winter 2012, Volume 2, Issue 4

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by staff at StatsWhisperer.™ For many more free resources in learning statistics, including webinars and subscribing to this newsletter, visit us on the web at: www.statswhisperer.com

Introduction to this Issue

In order to perform a data analysis, you first need data analyze. The first inclination of many a researcher is to gather data themselves to use in their study. However, gathering data can be expensive and time consuming. Also, if resources are limited, it is difficult to gather data upon a large sample or pool of variables.

The use of secondary data (data gathered by other researchers) in a data analysis is often a viable alternative to gathering one's own data. Besides saving a great deal of time and money, this option may enhance the quality of an empirical study.

In this newsletter issue we will discuss the advantages/disadvantages of using secondary data, as well as where to find secondary data sets. You can also use our mnemonic O.P.D.™ (Other People's Data), taken from the old 1990's OPP song, to remember this useful resource.

INSIDE THIS ISSUE ON THE USEFULNESS OF SECONDARY DATA (A.K.A., O.P.D.™)

Defining Secondary Data

Secondary Data simply refers to data gathered by someone other than the user. *Primary Data* refers to data gathered by the investigator performing the research.

So here we have a bit of a reference to the old line in Romeo and Juliet "**A rose by any other name would smell as sweet.**" Specifically, it is the same data, just by another name. And yes, **the data by any other name would still yield the same results.**

Where are some common sources of secondary data? There are actually many! Some of the most common are large government-funded datasets, university/colleges, and author's websites.

Yes, author's websites, meaning private researchers can post their own data! Of course, you must be scrupulous about data quality, but this suggests that you could collect and post your own data and **be a source of secondary data yourself!**

Advantages and Disadvantages of Using Secondary Data

Doubtlessly, one of the great advantages of using secondary data in contrast to primary data is all the time, effort, and money saved. Figure 1, while a generalization, illustrates the differences in the amount of time, money, and effort required when using primary or secondary data.

In the right hand column, we see that one could conceivable decide to use secondary data, find a data set that matched the study they had conceived, and begin analysis that day! The time may be extended a bit if the researcher needed IRB approval and/or needed to search several days for the appropriate secondary data set.

On the left hand column, we see the tasks required if one decided to collect primary data for their study. Of course, the estimates of time for each task would vary significantly depending on the size of the study. But in many ways the estimated amount of time needed is quite conservative, especially in the tasks of data collection. Data collection may go on for months or even years!

Those who have traversed the tasks listed in the left hand column will attest that it is not just time that is expended in this process. It is also a great amount of effort and attention in attending to planned and unplanned details and tasks included in the process.

Even though Figure 1 is an oversimplification of the research study process, I believe the point suggested is quite clear. Using a secondary data set can make your life much easier!

Figure 1. Study Tasks Using Primary and Secondary Data

Tasks When Gathering Primary Data

1) Conceptualize Study (Time?)

2) Gather Team/Resources (6 months)

3) Write Grant (6 months)

4) Pass IRB (2 months)

5) Data Collection (#points=months/years)

6) Data Entry (4 months)

7) Begin Statistical Analysis

Tasks When Using Secondary Data

1) Conceptualize Study (Time?)

2) Locate the Appropriate Dataset (1 week)

3) Pass IRB (if needed 2 months)

4) Begin Statistical Analysis

Secondary Data and Data Quality

Sure, using secondary data might make your life easier, but at what price? Specifically, are you sacrificing data quality to make life easier? Is it some sort of cop out?

Not necessarily, in fact the use of secondary data may in fact enhance the quality of your study significantly!

To illustrate this, let's take the case of a past co-worker of mine. He needed to complete his dissertation study, which examined level of depression within a specific NYC Community. His plan was to collect data on 100 residents of that community using a snowball sampling method.

For those not familiar with snowball sampling, it is the practice of locating potential study participants and having them refer you to other potential study participants they know of.

I call this method the *Faberge Shampoo* method of sampling. In the 1970's there was a commercial for this brand where the user liked the shampoo and reported that she "told two friends, and they told two friends, and so on and so on." Each time she/they reported telling two friends the screen would divide into more people. It is rather like a snowball sample of users of that brand of shampoo.

What is the flaw in his sampling method? It relates to the old "Birds of a feather flock together" rule. Specifically, it could be that those with high levels of depression are likely to associate with others who have high levels of depression (and vice versa). So his sample might be biased in either having very high or very low levels of depression.

In other words, his sample was likely not to reflect the actual levels of depression of residents within that NYC community. What his sample was likely to reflect was a segment of that population that was likely biased in having high or low levels of depression.

So what to do? As a graduate student he did not have a great deal of time or money to devote to collecting a rigorous random sample of community residents.

Secondary Data and Data Quality (continued)

Subsequently, I asked him if he was aware that one of his professors had been funded by the National Institute of Health (NIH) to collect community level data (that included his target community) as part of a prevention program. The rest as they say is history.

The NIH study conducted by his professor was a rigorous study that involved a random selection of study participants, a much bigger sample than he had planned on recruiting, and a long list of study variables.

Ultimately, through using secondary data, he produced a much better quality dissertation project.

The Disadvantages of Using Secondary Data

In research and statistics whenever you opt for the benefits of one method, you are giving up the benefits of another. Of course, opting to use secondary data over primary data is no exception.

What is the biggest sacrifice made when you opt for using secondary data? Control! Specifically, if you were to gather primary data, you could dictate what type of data would be gathered, how it would be gathered, and so on. However, when you use secondary data you must use the study and variables presented. This could result in problems such as:

1) **The study variables may not fit your framework or they may lack depth**

   For example, if you have a conceptual framework and the secondary dataset you have is missing a variable of two, you cannot change that.

   Also, if a variable is measured superficially, such as with a single item, rather than a full scale, you cannot change that either.

2) **Accuracy not known/Methods unclear**

   If you gather your own data you will likely have a degree of control over how that is accomplished. However, regarding another dataset, you may not fully approve of or be able to discern the methods used to gather data.

3) **Some fields/depts. don't value secondary data**

   Some fields and departments may not value the use of secondary data as much as others. If this is the case the smart thing to do is typically opt to gather primary data.

4) **May be outdated**

   Human society is changing so rapidly, that data are likely becoming outdated at a rate faster than ever before. So even if you found your perfect dataset, with all the variables you need, that are measured precisely, if the data are 20 or even just 10 years old, you must wonder if they are outdated. If the data are outdated, it could be said that they may not reflect the existing study population you are examining.

Where to Find Secondary Data

Sources of secondary data are actually quite easy to find. Although I have never seen a compiled list of the mass of sources of secondary data, your classic google search will provide you with a great many leads.

I have listed some of my favorite sites for accessing secondary data in Table 1 below. The best way to learn about these sites is not to be told about them, but to access them and have a look around.

For example, the first link takes you to the ICPSR, which is part of the Institute for Social Research at the University of Michigan (one of the best sources of data in the country). At this site you can browse data by keywords, topic, etc….

The last link is to "Sodapop," which is an acronym for Simple Online Data Archive for POPulation studies (hosted by Penn State University).

When you access Sodapop, you will see a long list of studies. Next to many of them are the words "Data Download Available." Thus, you can scroll up and down the list of studies and know immediately which studies can provide you with data for your analysis.

Please note that these types of sites can be rather overwhelming and pull you in many directions. A good strategy to combat this is to write down some keys words describing the type of secondary dataset you want before you access the first site.

Table 1. Sources of Secondary Data Sets

1) **Inter-University Consort. for Polit. & Soc. Research:**
   http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/index.jsp
2) **Data.gov:**
   http://www.data.gov
3) **National Center for Education Statistics:**
   http://nces.ed.gov
4) **U.S. Census Bureau:**
   http://www.census.gov
5) **Simple Online Data Archive for Pop. Studies:**
   http://sodapop.pop.psu.edu/data-collections

Examples of Large Widely Used Secondary DataSets

Table 2 presents a few examples of large widely used secondary data sets. Many of these types of studies have websites devoted entirely to presenting the details of the project, as well as making the data available in several formats.

This is essential as <u>if you are going to use a secondary dataset, you must familiarize yourself with the original study</u>. Specifically, you must examine the study questionnaires, interview protocols, study design, study aims, etc. <u>Optimally, you should understand that study as if you had gathered the data yourself!</u>

As an example, within Table 2, please note underlined dataset number 3, the General Social Survey (GSS). A description of this project, as well as all data gathered can be accessed at: http://www3.norc.org/gss+website/

When you access this website, you will first read a study description "The GSS contains a standard 'core' of demographic, behavioral, and attitudinal questions, plus topics of special interest." There are also a great many subpages and links listed.

Of particular interest, in the center of the *Home* page, you will see one option for *Data* and one option for *Documentation*. Under the *Data* option, you may actually download survey data (for SPSS or SAS) and begin actual analysis! Under the *Documentation* option, you can access codebooks, reports, questionnaires, and a bibliography of studies conducted using the GSS data!

In essence you have the data and documentation neatly organized for your use. Furthermore, you even have samples of other published studies. Needless to say, these sites and datasets can be a tremendous resource!

Table 2. Examples of Widely Used Secondary Data Sets

1) **National Education Longitudinal Study (NELS)**

2) National Household Education Surveys (NHES)

3) **<u>General Social Survey (GSS)</u>**

4) **Current Population Survey (CPS)**

5) **Monitoring the Future (MTF)**

Final Comments

After one considers all we have discussed in this issue of our newsletter, there is still one important point to make. The mass of people do not use secondary data sets for two reasons.

First, many people simply do not realize these datasets exist. I can still remember being told about the availability of secondary datasets after I had received my Masters Degree. What is surprising is that I was a research concentration major in that Masters program and it was terrific program! So I can understand how most people are unaware of these datasets.

Second, many people believe these datasets are difficult to find, access, and use. My hope is that our short newsletter has somewhat addressed these potential barriers.

So if we break down this issue, what have we learned? I would say, we have established that secondary datasets offer advantages and disadvantages relative to gathering and/or using primary data.

We identified several sites that list and make secondary datasets available. We also identified specific widely used available secondary datasets. Lastly, we visited the website devoted to one of these studies, the GSS, and noted the resources already available to us as researchers at this website.

One could say that this newsletter was as much about becoming aware of the resources already available to us, as was a discussion of the usefulness of secondary data. But that did not fit in the title!

Dr. William M. O'Bannon, Jr., the founder and CEO of StatsWhisperer can be reached at:

wb@statswhisperer.com

Thanks for your interest in our newsletter!

REGISTER NOW!

Our 6-Week Webinar Class on

The Logistics of Statistics:

<u>The Steps of Data Analysis</u>

This class instructs participants on:

1) The specific 6-step process involved in conducting a data analysis project (i.e., what does one do 1st, 2nd, 3rd, and so on)
2) The essential "need to know" facts within a data analysis project

The goal of this class is to enable participants to conduct their own data analysis project, as well as critique and recognize high quality research (in peer-reviewed journals, research reports, presentations, etc.).

The next class will begin on Apri1 1st 2013 and be completed on May 16th 2013.

Seminar information and registration can be accessed on our website at: http://www.statswhisperer.com/seminars/