DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

Klaus Röder
AECOM
www.klaus-roeder.com

Sampling and improving quality for testing food and food products

Workshop Baguio Dec. 01.-03. 2015

28.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

Workshop schedule

Dec.01: 8:30-10:00 1st Morning session: Opening, Lectures
10:00-10:30 Coffee break
10:30-12:30 2nd Morning session: Lectures/group work participants
12:00-13:30 Lunch break
13:30-15:00: 1st Afternoon session: Lectures/group work participants
15:00-15:30 Coffee break
15:30-17:30 2nd Afternoon session: Day's synopsis by lecturer/group work participants
Dec.02: as above
Dec.03:
8:30-10:00 1st Morning session; 10:00-10:30 Coffee break
10:30-12:30: 2nd Morning session: Workshop's synopsis by lecturer and group assessment by participants

28.11.15

EU-Philippines
Trade Related Technical Assistance Project 3

Page 2

AECOM

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

AECOM

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

Deduction and Induction (1)

Example: Just before an 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 Solution: for candidate A, that is with 95% confidence, the proportion for candidate A among the whole population of voters will be between 36% and 40%.

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 supports 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

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)

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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)	(4) Relative Frequency $\left(\frac{f}{n}\right)$
0	卌 卌 卌	15	.30
1	卌 卌	10	.20
2	卌 卌 |||	13	.26
3	卌		6	.12
4	|||	3	.06
5	|||	3	.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/n), and is recorded in the last column.

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

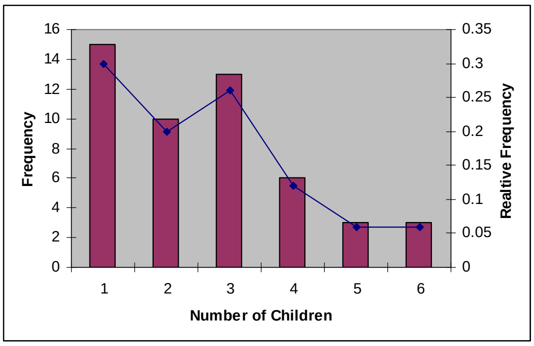
AECOM

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

$$\sum f = 50 \qquad \sum \frac{f}{n} = 1$$

where Σ means "the sum of." Thus, for example, Σ f means "the sum of the frequencies."

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).



DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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)	*(2)*	*(3)*	*(4)*	*(5)*
Cell No.	*Cel Boundaries*	*Cel Midpoints*	*Frequency*	*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.

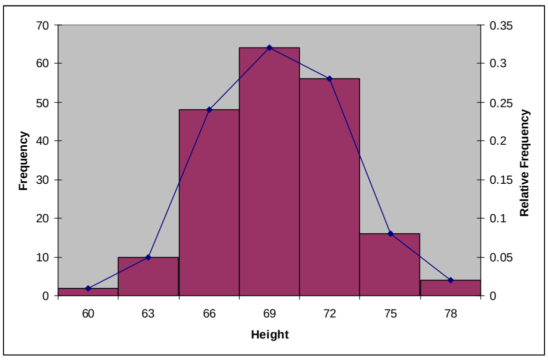
The grouping of the 200 observations into cells is illustrated.

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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.

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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. Bui 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.

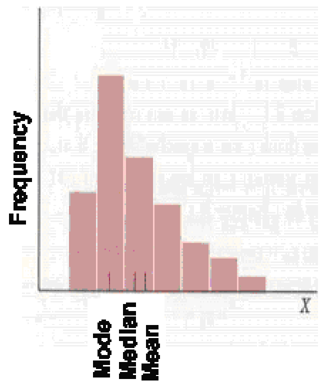
- **The Mean**
  This sometimes is called the arithmetic mean, or simply the average, and is the most common central measure. The original observations (X1, X2, . . . , Xn,,,) simply are summed, then divided by n.

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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 hall 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.



DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

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 a 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

$$Variance, s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

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

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Spread of a Distribution (2)

- Kurtosis

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:

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

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.

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

DEPARTMENT OF AGRICULTURE 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:

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

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 15

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY AECOM

First Group Exercise:

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

Exercise:
1. Please use the file “ExD01_1_Mango Yields Region IX.xlsx”, a list of farmers from Region IX with supposed mango yields
2. Please calculate Mean. Variance and Standard Deviation
3. Try also Kurtosis and Skewness

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

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

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 16

DEPARTMENT OF AGRICULTURE 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)

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

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

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 lo roll it over and over again, observing the relative frequency of ones is 1/6

$$\Pr(e_1) = \lim \frac{n1}{n}$$

Of course, rolling if five or ten tines 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

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

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

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 lo roll it over and over again, observing the relative frequency of ones is 1/6. Of course, rolling if five or ten tines 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

$$\Pr(e_1) = \lim \frac{n1}{n}$$

e1 is the outcome ("1")
n is the total number of times that the trial is repeated (die is thrown)
n1 is the number of times that the outcome el occurs (frequency,)
n1 / n is therefore the relative frequency of e1
lim is "the limit of . . . , as n approaches infinity."

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Elementary Properties of Probability

- 0<= ni/n <= 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)	e		
´1/8	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

Smaller derived sample space

X	p (x)
0	´1/8
1	´3/8
2	´3/8
3	´1/8

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Mean and Variance of Population an 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 s2 of the sample. It is natural to calculate analogous values from the probability distribution and call them the mean μ and variance σ2 of the probability distribution p(x), or of the random variable X itself

So the population mean is $$\mu \equiv \sum_x x p(x)$$

and the population variance $$\sigma^2 \equiv \sum_x (x-\mu)^2 p(x)$$

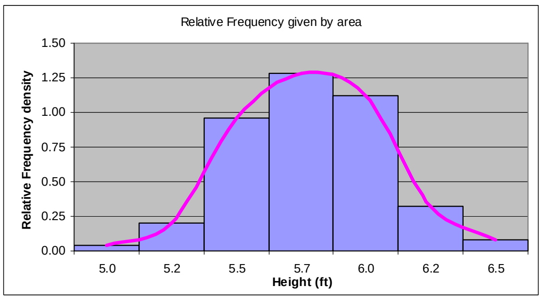
We are following the usual custom of reserving Greek letters for population values. In Greek μ is the equivalent of m for mean, and σ is the Greek equivalent of s for standard deviation.
A clear distinction must be made between sample and population values: μ is called the population mean since it is 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, σ2 and s2 represent population and sample variance, respectively

DEPARTMENT OF AGRICULTURE 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])

DEPARTMENT OF AGRICULTURE 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.

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 23

DEPARTMENT OF AGRICULTURE 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.

Systematic sampling						
Population = 100 Food Stores						
Sample desired = 20 Food Stores						
a. Draw a random number 1-5.						
b. Sample every X[th] store.						
Sample	Numbered Stores					
1	1,	6,	11,	16,	21...	96
2	2	7,	12	17,	22...	97
3	3,	8,	13	18,	23...	98
4	4,	9,	14	19,	24...	99
5	5,	10,	15,	20,	25...	100

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 24

DEPARTMENT OF AGRICULTURE 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?

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 25

DEPARTMENT OF AGRICULTURE 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.

28.11.15 EU-Philippines *Trade Related Technical Assistance Project 3* Page 26

DEPARTMENT OF AGRICULTURE 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:

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

So this method of allocation would result in:
Stratum A (10,000 × 0.005) = 50
Stratum B (90,000 × 0.005 = 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

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

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

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

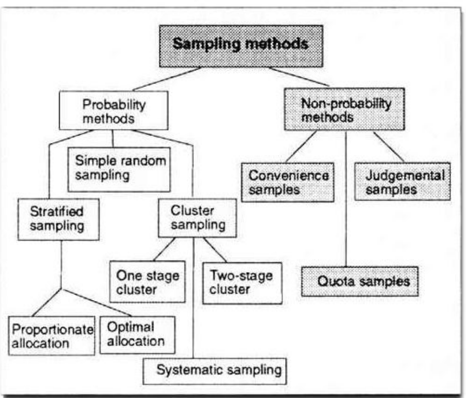
AECOM

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 [3] for further information)

Sampling methods in overview

Source http://www.fao.org/docrep/w3241e/w3241e0d.jpg



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

BUREAU OF PLANT INDUSTRY — DEPARTMENT OF AGRICULTURE   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.

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

BUREAU OF PLANT INDUSTRY — DEPARTMENT OF AGRICULTURE   AECOM

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. The labour of random selection is avoided
3 If fieldwork has to be done quickly, quota sampling may be the only possibility,
4. Quota sampling is independent of the existence of sampling frames.

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

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

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.

28.11.15 **EU-Philippines** *Trade Related Technical Assistance Project 3* **Page** 31

DEPARTMENT OF AGRICULTURE
BUREAU OF PLANT INDUSTRY

AECOM

Multistage sampling:

The population is regarded as being composed of a number of first stage or primary sampling units (PSU's) 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. PSU's 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 fields is available.

28.11.15 **EU-Philippines** *Trade Related Technical Assistance Project 3* **Page** 32

DEPARTMENT OF AGRICULTURE 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 (ha (!), 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)

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

DEPARTMENT OF AGRICULTURE BUREAU OF PLANT INDUSTRY

AECOM

Literature

[1] Klaus Röder : Handbook Introduction to Statistics: http://www.klaus-roeder.com/Ordner/PDFs/Projects/13WoD/13WoD_Handbook_WoD_and_Statistics_130208.pdf
[2] Introductory Statistics, 5th Edition 5th Edition, by Thomas H. Wonnacott (Author), Ronald J. Wonnacott (Author); ISBN-13: 978-0471615187
[3] Crawford, I. M. (1990), Marketing Research, Centre and Network for Agricultural Marketing Training in Eastern and Southern Africa, Harare, pp 36-48.
[4] FAO Sampling recommendations e.g.
http://www.fao.org/docrep/012/i1379e/i1379e05.pdf
http://www.fao.org/docrep/w3241e/w3241e08.htm

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