.e., from the population to the sample. Induction is the reverse—arguing from the specific to the general, i.e., from the sample to the population. The mentioned Equation represents inductive reasoning; we are arguing from a sample proportion to a population proportion. The inductive statement (that the population proportion can be inferred from the sample proportion) is based on a prior deduction (that the sample proportion is likely to be close to the population proportion).

**Sampling (1):**  
We draw a sample, rather than to examine the whole population, for several reasons:

- Limited resources.
- Scarcity. Sometimes only a small sample is available.
- Destructive testing. (light bulbs, NPAL samples)

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 4

## Day2

**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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 7

### Slide 7

**Discrete Example**

In a sample of 50 families, let us record the number of children,  $X$ , which takes on the values 0, 1, 2, 3, ... We call  $X$  a "discrete" random variable because it can take on only a finite number of values. Suppose that the 50 values of  $X$  turn out to be:

(1) Number of Children	(2) Tally	(3) Frequency ( $f_i$ )	(4) Relative Frequency ( $f_i/n$ )
0		15	0.30
1		10	0.20
2		13	0.26
3		6	0.12
4		3	0.06
5		3	0.06

In column (3) we record, for example, that 13 is the frequency that we observed for a two-child family. That is, we obtained this outcome on 13/50 of our sample observations; this proportion (26 or 26%) is called relative frequency ( $f_i/n$ ), and is recorded in the last column.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 7

**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  $\alpha$ ) 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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 8

### Slide 8

**Calculation of the Frequency and Relative Frequency of the Number of Children in a Sample of 50 families**

where  $\Sigma$  means "the sum of." Thus, for example,  $\Sigma f_i$  means "the sum of the frequencies."

$\sum f_i = 50$        $\sum \frac{f_i}{n} = 1$

The information from column (3) is called a "frequency distribution," which is graphed in the following figure. The "relative frequency distribution" in the last column could be graphed similarly; note that the two graphs are identical except for the vertical scale. Hence, a simple change of vertical scale transforms the figure's left side into a relative frequency distribution (right side).

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 9

**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

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 9

### Slide 9

**Continuous Example**

If we take a sample of 200 men, each of whose height is recorded in inches. We call height  $X$  a "continuous" random variable, since an individual's height might be any value, such as 64.328 inches. It no longer makes sense to talk about the frequency of this specific value of  $X$ , since never again will we observe anyone who is exactly 64.328 inches tall. Instead we can tally the frequency of heights within a class or cell (e.g., 58.5" to 61.5"), as in the following table. Then the frequency and relative frequency are tabulated, as before.

(1) Cell No.	(2) Cell Boundaries	(3) Cell Midpoint	(4) Frequency	(5) Relative Frequency
1	58.5-61.5	60	2	0.01
2	61.5-64.5	63	10	0.05
		66	48	0.24
		69	64	0.32
		72	56	0.28
		75	16	0.08
7	76.5-79.5	78	4	0.02

Number of cells is a reasonable compromise between too much detail and too little. Usually, 5 to 15 cells is appropriate.

Each cell midpoint, which will represent all observations in the cell, is a convenient whole number.

The grouping of the 200 observations into cells is illustrated.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 9

**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 cooperate.

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  $n$ th 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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 10

## Day1

### Slide 10

**Continuous Example**

The grouped data are graphed in the following figure. We use bars to represent frequencies as a reminder that the observations occurred throughout the cell, and not just at the midpoint. Such a graph is called a bar diagram or histogram.

But how may characterize a sample frequency distribution with a single descriptive number. There are two very useful concepts: the first is the centre of the distribution, and the second is the spread. These concepts will be illustrated with the continuous distribution of men's heights; but their application to discrete distributions (such as family size) is even more straightforward.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 10

## Day2

### 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 to 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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 11

### Slide 11

**Centre of a Distribution**

There are many different ways to measure the centre of distribution. Three of these—the mode, the median, and the mean—are discussed as follows:

- The Mode**  
Since mode is the French word for fashion, the mode of a distribution is defined as the most frequent (fashionable) value. In the example of men's heights, the mode is 69 inches, since this cell has the greatest frequency or highest bar. Generally, the mode is not a good measure of central tendency. Why?
- The Median**  
The median is just the 50th percentile, i.e., the value below which 50% of the values in the sample fall. Since it splits the observations into two halves, it sometimes is called the middle value. In the sample of 200 detailed heights, the median (say, 69.3) easily is found by reading the 100th value from the left. But if the only information available is the grouped frequency distribution, the median can only be approximated, by choosing an appropriate value within the median cell.
- The Mean**  
This sometimes is called the arithmetic mean, or simply the average, and is the most common central measure. The original observations ( $X_1, X_2, \dots, X_n$ ) simply are summed, then divided by  $n$ .

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 11

**Introduction to Statistical Regression (2)**

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

Fertilizer (kg/ha)	Yield (kg/ha)
100	2.700
200	2.850
300	3.000
400	3.150
500	3.300
600	3.450
700	3.600

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 experimenter, we refer to it as an "independent variable" or "factor", or "regressor"  $X$ .

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 12

### Slide 12

**Comparison of Mean, Median, and Mode**

We showed in the previous slide a distribution that has a single peak and is symmetric (i.e., one half is the mirror image of the other); in this case, all three central measures coincide. But when the distribution is skewed to the right, as in in the following figure the median falls to the right of the mode; with the long scatter of observations strung out in the right-hand tail, we have to move from the mode to the right to pick up half the observations.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 13

**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  $Y_i$  to the fitted value  $\hat{Y}_i$  on the line, that is,  $(Y_i - \hat{Y}_i)$ . We note that this error is positive when the observed  $Y_i$  is above the line and negative when the observed  $Y_i$  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 to 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.

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 13

### Slide 13

**Spread of a Distribution (1)**

Although average height may be the most important single statistic, it also is important to know how spread out or varied the observations are. As with measures of centre, we find that there are several measures of spread.

- The Range**  
The range is simply the distance between the largest and smallest value. Range = largest — smallest observation. For men's heights, the range is 21 (i.e., 79.5-58.5).
- Mean Absolute Deviation (MAD)**  
The average deviation, as its name implies, is found by calculating the deviation of each observation from the mean.
- Variance and Standard Deviation**  
MSD is ok, provided that we only wish to describe the sample. But typically we shall want to go one step further and use this to make a statistical inference about the population. The Standard Deviation  $s$  is the square root of the Variance

$$\text{Variance, } s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 14

**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  $P_1$  to another point  $P_2$ , we denote the horizontal distance by  $\Delta X$  (where  $\Delta$  means change, or difference), and the vertical distance by  $\Delta Y$ . Then the slope  $m$  is defined as:

$$\text{slope} = \Delta Y / \Delta X$$

The characteristic of a straight line is that this slope remains the same everywhere:  $\text{slope} = \Delta Y / \Delta X = b$  (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  $P_0$ , the Y-intercept: since its coordinates are  $P_0(0, a_0)$ . In moving to any other point  $P(X, Y)$  on the line, we may write:  $\text{slope} = \Delta Y / \Delta X = b = (Y - a_0) / (X - 0)$ .

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 14

## Day1

### Slide 14

**BUREAU OF PLANT INDUSTRY** **AECOM**

Spread of a Distribution (2)

- Kurtosis is based on the size of a distribution's tails. Distributions with relatively large tails are called "leptokurtic"; those with small tails are called "platykurtic." A distribution with the same kurtosis as the normal distribution is called "mesokurtic."
- The following formula can be used to calculate the Kurtosis of a sample:

$$Kurtosis = \frac{1}{(n-1)s^4} \sum_{i=1}^n (X_i - \bar{X})^4 - 3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with this definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 14

## Day2

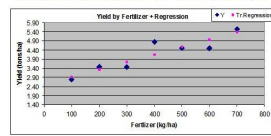
**BUREAU OF PLANT INDUSTRY** **AECOM**

Introduction to Statistical Regression (5)

After transformation we arrive at the equation of a line  $Y=a+bX$ , where  $a$  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:  $Y = a + bX$ . The geometry of lines.



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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 15

### Slide 15

**BUREAU OF PLANT INDUSTRY** **AECOM**

Spread of a Distribution (3)

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail, skewed right means that the right tail is long relative to the left tail.
- The following formula can be used to calculate the Skewness of a sample:

$$Skewness = \frac{1}{(n-1)s^3} \sum_{i=1}^n (X_i - \bar{X})^3$$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 16

**BUREAU OF PLANT INDUSTRY** **AECOM**

Day2: First Group Exercise:

Questions:

- What does a Test Hypothesis mean in the context of sampling of NPAL and other organizations, please name one?
- 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 file EX\_D02\_Sour.c.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:

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 16

### Slide 16

**BUREAU OF PLANT INDUSTRY** **AECOM**

First Group Exercise:

Questions:

- What Mean, Mode and Median would signify in the context of sampling of NPAL and other organizations.
- Why and how are these concepts important and apply for the work of NPAL and other organizations

Exercise:

- Please use the file "EXD01\_1\_Mango Yields Region IX.xls", a list of farmers from Region IX with supposed mango yields
- Please calculate Mean, Variance and Standard Deviation
- Try also Kurtosis and Skewness
- What would you say about the data. Are these yields?

Presentation:

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 17

**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, Tests, confidence interval, sample size, 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 – Exercises on Regression Analysis for Agriculturalists, Chemists, Analysts prepared but not mandatory (if possible for the 4 different types of sampling points)

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 17

### Slide 17

**BUREAU OF PLANT INDUSTRY** **AECOM**

Workshop program Day 1 Afternoon

Day 1/ 1. Afternoon Session:

- Probability
- Introduction to Probability
- Concept of Probability
- Elementary Properties of Probability
- Probability Distributions

Day 1/ 2. Afternoon Session:

- Random sampling
- Systematic sampling
- Stratified samples
- Sample sizes within strata
- Quota sampling
- Cluster and multistage sampling
- Area sampling

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)

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 17

**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 " $\pi$ " and " $e$ " denote important mathematical constants, approximately 3.14 and 2.72 respectively.

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

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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 18

## Day1

### Slide 18

**Probability**

In the next slides, we make deductions about a sample from a known population. For example, if the population of voters is 55% in favour of one candidate, we can hardly hope to draw exactly that same percentage in a random sample. Nevertheless, it is "likely" that "close to this percentage will turn up in our sample. Our objective is to define likely" and "close to" more precisely; in this way we shall be able to make useful predictions. First, however, we must lay a good deal of groundwork. Predicting in the face of uncertainty requires a knowledge of the laws of probability, and the slides of today are devoted exclusively to their development. We shall begin with the simplest example - rolling dice - which was also the historical beginning of probability theory, several hundred years ago.

**Concept of Probability**

Suppose that a gambler has a die he suspects is loaded, and asks us the probability that it will come number one. One solution would be to roll it over and over again, observing the relative frequency of ones is  $1/6$ . Of course, rolling it five or ten times would not be enough to average out chance fluctuations. But over the long run, the relative frequency would settle down to a limiting value, which is probability.

**Probability = proportion, in the long run or in mathematical terms**

$$\Pr(e_1) = \lim_{n \rightarrow \infty} \frac{n_1}{n}$$

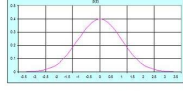
08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 19

## Day2

**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 specifies 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  $+z$ .



08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 19

### Slide 19

**Probability**

In the next slides, we make deductions about a sample from a known population. For example, if the population of voters is 55% in favour of one candidate, we can hardly hope to draw exactly that same percentage in a random sample. Nevertheless, it is "likely" that "close to this percentage will turn up in our sample. Our objective is to define likely" and "close to" more precisely; in this way we shall be able to make useful predictions. However, predicting in the face of uncertainty requires a knowledge of the laws of probability. We shall begin with the simplest example - rolling dice - which was also the historical beginning of probability theory, several hundred years ago.

**Concept of Probability**

Suppose that a gambler asks us the probability that it will come number one. One solution would be to roll it over and over again, observing the relative frequency of ones is  $1/6$ . Of course, rolling it five or ten times would not be enough to average out chance fluctuations. But over the long run, the relative frequency would settle down to a limiting value, which is probability.

**Probability = proportion, in the long run or in mathematical terms**

$$\Pr(e_1) = \lim_{n \rightarrow \infty} \frac{n_1}{n}$$

$e_1$  is the outcome ("1")  
 $n$  is the total number of times that the trial is repeated (die is thrown)  
 $n_1$  is the number of times that the outcome  $e_1$  occurs (frequency)  
 $n_1/n$  is therefore the relative frequency of  $e_1$   
limit is "the limit of ... as  $n$  approaches infinity."

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 19

**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  $\bar{X}$  is then called a derived distribution or a sampling distribution.

We concluded so far that  $\bar{X}$  was a good estimator of  $\mu$  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  $\mu$  is precisely equal in the observed sample mean. Instead, we must construct an interval estimate or confidence interval of the form:  $\mu = \bar{X} \pm \text{sampling error}$ .

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  $\mu$ . 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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 20

### Slide 20

**Elementary Properties of Probability**

- $0 \leq \Pr(e) \leq 1$  and
- Any probability lies between 0 and 1
- The sum of all relative frequencies adds up to 1

**Discrete Random Variables**

Suppose that a couple is planning three children and is primarily interested in the number of boys. This is an example of a random variable and is usually denoted by a capital letter:  $X$  = the number of boys

The possible values of  $X$  are 0, 1, 2, 3; however, they are not equally likely. To find the probabilities are, we must examine the original sample space. Thus, for example, the event "one boy" ( $X=1$ ) consists of three of the equally probable outcomes and its probability is  $3/8$ . Similarly the probability of each of the other events is computed

Original Sample Space	$\Pr(e)$
G G G	$1/8$
G G B	$1/8$
G B G	$1/8$
G B B	$1/8$
B G G	$1/8$
B G B	$1/8$
B B G	$1/8$
B B B	$1/8$

Smaller derived sample spaces	$\Pr(X)$
0	$1/8$
1	$3/8$
2	$3/8$
3	$1/8$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 20

**Confidence intervals and t-Test**

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

The confidence interval  $P(\mu - 1.96 \text{ sSME} < \mu < 1.96 \text{ sSME}) = 95\%$

Since  $\sigma$  is unknown, the statistician who wishes to evaluate the confidence interval (95%) must use some estimator of  $\sigma$ . 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

**Formulas for confidence intervals**

The general  $\mu = \bar{X} \pm z_{0.025} \frac{s}{\sqrt{n}}$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 21

### Slide 21

**Mean and Variance of Population and Sample**

If the sample size were increased without limit, the relative frequency distribution would settle down to the probability distribution. Relative frequency becomes probability in the limit.

From the relative frequency distribution, we calculated the mean  $\bar{X}$  and variance  $s^2$  of the sample. It is natural to calculate analogous values from the probability distribution and call them the mean  $\mu$  and variance  $\sigma^2$  of the probability distribution  $p(x)$ , or of the random variable  $X$  itself

So the population mean is  $\mu = \sum x p(x)$

and the population variance  $\sigma^2 = \sum (x - \mu)^2 p(x)$

We are following the usual custom of reserving Greek letters for population values. In Greek  $\mu$  is the equivalent of  $m$  for mean, and  $\sigma$  is the Greek equivalent of  $s$  for standard deviation.

A clear distinction must be made between sample and population values:  $\mu$  is called the population mean since it is based on the population of all possible repetitions of the experiment; on the other hand, we call  $\bar{X}$  the sample mean since it is based on a mere sample drawn from the parent population. Similarly,  $\sigma^2$  and  $s^2$  represent population and sample variance, respectively

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 21

**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  $\sigma$  was known, we formed the standardized normal variable, which is the transformed general formula for confidence intervals

**By analogy we introduce "Student's t" variable**

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

The similarity of these two variables immediately is evident. The only difference is that  $Z$  involves  $\sigma$ , which usually is unknown; but  $t$  involves  $s$ , which always can be calculated from the sample. The distribution of  $t$  is similar to the normal distribution. The  $t$  distribution has a wider spread than the normal, of course, since the use of  $s$  instead of  $\sigma$  introduces additional uncertainty. Moreover, while there is one standard normal distribution, there is a whole family of  $t$  distributions.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 22

## Day1

### Slide 22

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Continuous Distributions**

In our previous example, we saw how a continuous quantity such as height could be nicely represented by a bar graph showing relative frequencies. This graph is reproduced in the following figure, below (with height now measured in feet, rather than inches; furthermore, the y-axis has been shrunk to the same scale as the x-axis). The sum of all the relative frequencies (i.e., the sum of all the heights of the bars) is of course 1, as noted before. We have then changed the vertical scale to relative frequency density. This rescaling is designed specifically to make the total area equal to 1.

All the theorems that we stated about discrete random variables are equally valid for continuous random variables, with summations replaced by integrals. Therefore theorems are giving for discrete random variables only. (For more explanation see documentation [1])

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 22

## Day2

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Confidence intervals and t-Test (3)**

Note that with small sample size, the t distribution differs substantially from the normal; but as sample size increases, it approaches the normal. The distribution of t is not tabulated according to sample size n (remember n-1), but rather according to the divisor in s2, which now is called "degrees of freedom". So for a sample size n we calculate the expected frequency and thus the confidence intervals in general terms for the population mean based on a sample:

$$\mu = \bar{x} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

where 1.925 is the critical t value leaving 2.5% of the probability in the upper tail, with n - 1 degrees of freedom. (<https://www.bartleby.com/7/7e/7e1e87>)

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 23

### Slide 23

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Sampling**

Up to now we have studied probability and random variables so that we can now answer the basic deductive question in statistics: What can we expect of a random sample drawn from a known population?

Moreover we will try in this section to:

- Distinguish between probabilistic and non-probabilistic sampling methods
- Understand the bases for stratifying samples
- Make an informed choice between random and quota samples
- Comprehend multistage sampling, and
- Appreciate the use of area or aerial sampling.

Two major principles underlie all sample design. The first is the desire to avoid bias in the selection procedure; the second is to achieve the maximum precision for a given outlay of resources. Bias in the selection can arise:

1. If the selection of the sample is done by some non-random method i.e. selection is consciously or unconsciously influenced by human choice
2. If the sampling frame (i.e. list, index, population record) does not adequately cover the target population
3. If some sections of the population are impossible to find or refuse to co-operate.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 24

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Day2: Second Group Exercise (1):**

**Exercise:**

- In a census taken 6 years ago, 60% of farms were found to be selling vegetable produce direct to urban markets. Recently a sample survey has been carried out on 1000 farms and found 70% of them were selling their vegetable produce to urban centres direct.
- Situation: Population statistics (P = 60%) are known
- Question: Has there been a change in 6 years or is the higher percentage (p = 70%) found due to sampling error?
- When the population value is known, we can know the sampling error and we use this error for the purpose of our statistical test. The standard error of a percentage is always  $\sqrt{p(1-p)/n}$ . In this case p = population value 2, and the size of the [1] sample, n, to ascertain SE, (but the 2-sample to compare p = 70%).
- The null hypothesis for this case is: "There is no difference between the sample percentage of farms selling direct to urban areas and the population percentage of farms found to be selling direct 6 years ago" (i.e. no change in the 6 years).
- This must be a 2-tailed test as it could not be assumed that there would either be more or less farms selling produce direct six years later.
- Please use „Excel t-Test2\_Farmers.xls“ also with Probabilities 65% and 80%.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 24

### Slide 24

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Random sampling**

Random, or probability sampling, gives each member of the target population a known and equal probability of selection. The two basic procedures are:

- the lottery method, e.g. picking elements out of a hat or bag
- the use of a table of random numbers.

**Systematic sampling**

Systematic sampling is a modification of random sampling. To arrive at a systematic sample we simply calculate the desired sampling fraction, e.g. if there are 100 distributors of a particular product in which we are interested and our budget allows us to sample say 20 of them then we divide 100 by 20 and get the sampling fraction 5. Thereafter we go through our sampling frame selecting every 5th distributor. In the purest sense this does not give rise to a true random sample since some systematic arrangement is used in listing and not every distributor has a chance of being selected once the sampling fraction is calculated.

Sample	Numbered Stores
1	1, 6, 11, 16, 21, 26
2	2, 7, 12, 17, 22, 27
3	3, 8, 13, 18, 23, 28
4	4, 9, 14, 19, 24, 29
5	5, 10, 15, 20, 25, 30

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 25

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Day2: Second Group Exercise (2):**

**Udacity Video on using t-tables:**

Udacity is a for-profit educational organization founded by Sebastian Thrun, David Stavens, and Mike Sokolsky offering massive open online courses. According to Thrun, the origin of the name Udacity comes from the company's desire to be "audacious for you, the student". [www.udacity.com](http://www.udacity.com)

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 26

### Slide 25

**BUREAU OF PLANT INDUSTRY** **AECOM**

**Stratified samples**

Stratification increases precision without increasing sample size. Stratification does not imply any departure from the principles of randomness. It merely denotes that before any selection takes place, the population is divided into a number of strata, then random samples taken within each stratum. It is only possible to do this if the distribution of the population with respect to a particular factor is known, and if it is also known to which stratum each member of the population belongs.

Random stratified sampling is more precise and often more convenient than simple random sampling. When stratified sampling designs are to be employed, there are 3 key questions which have to be immediately addressed:

- 1 The bases of stratification, i.e. what characteristics should be used to subdivide the universe/population into strata?
- 2 The number of strata, i.e. how many strata should be constructed and what stratum boundaries should be used?
- 3 Sample sizes within strata, i.e. how many observations should be taken in each stratum?

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 26

**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, the male and female professors were sampled independently, yielding the following annual salaries (in ten-thousands of dollars, rounded):

Male (X1)	Female (X2)
12	10
14	12
16	14
18	16
20	18
22	20
24	22
26	24
28	26
30	28

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:  $H = (\bar{X}_1 - \bar{X}_2) \pm 2.16 \cdot \sqrt{SE^2} = 5.0 \pm 2.16(1.67)$  this means:  $5.0 \pm 3.6$

Thus, with 95% confidence, H is estimated to be between 1 and 9. Thus the claim  $H = 0$  seems implausible, because it falls outside this confidence interval

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 26

## Day1

### Slide 26

**Bureau of Plant Industry** **AECOM**

**Bases of stratification**

In general, it is desirable to make up strata in such a way that the sampling units within strata are as similar as possible. In this way a relatively limited sample within each stratum will provide a generally precise estimate of the mean of that stratum. Similarly it is important to maximise differences in stratum means. This is desirable since stratification has the effect of removing differences between stratum means from the sampling error.

Total variance within a population has two types of natural variation: **between-strata variance** and **within-strata variance**. Stratification removes the second type of variance from the calculation of the standard error. Stratification ensures that variation between strata does not enter into the standard error by taking account of this source in drawing the sample.

**Number of strata**

As regards number of strata, as many as possible could/ should be used. But some practical problems limit the desirability of a large number of strata:

- 1 Past a certain point, the "residual" or "unexplained" variation will dominate, and little improvement will be effected by creating more strata.
- 2 Depending on the costs of stratification, a point may be reached quickly where creation of additional strata is economically unproductive.

If a single overall estimate is to be made like pesticide contents we would normally use no more than about 6 strata. If estimates are required for population subgroups (e.g. by region and/or age group), then more strata may be justified.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 34

## Day2

**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  $p(Y|x)$  have the same variance  $\sigma^2$  for all  $x$ .
2. The means  $E(Y|x)$  lie on a straight line, known as the true (population) regression line: The population parameter  $\alpha$  and  $\beta$  specify the line; they are to be estimated from sample information.
3. The random variables  $Y_i$  are statistically independent (i.e.  $Y_2$  is "unaffected" by  $Y_1$ )

It is useful to describe the deviation of  $Y_i$  from its expected value or disturbance term  $e_i$  so that the model  $Y_i = \alpha + \beta x_i + e_i$  alternatively may be written as: Obviously  $\alpha$  and  $\beta$  correspond to intercept  $a$  and slope  $b$  from our sample

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 37

### Slide 27

**Bureau of Plant Industry** **AECOM**

**Sample sizes within strata**

**Proportional allocation:**

Once strata have been established, the question becomes, "How big a sample must be drawn from each?" Consider a situation where a survey of a two-stratum population is to be carried out:

Stratum	Number of Items in Stratum
A	10,000
B	90,000

If the budget is fixed at \$3000 and we know the cost per observation is \$6 in each stratum, so the available total sample size is 500. The most common approach would be to sample the same proportion of items in each stratum. This is termed **proportional allocation**. In this example, the overall sampling fraction is:

$$\frac{\text{Sample Size}}{\text{Population Size}} = \frac{500}{100,000} = 0.005$$

So this method of allocation would result in:

Stratum	Sample Size
A	500
B	450

The major practical advantage of proportional allocation is that it leads to estimates which are computationally simple. Where proportional sampling has been employed we do not need to weight the means of the individual stratum when calculating the overall mean, so

$$\frac{\bar{x}_1 + \bar{x}_2}{2} = \bar{x}_{\text{total}}$$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 37

**Bureau of Plant Industry** **AECOM**

**Regression: Confidence intervals and hypothesis tests for  $\beta$**

**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  $\beta$ ), and similarly  $a$  is the minimum variance estimator of  $\alpha$ .

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],[3],[7]

**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  $Y_i$  are normal. Since  $b$  is a linear combination of the  $Y_i$ , it follows that  $b$  also will be normal. But even without assuming that the  $Y_i$  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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 38

### Slide 28

**Bureau of Plant Industry** **AECOM**

**Optimum allocation:**

Proportional allocation is advisable when all we know of the strata is their sizes. In situations where the standard deviations of the strata are known it may be advantageous to make a disproportionate allocation.

Suppose that, once again, we had stratum A and stratum B, but we know that the individuals assigned to stratum A were more varied with respect to their opinions than those assigned to stratum B. Optimum allocation minimises the standard error of the estimated mean by ensuring that more respondents are assigned to the stratum within which there is greatest variation. Then the obtained means would have to be weighted (See documentation [2] for further information)

**Sampling methods in overview**

Source: <http://www.fao.org/docrep/012/a5342e04.jpg>

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 39

**Bureau of Plant Industry** **AECOM**

**Confidence intervals for  $a$  and  $b$  (and so for  $\alpha$  and  $\beta$ )**

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

$$\beta = b \pm t_{0.025} s_b$$

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

$$\alpha = a \pm t_{0.025} s_a / \sqrt{n}$$

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  $Y_0$ , you can refer to Literature [1] but we will retain, that the combination of the above formulas (together with combinations of sampled values of  $x_i$ ) will allow to calculate confidence intervals for the regression values  $Y_i$

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 39

### Slide 29

**Bureau of Plant Industry** **AECOM**

**Quota sampling (1)**

Quota sampling is a method of stratified sampling in which the selection within strata is non-random. Selection is normally left to the discretion of the interviewer and it is this characteristic, which destroys any pretensions towards randomness.

**Quota v random sampling**

The advantages and disadvantages of quota versus probability samples has been a subject of controversy for many years. Some practitioners hold the quota sample method to be so unreliable and prone to bias as to be almost worthless. Others think that although it is clearly less sound theoretically than probability sampling, it can be used safely in certain circumstances. Still others believe that with adequate safeguards quota sampling can be made highly reliable and that the extra cost of probability sampling is not worthwhile.

Generally, statisticians criticise the method for its theoretical weakness while market researchers defend it for its cheapness and administrative convenience.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 39

**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  $\alpha$  and  $\beta$ ). In this section, we shall consider two narrower problems:

- (a) For a given value  $x_0$ , what is the interval that will predict the corresponding mean value of  $Y_0$ . 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  $Y_0$  (referred to as the prediction interval for an individual  $Y_0$ ). 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:

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 39

## Day1

### Slide 30

**Quota sampling (2)**

**Contra**

- 1 No estimates of sampling errors
- 2 The interviewer often fail to secure a representative sample of respondents in quota sampling.
- 3 Strata controls leave a lot to the interviewer's judgement.
- 4 Strict control of fieldwork is more difficult, or impossible

**Pro**

- 1 Less costly.
- 2 It is easy administratively.
- 3 The labour of random selection is avoided
- 4 If fieldwork has to be done quickly, quota sampling may be the only possibility.
5. Quota sampling is independent of the existence of sampling frames.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 30

## Day2

**Example of Interval estimates calculated**

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 31

### Slide 31

**Cluster and multistage sampling**

**Cluster sampling:**

For example, a survey is to be done in a large town and that the unit of inquiry (i.e. the unit from which data are to be gathered) is the individual household. Suppose further that the town contains 20,000 households, all of them listed on convenient records, and that a sample of 200 households is to be selected. One approach would be to pick the 200 by some random method. However, this would spread the sample over the whole town, with high fieldwork costs. (think of rural areas in developing countries). One might decide therefore to concentrate the sample in a few parts of the town and it may be assumed for simplicity that the town is divided into 400 areas with 50 households in each. A simple course would be to select say 4 areas at random (i.e. 1 in 100) and include all the households within these areas in our sample. The overall probability of selection is unchanged, but by selecting clusters of households, one has materially simplified and made cheaper the fieldwork.

**A large number of small clusters is better, all other things being equal, than a small number of large clusters. Whether single stage cluster sampling proves to be as statistically efficient as a simple random sampling depends upon the degree of homogeneity within clusters.**

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 31

**Example of Interval estimates calculated cont'd**

The above mentioned and calculated example [1] was for one observation  $x_0 = 559$ . 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  $x_0$  moves farther away from its central value (the mean) this reflects the fact that  $x_0$  appears in both variances. This example gives reason for some general remarks about predictions based on statistical regression.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 32

### Slide 32

**Multistage sampling**

The population is regarded as being composed of a number of first stage or primary sampling units (PSUs) each of them being made up of a number of second stage units in each selected PSU and so the procedure continues down to the final sampling unit, with the sampling ideally being random at each stage.

The necessity of multistage sampling is easily established. PSUs for national surveys are often administrative districts, urban districts or parliamentary constituencies. Within the selected PSU one may go direct to the final sampling units, such as individuals, households or addresses, in which case we have a two-stage sample.

**Area sampling**

Area sampling is basically multistage sampling in which maps, rather than lists or registers, serve as the sampling frame. This is the main method of sampling in developing countries where adequate population lists are rare. The area to be covered is divided into a number of smaller sub-areas from which a sample is selected at random within these areas; either a complete enumeration is taken or a further sub-sample. **Think of this if no complete list of okra fields is available.**

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 32

**Dangers of extrapolation**

We emphasize that  $x_0$  may be any value of  $x$ . If  $x_0$  lies among the observed values  $x_1 \dots x_n$ , the process is called interpolation. If  $x_0$  is out beyond the observed values  $x_1 \dots x_n$ , 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  $x_0$  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  $x_0$  moves away from the centre. This is true, even if all the assumptions underlying our mathematical model hold exactly.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 33

### Slide 33

**Example of Area sampling**

A grid, such as that shown above, is drawn and superimposed on a map of the area of concern. Sampling points are selected on the basis of numbers drawn at random that equate to the numbered columns and rows of the grid.

If the area is large, it can be subdivided into sub-areas and a grid overlayed on these. The second figure depicts the procedures involved. As in the first figure the columns and rows are given numbers. Then, each square in the grid is allocated numbers to define grid lines. Using random numbers, sampling points are chosen within each square. The second gives an impression of the pattern of sampling which emerges from this multistage aerial sampling.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 34

**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 over-dosed. 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  $x_0$  moves away from the centre. This is true, even if all the assumptions underlying our mathematical model hold exactly.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 34

## Day1

### Slide 34

**BUREAU OF PLANT INDUSTRY** **AECOM**

Second Group Exercise:

Exercise:

- Calculation of Probabilities Distributions characteristics of samples of NPAL and other organizations - (if possible for the 4 different types of sampling points)
- Please take a typical sample size of your working area and divide it by the population size (the 1), you don't know the population size because you don't have a sample frame - please guess the population size, also in several stages (e.g. vegetables in markets (1.stage), markets in cities / regions (2.stage))

Questions:

Group Work on 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). First assessment of NPAL and other staff members: What are our needs? What do we want to improve?

Presentation:

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

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 34

## Day2

**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:  $\sigma^2$  (and  $\alpha$  and  $\beta$ ) 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])**

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 35

### Slide 35

**BUREAU OF PLANT INDUSTRY** **AECOM**

Literature

[1] Klaus Röder : Handbook Introduction to Statistics: [http://www.klaus-roeder.com/Ordner/PDFs/Projects/13WoD/13WoD\\_Handbook\\_WeD\\_and\\_Statistics\\_130208.pdf](http://www.klaus-roeder.com/Ordner/PDFs/Projects/13WoD/13WoD_Handbook_WeD_and_Statistics_130208.pdf)

[2] Introductory Statistics, 5th Edition 5th Edition, by Thomas H. Wonnacott (Author), Ronald J. Wonnacott (Author); ISBN-13: 978-0471615157

[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/012/1379e/1379e05.pdf>

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 36

**BUREAU OF PLANT INDUSTRY** **AECOM**

Day2: Second Group Exercise (3):

Exercise:

- Group Work on what Confidence Intervals and t-Test could signify in the context of sampling of NPAL.
- Use File ExD02\_1\_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-Mining). 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.

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 36

### Slide 36

**BUREAU OF PLANT INDUSTRY** **AECOM**

Day2: Second Group Exercise (4):

Questions:

Group Work on which test would apply for confidence intervals and 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)

08.12.15 EU-Philippines Trade Related Technical Assistance Project 3 Page 37

The Workshop PowerPoint Presentations and Exercises and test results will all be available from this website:

[http://www.klaus-roeder.com/4\\_Projekte/projekte.html](http://www.klaus-roeder.com/4_Projekte/projekte.html)