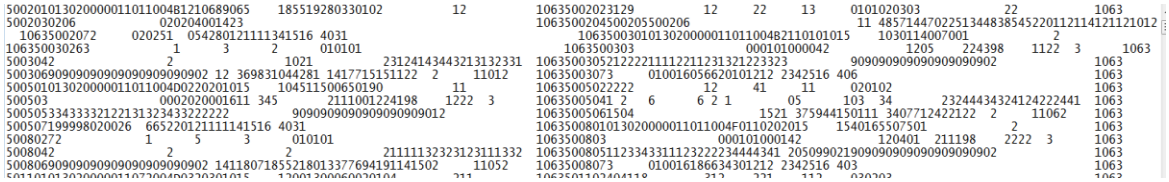
About Social Science Data

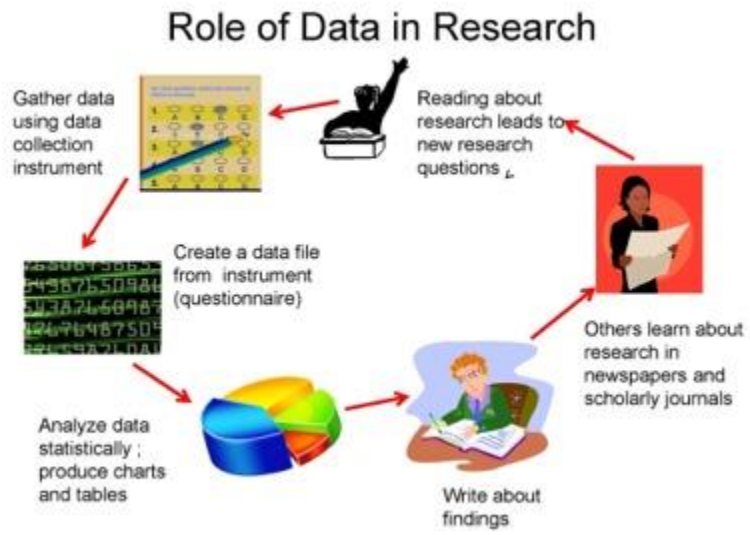
What is Social Science Data?

Social Science Data Files are primary source materials encompassing raw data files from surveys and textual or electronic format documentation, normally called *codebooks*. Data files are not *eye-readable*. A *raw* data file is composed of characters or numbers on which mathematical operations are carried out using a statistical program (for example, using SPSS, SAS, STATA or R), on a computer, and interpreting the results of statistical processing. If you printed out a raw data file, it would look like



How are data files used in research?

Below is a visual on how data is incorporated into the research process.



Secondary Analysis

Since the early 1940's social scientists have collected and analyzed data in surveys as part of a *quantitative* approach to research. When you use data from an archive of previously collected surveys you are engaged in ***secondary analysis*** of primary information. Secondary analysis is a great way to do research, especially if you are a student. Designing your own survey is not trivial and the cost to carry out even a simple survey can be quite prohibitive in terms of cost. One of the ways that secondary analysis of data is supported is through the Berkeley Initiative for Transparency in the Social Sciences (BITSS).

Replication
Beyond cost, there are other excellent reasons to do research using previously collected data. For example, you might read about the results of a survey on a subject of interest to you. The public opinion polls conducted by media organizations such as the New York Times or CBS News are examples of such surveys. If the data are available from an archive, (such as ICPSR) you might decide to analyze the data with the same statistical procedures used by the original data collectors. By ***replicating*** the work of another researcher, you can test the validity of the previous analyses.

Test Hypotheses
Data files are used to ***test research hypotheses*** about people's attitudes and behaviors. For example, you might want to test your idea that a higher level of education leads to a higher income level. Comparing years of schooling with a person's wage or salary, is one way to better understand the relationship between education and income attainment. Of course this is a very simple comparison and there are many other factors to consider. If you are interested in this topic, take a look at the data available from High School and Beyond: A Longitudinal Survey of Students in the United States.

Evaluate Methodology
There are many methods used to conduct surveys. Questionnaires can be organized in a number of ways, and the design can affect how people respond to questions. Some social scientists analyze similar data collected in several surveys to understand how ***different survey methodologies*** change the kind and quality of data collected. Other researchers study the methods used to identify sample populations, and still others are interested in how the interview setting can affect people's responses. For example, there may be differences depending on whether the questionnaire is sent by mail, is conducted on the telephone, or if the survey is conducted through the internet or using a mobile device. Data are gathered in the UCLA Los Angeles County Social Survey by telephone. The Hispanic Health and Nutrition Examination Survey is conducted by personal interview. Questionnaires are sent by mail in conducting the U.S. Decennial Census.

Compare Population Groups
Research using surveys is useful for studying the attitudes and behaviors of ***different population groups***. For example, you might want to study the music preferences of teenagers vs those who are age 65 and above. Or, perhaps you have an interest in understanding the differences between Republican and Democratic voters. Populations from different countries and cultures can also be studied using secondary analysis techniques. Some examples of such studies are the American National Election Study (ANES), the General Social Survey (GSS), or the International Social Survey program (ISSP).

Longitudinal Analysis
In another area of social science research, changes in attitudes and behavior are studied over time. There are two approaches commonly used. In ***longitudinal analysis***, the same respondents are studied repeatedly for a specified period of time. Some data files in the Archive are longitudinal studies of high school students, of the occupational change of members of the same households, or of changes in the amount and kind of income people receive over time. Examples of these are National Education Longitudinal Survey (NELS), Panel Study of Income Dynamics (PSID), or Survey of Income and Program Participation (SIPP).

Cross-Sectional Analysis
A second approach is called ***cross-sectional analysis***. Using this method, people are interviewed only one time, but the same questionnaire is used repeatedly, with different respondents, over a specified period of time. Here, the method used in determining the group of people or sample to be interviewed is of great importance. Special

CC BY NC SA

care is taken to be sure the groups are similar in number, gender, race/ethnicity, age, and in geographic location. In cross-sectional studies, changes in attitudes and behavior are studied by analyzing data gathered at discrete points. Examples of data files which can be used for cross-sectional analysis are The California Polls, the Current Population Survey: Annual Demographic File, and the National Survey of Family Growth.

Social science research uses many terms which may be new to you. For more information on terms such as are used in this web site, you might want to explore a glossary created by Jim Jacobs, Data Consultant.

When should I use data instead of a printed resource?

Data files do not always need to be used to answer research questions. In fact, you may want to use a range of data files, published statistics, and reports for your research. Information from raw data files is often already processed into published tables and explanatory reports. Tables and reports can be found in publications stored in libraries, and increasingly, over the Internet. Many numeric information resources are available through online tools for creating customized tables. An example of an online tool is the Census Bureau's American Factfinder. However, for detailed answers to questions, statistically analyzing data can be the best approach. Here are some points to consider when deciding what kind of format will work best for your research:

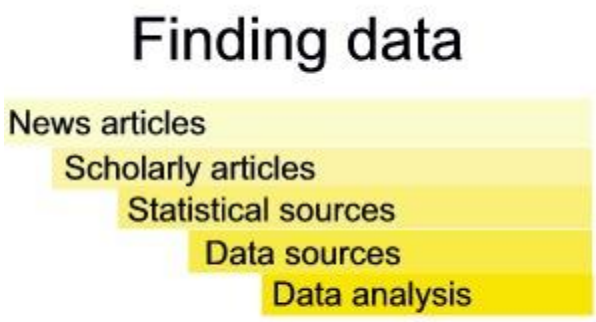
- Several layers of geography are easier to process in a data file.
- Small geographic areas, such as tracts or blocks are usually only found in data files.
- Data files are usually needed for longitudinal and cross-sectional analyses.
- Data files are best for detailed race and ethnicity tabulations.
- Printed reports will usually contain the most recent information.
- Historical information will often only be found in printed tables and reports.
- Online tools will usually generate aggregate tables as well as maps.
- Online tools are best for one or two specific geographic areas, such as two counties, or two cities.

For more information, be sure to review the sections of document on “What is a Codebook” and on “Searching for Data”.

How do I look for data?

There are different places you can look in order to find and use data. First you should to consider the type of data needed to address a research question. Next a review of the study documentation such as questionnaires and codebooks will help determine which data will useful. Here is a visual of ways to think about the kinds of sources where you can find data, followed by descriptions of each of these sources.

CC BY NC SA



News Articles
You can look for information in the popular press discussing recent studies. You can find news articles in indexes of newspapers and by using Lexis Nexis. The UCLA Library has access to LexisNexis – you can also search by various newspaper indexes, such as the New York Times or use periodical indexes. Check the UCLA Library website under databases and use the search term "newspaper".

Scholarly Articles
Sometimes we read about things in the news and we want something more extensive to use in research. For this, we can use the article databases to locate analyses and find research about studies discussed in the news. Example: Social Sciences Citation Index.

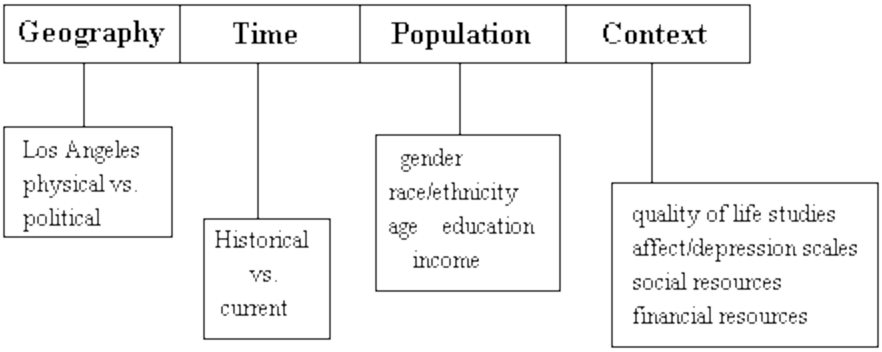
Statistical Sources
But maybe we want to find some more detail from the scholarly articles we read – sometimes information in tabular form – tables, charts, etc. can be useful. Government sites provide a great deal of statistical information – remember to use local, state, and national government sites. Example: the LA County Election Office you might use a statistical abstract for both the US and/or for individual states.

Data Sources
Sometimes we look at a chart or a table and we want to know about the data that created the table – so here we are talking about the raw data. The codebook or other materials that are the eye-readable guide to the data can be as useful as the data files.

Data Analysis
Finally, sometimes you won't have the data points you seek until you actually analyze your data with a software tool or statistical package.

Refining Research

You probably have in mind a general area or research topic you want to pursue. When you talk with one of the Archive staff we will help you further refine your ideas by narrowing in on four areas: Geography, Time, Population and Context.



Geography

Let's say your topic is about Earthquakes. Think about what geographic areas you'd like to cover with this topic. You should consider this in terms of the following:

- **Physical Areas:** Fault zones, mountain regions, water areas, etc.
- **Governmental Units:** Cities, counties, states, countries
- **Political Boundaries:**Congressional districts
- **Other Units:** School districts, hospital or fire zones, telephone service areas, water districts

Note: The Archive has reference sources to use in understanding geographic units and boundaries.

Time

Consider whether your research area covers recent events, or is historical or look at changes over a specified range of time.

- **Current**: Many data files are about a year old before they are released to the public. If you need very recent information, you will need to use other sources, such as newspapers and journal articles with tables or other quantitative information rather than raw data.
- **Historical**: Computers have been popularly used in survey research since the 1960's. If you need data for earlier time periods it may be more difficult to locate in an electronic format.

Population

What is the population or case group you wish to study? This can be a group or groups of people, particular events, official records, etc. In addition you should consider whether you will look at a specific sample or subset of people, events, records, etc.

Examples :

- You may be able to describe a sample of people buy specifying gender, race, age, ethnicity, etc.
- An example of events might be occurrences of earthquakes. You can describe a sample of these occurrences by looking at only those registering magnitude 5 and above on the Richter scale.
- Within the universe of all official records, you might want to look oat only birth certificates, perhaps for only those issued for children born to mothers under 15 years of age.

Context

Raw data can be used in a variety of ways. Continuing with the earthquake example, even after you have specified the geographic region, the time period, the population and sample you desire, you will want to look at the data from a certain perspective or within a context relevant to your area of study. This context will determine the type of data most useful for analysis.

For example, if you want to study people who have experienced high magnitude earthquakes in Los Angeles in the last 20 years, do you want to look at how quality of life is affected? Do you want to examine scales to determine emotional response to earthquakes? Do you want to study the relationship of quality of life and social or financial resources?

Next you will need to identify the units of analysis and the types of variables you want to use.

Units of Analysis

Here we want to know the unit of measurement in the data file. Examples might be:

- Persons
- Households
- Records
  - Official
  - Civil
- Events

Why is this important? Let's say you want to look at the income of each person in a household. A study which is organized with only household level of information will not be useful to you since it will not provide income amounts for individuals. Keep in mind that some data files have more than one type of unit of analysis or measurement.

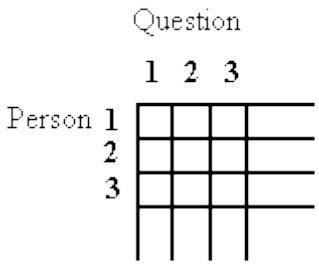
Types of Variables

Alpha: data that is in text format.	vs	**Numeric**: data that are coded responses to questions.
Continuous: represents a 'real' piece of information, such as age in actual years, or income in actual dollars.	vs	**Discrete**: data that are coded responses to questions.
Summary: data that have been tabulated or averaged or in some way combined such that single units of data cannot be evaluated.	vs	**Micro**: a data file containing information about individuals or individual units; organized for analysis by single units.

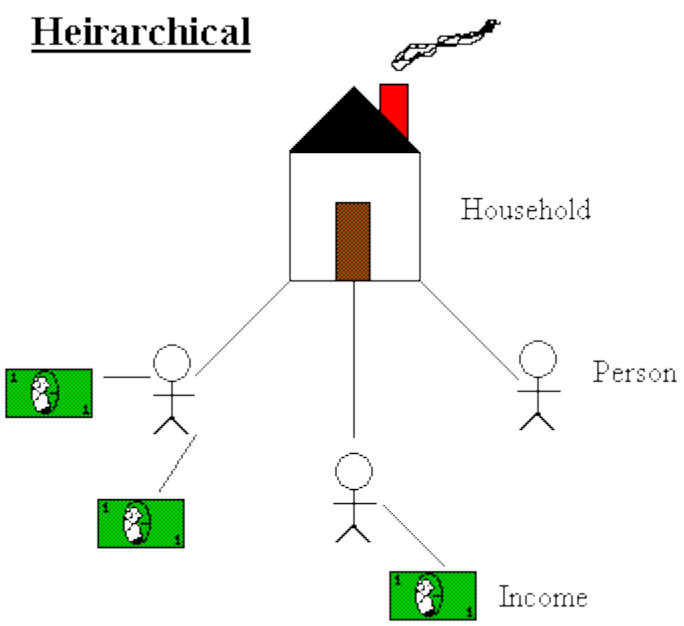
Data Structures

Finally, you need to study the codebook to determine the physical format of the file as it is arranged electronically. You will need to describe this arrangement in your statistical analysis. The three most common structures are: Rectangular, Hierarchical, and Relational.

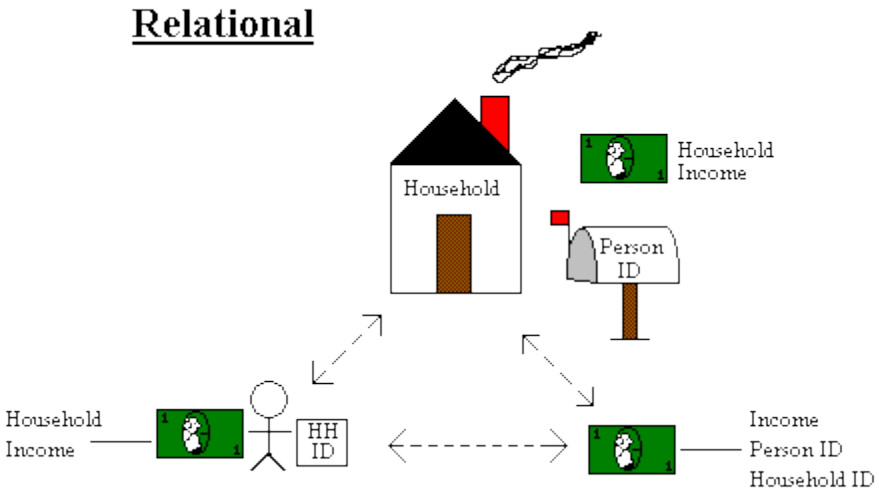
Rectangular



This is the most common form of data structure. In a survey, the answers given by each respondent are arranged in the same order. If the data were printed, it would resemble an array of persons and responses to questions, as illustrated.



These files are also described as having a tree-like structure. Data in this example are organized by household. Within households, it is possible to study individuals, and for each individual is possible to study sources of income. Data items are linked via their household identification.



The most common form of a relational file is one created using a database, such as Microsoft Access. In this example, there is a household file, a person file, and an income file. Each of these may be analyzed separately from the others. The files are linked by keys, or identification pointers. For example, there is an identifier for persons and income in the household file, and so forth.