POLI 343
Introduction to Political Research

Session 13-Sources of Error in Research

Lecturer: Prof. A. Essuman-Johnson, Dept. of Political Science
Contact Information: aessuman-johnson@ug.edu.gh

UNIVERSITY OF GHANA
College of Education
School of Continuing and Distance Education
2014/2015 – 2016/2017

Introduction

One major problem that researchers have to deal with in the research process is the extent to which errors or bias has been kept out of the research. In general a good research, design which is efficiently executed, can save time and cost. But undue savings will affect the quality of the data. When resources are limited it is tempting to cut short the preliminary design work e.g. to omit the briefing of interviewers, to relax quality control checks, to dispense with supervision.

Sources of Error

- Any of these actions is likely to increase the number of errors contained in the data. Some errors are an inherent feature of all surveys and it is risky to introduce new ones.
- Survey errors vary in their cause/source, nature and importance. They can arise because surveys collect information from only a sample of the total population. Probability sampling is rarely able to represent a population perfectly.

SourĐes of Error ;CoŶt'd⍰:

- There might, by chance, be slightly too few men or too few young people or a particular occupational group might over-represented. But errors deriving from chance (probability) will be small and their approximate size is calculable.
- Errors can also arise because the list (sampling frame) used as a basis for selection is deficient. It might not, for example, contain immigrants in their true proportions. These deficiencies can be overcome or reduced, but it is usually costly to do so.

SourĐes of Error ;CoŶt'd:

- Where the sampling frame is accurate, errors will occur through non-response, because a sample will almost always fall short of the selected ones in some members of the sample will be away from home, some will refuse to participate some too might be ill. Survey response rate of 85% are rare. Often they range from 75% to 80%. Unless the characteristics of non-respondents are identical to those of respondents, the sample will contain bias (errors).

SourĐes of Error ;CoŶt'd:

❖And the bias will be serious when the characteristics of non-respondents are related to the subject under study.

❖The way questions have been designed, administered and answered can also produce errors e.g. respondents may misunderstand a question, interviewers may make recording errors, and answers may not be truthful. Errors are therefore difficult to identify.

SourĐes of Error ;CoŶt'd:

- ❖Coders may make errors as well as by misclassifying a response, data entry personnel may mistakenly hit the wrong key when transferring the response into computer for analysis.

These sources of errors can be identified by clerical or computer editing checks though a few will escape detection. Surveys can thus produce only approximate not precise measurements.

Human Error in phases of a Survey

Throughout the various phases of a survey, human error can and will be made. These phases are as follows:

❖Exploratory phase to help in design of questionnaires content and construction.

❖Deciding on the questions to be asked, the wording, sequence of questions and layout.

❖Sampling – determining those to be interviewed, how many to interview and method of selecting respondents.

Human Error in phases of a Survey ;CoŶt'd:

- Interview or mail questionnaire – These two are main means of data collection and the core of a sample survey.
- Field organization – This demands good management skills in recruiting, training and controlling an interviewing team.
- Data preparation – problems of coding the questionnaire for analysis.

Other Sources of Error in Research

Another source of error is provided by the measurement procedures adopted. The selection of methods to be used in research will depend on the questions to be answered. Often several methods are used in a single project to provide a wide range of evidence on the selected topic. The initial hypotheses specify the data to be collected. Operational definitions of concepts and models of the relationships between variable show how the data will be organized.

Other SourĐes of Error ;CoŶt'd:

The goal is obtain reliable and valid data, as free from bias as possible, which will provide an unambiguous response to the research questions. Measures are considered reliable if the results are consistent i.e. if the same people are asked the same questions again, they will give the same answers. A measurement is valid if they represent the true position e.g. the observer reports correctly on what happened, the strength of attitudes is accurately recorded etc.

Other SourĐes of Error ;CoŶt'd�:

A finding may be reliable but invalid or unreliable but |alid e.g. a ŵeasure of satisfaĐtioŶ ‖ith the preseŶt go|erŶŵeŶt ŵay ĐoŶsisteŶtly i.e. reliaďly, sho‖ that certain people are more satisfied than others even though both groups have the same level of satisfaction. In this case the instrument used to measure satisfaction is not valid. Reliability is especially difficult in a rapidly changing situation.

Other SourĐes of Error ;CoŶt'd:

Someone who says he plans to vote for candidate X may answer the same question a week later by saying he intends to vote for candidate Y. In such a situation, prediction of the outcome of the election may be impossible; here the measuring instrument is valid but the results are not reliable.

Factors accounting for Invalid/ Unreliable Measurement

Measurements may be unreliable or invalid due to any of
the following:

❖Defects in the measuring instrument e.g. concepts may be poorly operationalized, the wrong questions or not enough questions are asked.

❖The circumstances of data collection e.g. the people observed or questioned are in some way unusual, not their normal selves or are affected by the environment

❖Inadequate methods are used e.g. too small a sample chosen, answers wrongly recorded, analysis is carelessly done. Such defects introduce bias/errors i.e. systematic errors into the results.

Overcoming Measurement Problems:

Careful planning can help to avoid some of these problems, but their effect must be taken into account when the report is written. There will also be some random error i.e. some effects, which vary from one measurement to the next and thus lower the reliability of the findings somewhat no matter how much care is taken. Nothing can be done about random errors except to observe statistical safeguards. All research is involved in the never-ending fight against error.

Overcoming Measurement Problems:

Sometimes it is possible to compare observed diﬀereŶĐes ďetǁeeŶ groups agaiŶst ĐhaŶĐe errors ďy use of statistical tests of significance e.g. in designs in which respondents have been assigned to one of two or more groups at random.

Sources of Error Unaccounted for when Controlling Variables

More often, however, attempts at controlling variables have left a host of possible sources of error unaccounted for such as:

- ❖Faults in the design of the survey
- ❖Sampling errors
- ❖Errors due to non response
- ❖Bias due to questionnaire design and question wording

Sources of Error Unaccounted for when Controlling Variaďles ;CoŶt’d:

- ❖Unreliability or lack of validity of various measuring techniques used
- ❖Interviewer bias
- ❖Respondent unreliability, ignorance and misunderstanding.

Sampling Methods and Assessment of Sampling Error:

Big strides have been made in recent years in the improvement of sampling methods and the assessment of sampling error limits, but the remaining sources of error remain and any of them could easily outweigh the gains from improved sampling techniques. It becomes the burden for the researcher to remain severely critical, to search out biases in others and in himself and to avoid giving the appearance of spurious exactitude.

Non-Sampling Errors

The Sampling Distribution

So how do we get from our sample statistic to an estimate of the population parameter? A crucial midway concept you need to understand is the **sampling distribution**. In order to understand it, you have to be able and willing to do a thought experiment. Imagine that instead of just taking a single sample like we do in a typical study, you took three independent samples of the same population.

Non-SaŵpliŶg Errors ;CoŶt'd:

Furthermore, imagine that for each of your three samples, you collected a single response and computed a single statistic, say, the mean of the response. Even though all three samples came from the same population, you would not expect to get the exact same statistic from each. They would differ slightly just due to the random ͞luĐk of the draǁ͟ or to the natural fluctuations or vagaries of drawing a sample. But you would expect that all three samples would yield a similar statistical estimate because they were drawn from the same population.

Non-SaŵpliŶg Errors ;CoŶt'd:

Now, for the leap of imagination! Imagine that you did an *infinite* number of samples from the same population and computed the average for each one. If you plotted them on a histogram or bar graph you should find that most of them converge on the same central value and that you get fewer and fewer samples that have averages farther away up or down from that central value. In other words, the bar graph would be well described by the *bell curve* shape that is an indication of a "normal" distribution in statistics.

Non-SaŵpliŶg Errors ;CoŶt’d:

The distribution of an infinite number of samples of the same size as the sample in your study is known as the **sampling distribution**. We don't ever actually construct a sampling distribution. Why not? You're not paying attention! Because to construct it we would have to take an *infinite* number of samples and at least the last time I checked, on this planet infinite is not a number we know how to reach. So why do we even talk about a sampling distribution?

Non-SaŵpliŶg Errors ;CoŶt'd�:

Now that is a good question! Because we need to realize that our sample is just one of a potentially infinite number of samples that we could have taken. When we keep the sampling distribution in mind, we realize that while the statistic we got from our sample is probably near the center of the sampling distribution (because most of the samples would be there) we could have gotten one of the extreme samples just by the luck of the draw.

Non-SaŵpliŶg Errors ;CoŶt'd:

If we take the average of the sampling distribution (the average of the averages of an infinite number of samples), we would be much closer to the true population average-the parameter of interest. So the average of the sampling distribution is essentially equivalent to the parameter. But what is the standard deviation of the sampling distribution? You will need to brush up your Numeracy Skills if you have gotten rusty. From Numeracy Skills, you know what a standard deviation is.

Standard Deviation

The standard deviation of the sampling distribution tells us something about how different samples would be distributed. In statistics it is referred to as the **standard error** (so we can keep it separate in our minds from standard deviations). A standard deviation is the spread of the scores around the average in a single sample. The standard error is the spread of the averages around the average of averages in a sampling distribution).

Sampling Error

In sampling contexts, the standard error is called **sampling error**. Sampling error gives us some idea of the precision of our statistical estimate. A low sampling error means that we had relatively less variability or range in the sampling distribution. But here we go again-we never actually see the sampling distribution!

So how do we calculate sampling error? We base our calculation ***on the standard deviation of our sample***.

Standard Deviation and Sampling Error

The greater the sample standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size.

The greater your sample size, the smaller the standard error. Why? Because the greater the sample size, the closer your sample is to the actual population itself. If you take a sample that consists of the entire population you actually have no sampling error because you don't have a sample, you have the entire population. In that case, the mean you estimate is the parameter.

The 68, 95 and 99 Percent Rule

There is a general rule that applies whenever we have a normal or bell-shaped distribution. Start with the average-the center of the distribution. If you go up and down (i.e. left and right) one standard unit, you will include approximately 68% of the cases in the distribution (i.e. 68% of the area under the curve). If you go up and down two standard units, you will include approximately 95% of the cases. And if you go plus-and-minus three standard units, you will include about 99% of the cases.