Abstract

Marketing research uses two sources of data: primary and secondary. There are many advantages in use of secondary data but also there are many limitations such as different types of errors and biases that can arise in these data. Secondary data should be accurate, reliable, precise, unbiased, valid, appropriate and timely. Four categories of potential errors can reduce accuracy of secondary data: sampling and non-sampling errors, errors that invalidate the data, errors that require data reformulation and errors that reduce reliability. All sources of errors can lower the reliability and validity of results. This implies that secondary data have to be treated carefully.

Ključne reči: Secondary data, errors, bias, accuracy.

1. Introduction

Marketing concept requires that customer satisfaction rather than profit maximization be the goal of an organization. In other words, the organization should be consumer oriented and should try to understand consumer’s requirements and satisfy them quickly and efficiently, in ways that are beneficial to both the consumer and the organization. This means that any marketing oriented organization should try to obtain information on consumer needs and gather marketing intelligence to help satisfy these needs efficiently. Marketing research is a critical part of such a marketing intelligence system. It helps to improve management decision making by providing relevant, accurate and timely information. Every decision needs information and relevant strategies are based on the information collected through marketing research in action. Marketing research is an information input to decisions.

Marketing research uses two sources of data: primary and secondary. Primary data are gathered through ad-hoc surveys specifically built for the purposes of the study. Secondary data are assembled by accessing available information like previous studies, official statistics etc. Secondary data are information from secondary sources, i.e., they are not directly complied by the analyst. They may include published or unpublished work based on research that relies on primary sources of any material other than primary sources used to prepare a written work. Secondary data are collected by persons or agencies for purposes other than solving the problem at hand.

Use of secondary data involves many advantages (they are one of the cheapest and easiest means of access to information; there is huge amount of available secondary data) but also there are many limitations, problems and concerns in application of secondary data that has to be carefully and thoroughly considered and analysed; otherwise, the complete analysis can be seriously invalidated and its results will not be valid and reliable.

2. Sources and properties of secondary data in marketing research

Secondary data have been gathered by others for their own purposes but the data could be useful in the analysis of a wide range of marketing and consumer research. Secondary data are generated by means of primary data gathering techniques. One person or entity’s primary data become another person or entity’s secondary data. For example, when census data are used in a market evaluation for a market study they become secondary data. Marketing analysts use many data sets that are gathered as primary data. Any demographic and economic data generated by any government agency (federal, state, or local) for whatever purpose they need, necessary become someone else’s secondary data. Marketing researchers and analysts need to understand the issues involved in primary data gathering because the secondary data they use from published sources were previously gathered as primary data.

There are primary sources of secondary data and secondary sources of secondary data. Census publications are primary sources of secondary data. When these secondary data are taken into a process that augments, modifies, summarizes, synthesizes, updates, or in any way manipulates the data, the output of that process is a secondary source of secondary data (Rabianski, 2006, pp.44).
The significance of this distinction is simply that each time data passes through a process, the chance of error increases.

Secondary data are significant input into marketing analysis. They can be used by researchers in many ways. A company’s internal records, accounting and control systems, provide the most basic data on marketing inputs and the resulting outcomes. Data on inputs can range from budgets and schedules of expenditures to salespeople’s call reports. Data on outcomes can be obtained from the billing records on shipments maintained in the accounting systems. In many industries the resulting sales reports are the most important items of data used by marketing managers. Published data are the most popular source of marketing information. The major published sources are the various government publications, periodicals and journals, and publicly available reports from such private groups as foundations, publishers, trade associations and companies. Of all these sources the most valuable information comes from the government census information and various registration requirements that include births, deaths, marriages, unemployment records, export declarations, automobile registrations etc.

Secondary data might be available at no cost to the public or obtained by membership in a trade organization or through a subscription to its publications. Often private data sources provide summarized versions of raw data collected by organizations involved in various types of research. These groups generate and make available the data as a by-product of their work. In other instances, the private organizations compile and market a database specifically to supplement secondary data available from public sources.

Secondary data is often available from both the original source, which collects and organizes the data, and from sources that simply summarize data collected by others and market the information. For example, the original source of secondary data for population characteristics is the Census of population. When data is obtained directly from census publications, all the backup information is provided about data collection techniques and statistical methodologies used, possible inaccuracies, and other valuable background inputs.

There are some important properties that all secondary data should have if researchers tend to have reliable results of the analysis. Secondary data should be accurate, reliable, precise, unbiased, valid, appropriate and timely. Accurate data must accurately reflect what is being studied. Accurate data reflects the true population parameter. Reliability refers to this: If the same variable is measured several times, the data is reliable if the estimates are approximately the same. Bias is the deviation of a statistical estimate from the true parameter which is estimated by the selected statistical procedure. It is systematic error introduced into an analysis by the failure to follow proper procedure or by other errors in the database. Validation is the process of checking to make sure the proper procedures were followed in collecting, organizing, and analyzing the data. Data that has been validated is considered more accurate because more is known about its origin and characteristics. Consequently, more confidence can be placed in the use of validated data. The researcher is also concerned that the data are appropriate. They must measure what they are supposed to measure; the sample must be taken from the correct population. The data must reflect the time period that governs the analysis.
3. Errors and bias in secondary data

Secondary data exposes the analysis in which it is used to a variety of possible errors and bias, but precautions are available to deal with them. The analyst will not be able to remove or overcome some of the errors, but knowledge of their existence will help in drawing informed conclusions and establishing some level of confidence in the judgments that result. Secondary data should be checked for errors to verify their accuracy. If such validation cannot be accomplished, then secondary data should be regarded as suspect. Whenever it is possible, visual and statistical techniques should be employed to eliminate errors or to explicitly take them into account. The researcher tries to obtain accurate data by reducing the error in the research. There are 4 categories of potential error in secondary data (Rabianski, 2006, pp.49):

1. Sampling and non-sampling errors
2. Errors that invalidate the data
3. Errors that require data reformulation
4. Errors that reduce reliability

1. Sampling and non-sampling errors. Sampling error raises statistical issues surrounding the sample selection. It is a portion of survey error that is imputed to sampling, that is, observing a subset of the target population instead of collecting information on the whole set. Sampling error occurs when the sample chosen by the researcher doesn’t accurately reflect the total population that is studied. In a statistical sense, sampling error happens because the sample is not a random sample of a population meaning that each element of the population doesn’t have an equal chance of being selected. This error could arise if the population is stratified but the sample represents only one or some of the strata. For example, if the true population is bimodal regarding the age, but the sample contains only the young, there has been a sampling error. Sampling error can be dealt with through probabilistic sampling which allows measurement and control of this error component.

However, in many situations the sampling error is quite low compared to non-sampling and potentially systematic errors especially when the proportion of non-respondent is high. This is a serious threat to the successful completion of data collection (Assael and Keon, 1992, pp.119). Contrary to sampling errors, they cannot be quantified prior to the survey and sometimes it is difficult to detect them even after the field work. They represent one of those risks where a prevention is definitely better than cure. Prevention is based on an accurate inspection of the issues emerging during data collection. Non-sampling errors arise from problems during the observation or questioning phase of primary data gathering. Five general types of non-sampling errors could arise in this phase: frame error, measurement error, sequence bias, interview bias and non-response bias (Mazzocchi, 2008, pp.48).

Frame error occurs when the list that the analyst generates to represent the population omits certain individuals whose opinions, attitudes or other characteristics will not otherwise be represented. For example, a telephone survey cannot contact people who do not have access to a telephone. An email
survey cannot contact people who do not have access to a computer. Frame error can also occur when physical entities are considered.

Measurement error, i.e. response error arises when the individuals who respond to the questions give information that is not true. For example, if the total number of shoppers in a store is misinterpreted as the number of buyers (the true customers), the estimate of total sales volume and of value would be overstated.

Sequence bias occurs when the order of the questions on a questionnaire or in an interview suggests or induces an idea or opinion in the mind of the respondent as a direct consequence of the manner in which the questions are sequenced.

Interviewer bias occurs because of the presence or influence of an interviewer in a face-to-face or telephone interview. The interviewer might unknowingly bring out an untrue response to sensitive questions, e.g., the respondent may produce an answer to please the interviewer instead of answering truthfully or the interviewer might record a verbal response incorrectly because the statement is interpreted with the interviewer’s bias. Interviewer bias can also occur if the interviewer asks questions that are designed to generate a given response, i.e., leading questions.

Non-response bias occurs because individuals in the sample do not respond even though the analyst tried to contact them, or when individuals do not answer certain questions.

2. Errors that can invalidate data. Secondary data might be contaminated and represented as invalid because of actions or attitudes of the person(s) or the orientation of the organization that is collecting data. Data might reflect manipulation, contamination caused by inappropriateness, confusion or carelessness, or concept error (Iacobucci & Churchill, 2009, pp.201).

Errors caused by manipulation. The organization gathering data might manipulate or reorganize the data to meet a purpose that is unknown to others\(^2\).

The data could have been reorganized so that the collecting agency could show that its organizational goals were met. Similarly, the data might be manipulated to generate adverse conclusions about situations that the collecting agency opposes. If any such manipulation occurs, or even if there is a reasonable suspicion that it has occurred, the data should not be used.

Errors caused by inappropriateness, confusion or carelessness. Organizations might collect, organize, and distribute data without properly specifying the particulars of the collection process, their data assembly procedures or any data synthesis that was used. They also might not care about the data’s quality and validity. The organization’s staff may not know how to collect data.

\(^2\) In PR theory it is possible to find ideas considering manipulating data for the purpose of creating events, i.e. news. It is, for example, emphasized that “some news really happen, others are created” (Wilcox et al., 2005, pp.111). Furthermore, these theorists point out that “successful marketing and PR practitioners have to do more than just producing competent, accurate press releases on routine activities of their clients or employers. They have to use imagination and organizational skills in order to create events that attract the interests of media. Created events have an important role in corporate campaigns, but can also produce organized attacks on a company by the media, where activists try to redefine image and reputation of target company”. These attitudes clearly show in what measure data manipulation is common and present in today’s business, and that, for marketing purposes, there are situations when purposely manipulated data has an advantage over accurate one.
Whenever inappropriateness, confusion or carelessness is suspected, the researcher should not use the data. This situation can occur in organizations that have a well-defined primary function and collect data only as a secondary activity.

**Concept errors.** Concept errors represent a broad class of error that can significantly invalidate data. Data containing concept error can still be used, however, if the analyst can obtain information about the nature of the error. Concept error is defined as the error that arises because of the difference between the concept to be measured and indicator, or specific item, that is used to measure that concept (Berry & Linoff, 1997, pp.87). There are many indicator variables in market analysis that are surrogates for the data that the analyst cannot obtain. For example, the analyst may be seeking information about household income, which includes salaries, rental income, interest income and dividends. The indicator used to measure household income might report only salary data. In this case, the indicator contains a large component of household income but does not include all sources of income that the household can receive. Use of this indicator variable might cause only a small error among households which are headed by salary earners but it would cause a large error in retirement community. An error can result from an indicator variable that does not show complexity of the concept variable (Dehmater & Hanckok, 2001, pp.108).

Finally, market analyst sometimes try to measure the purchasing power in a retail trade area by multiplying the number of households by the median household income. Here, the appropriate income measure is the mean household income. Median income distorts the measure of purchasing power in an income distribution that is skewed. Concept error can, but does not necessarily, invalidate the data and the analysis. The analyst may decide to use the data even though concept error is present and handle it with a variety of techniques. The decision to use data depends on the following considerations (Houston, 2004, pp.159):

- The size of the discrepancy between the concept and the indicator. If the size of the discrepancy is small and the indicator responds similarly to the same casual factors that affect the concept, the data could be used
- The purpose of the analysis. An exploratory study is able to tolerate larger errors than a study that is designed to rest fairly explicit hypotheses
- The availability of valid and accurate data. If accurate data exists, it should be used. But, this statement must be tempered by a recognition of the cost of accurate data and the time constraints under which a study is being made. If accurate data is costly and cannot be obtained in a timely manner, the analyst might decide to use data that contains some degree of concept error. In such instances, the analyst should understand the nature of the concept error as well as its magnitude and direction.

In any event, the analyst needs to be aware that concept error may be present in the data used for the analysis. In too many instances the existence of concept error is never realized and its possible effect on the study is never considered.
3. **Errors that require data reformulation.** Secondary data is sometimes not directly useful to the analyst because it does not adequately measure the concept being studied. Errors commonly result from the following four types of situations (Patzer, 1995, pp.78):

- Changing circumstances
- Inappropriate transformations
- Inappropriate temporal extrapolations
- Inappropriate temporal recognition

**Errors caused by changing circumstances.** This type of error is caused by a change that affects a data series but is not readily apparent in that data series. For example, a change in the geographic boundaries of statistical areas can occur when official statistical institutes adds a county to the statistical area. A change in the underlying unit of measurement can also occur. For example, a data series that presented monthly statistics is now presented on a bimonthly basis. For the sake of consistency, the analyst might need to choose one or the other presentation format. The analyst must either combine previous monthly data into bimonthly groupings as currently used or split the current bimonthly data into monthly statistics. The unit of measurement could also have changed because of a shift in the collection time period. An error can also arise because the concept being measured is redefined over time and across space.

**Errors that arise from inappropriate transformations.** Original data is often presented in a secondary data sources in categories that were created to make the data more presentable in a tabular format, or the original categories do not reflect the analyst’s needs to handle the task at hand (Biemer et al., 2004, pp.431). Moreover, data may be presented as a ratio that made sense for their original purpose but do not make sense in the context of the analyst’s current study. The indicator variable can use the wrong base measurement. Secondary data can be presented in groupings such as household income, which distribute the population characteristic. The categories can change from one reporting to another. Unless the analyst transforms the data, the analysis can be flawed.

**Errors from inappropriate temporal extrapolations.** Secondary data are often not available for the intervening periods (months, quarters or years) between published reports. Data for intervening periods have to be interpolated from the two nearest reporting years. An example is the situation in which the analyst needs population information for a specific census tract for 1998, but the secondary data exist for only 1995 and 2000. Using only these two points, the interpolation for 1998 can be made as a straight line or as an exponential rate of change at an increasing or decreasing rate. Without knowing the true path of change between these two points, any one of three answers can be obtained for the 1998 figure. Typically, the interpolation would be made using an average, annual, straight-line rate of change between 1995 and 2000. The shape of the curve showing exponential change can be estimated by analyzing another related data series in which the variable moves in approximately the same direction and the same magnitude.
Errors from inappropriate temporal recognition. The most common error of this type arises from a misunderstanding of the time dimension of the secondary data. For example, if data are used in the year of their publication instead of the year for which they were gathered. There is always a time lag between the time the primary data are gathered and the time they are made available. A 2005 publication for instance, more than likely contains data gathered at an earlier date such as 2004; in some instances the time lag can be even longer.

4. Errors that reduce reliability. A data set is reliable if successive counts produce the same results. As explained before, reliability is not accuracy; the data set is accurate only if it is free from procedural and measurement errors. An inaccurate data set can be reliable if it maintains the same degree of inaccuracy. The reliability of data is a function of the organization that gathers, organizes, records, and publishes the secondary data. Several issues should be considered when evaluating the organization that is collecting and disseminating data such as whether data collection is the stated purpose of the organization or merely a secondary or adjunct function. Another issue is whether the individuals and staff that undertake data collection are trained and experienced in data collection procedures. The analyst should also determine if the organization has adequate resources to do a thorough job. Errors causing the analyst to question the reliability of the data can be divided into three categories (Baumgartner & Steenkamp, 2006, pp.438):

- Clerical
- Changes in collection procedures
- Failure to use correct data

Clerical errors. Clerical errors are a frequent occurrence and they happen even to the most careful people. To detect the existence of clerical errors, the data might be displayed in an easily comprehended manner (e.g., a scatter plot diagram or a simple table). In this way, outliers can be detected more easily. This procedure will allow the analyst to notice the misplaced decimal, the added zero, or the extra digit. Another error of this type is the transposition of numbers in a series with the same number of digits. A plot of the values will allow the analyst to catch this outlier error.

Errors due to changes in collection procedures. When error results from a change in collection procedures, the generated data may be quite different from previous data in the same data set. This error can arise because of different methods of collection or different circumstances surrounding the collection. For example, the time of collection (time of day, day of week, season, year, etc.) might have changed. The manner in which the data is summarized might also change. The use of the scatter plot or a simple review of the raw data could reveal discontinuity or a jump in the data points caused by the change in the collection procedures.

Errors due to corrected data. Data can be inconsistent from one report to another in the same published series because of errors that have been discovered, corrected and then reflected in subsequent versions of the data set. Most often these are clerical errors. The analyst needs to use the most recent version to reduce errors. Also, if possible, the analyst should know when data is checked and when a clean version of that data is printed. When using secondary data that has been
reorganized at some point, always check it against the newest versions of that data set. Another situation of data series correction occurs when secondary data providers are able to adjust their prior estimates or forecasts for a decennial year against actual census numbers. The previous estimates or forecasts are adjusted and corrected in subsequent publications. It is clear that data should always be checked against the newest versions of the data set.

4. Conclusion

"He who makes no mistakes makes nothing" is a saying that can definitely be applied to the issues of errors in secondary data used in all fields including marketing research. It is obvious that errors in secondary data are unavoidable. However, marketing analysts can undertake many actions in order to minimize these errors which will produce sufficiently accurate, reliable and appropriate secondary data. Sampling errors can be dealt with through adequate sampling process which allows evaluation and control of this error. Since non-sampling errors cannot be quantified prior to the survey and since it is rather difficult to detect them, the best tool for dealing with this type of errors is prevention that is based on an accurate inspection of the issues in the process of collecting the primary data. Analysts should also pay attention to other sources of errors (errors that invalidate data, errors that require data reformulation and errors that reduce reliability) that can seriously weaken data and therefore influence results of the whole analysis. Although there are many different sources of errors in secondary data, as presented in this paper, there are also many procedures for dealing with them which will result in more valid findings of the research that will provide reliable basis for the marketing decision making process.

References


Abstract

Istraživanje marketinga koristi dva izvora podataka: primarni i sekundarni. Postoji mnogo prednosti u korišćenju sekundarnih podataka, ali postoje mnoga ograničenja, kao što su različite vrste grešaka i predrasuda koje mogu nastati u ovim podacima. Sekundarni podaci treba da budu tačni, pouzdani, precizni, nepristrani, važni, odgovarajući i pravovremeni. Četiri kategorije potencijalnih grešaka mogu smanjiti tačnost sekundarnih podataka: utorkovanje i neuzorkovanje greške, greške koje poništavaju podatke, greške koje zahtevaju redefinisanje podataka i greške koje smanjuju pouzdanost. Svi izvori grešaka mogu smanjiti pouzdanost i validnost rezultata. To podrazumeva da se sekundarni podaci moraju pažljivo tretirati.

Keywords: Sekundarni podaci, greške, pristrasnost, tačnost.