



WORLD BANK  
ECONOMICS OF TOBACCO TOOLKIT

Editors: Ayda Yurekli & Joy de Beyer

---

Tool 2. Tobacco Data

# Data for Economic Analysis

Christina U. Ciecierski and Frank J. Chaloupka

# Contents

<b>I. Introduction</b>	<b>1</b>
Purpose of this Tool.....	1
Who Should Use this Tool.....	1
How to Use this Tool.....	1
<b>II. Key Information</b>	<b>3</b>
Defining Tobacco Use and Mortality.....	3
Smoking Prevalence.....	3
Conditional Demand.....	4
Assumptions and Requirements.....	4
<b>III. The Basics of Collecting Data</b>	<b>5</b>
What to Know about Data.....	5
Different Data Types.....	5
Aggregate Data.....	6
Where to Begin Looking.....	7
Central Data Collection Agency.....	8
Ministries or Departments of Government.....	8
Important Data to Collect.....	9
Consumption.....	9
Sales.....	11
Tobacco Price.....	12
Employment.....	14
Tobacco Production.....	17
Consumer Expenditures.....	18
Demographic Information.....	21
Economic Indices.....	21
Tobacco Trade Information.....	22
The Market for Tobacco.....	23
Examples of Market Analyses.....	23
Tobacco Price Measures.....	23
Tobacco Regulatory Environment.....	24
Aggregate Model Results.....	25
Multicollinearity.....	25
Bias.....	27
Simultaneity.....	27
Limitations from Units of Measure.....	27
Using Individual-Level Data.....	27
Consumption.....	29
Tobacco Price.....	29
Tobacco Taxes as a Proxy for Price.....	32
Measures of Income.....	33
Socio-Demographic Information.....	34

Use Caution when Interpreting Model Results Based on Individual-Level Data .....	34
<b>IV. Data Preparation and Management: Easy Steps to Building Your Own Database</b>	<b>38</b>
Choose a Software Package .....	38
Spreadsheets .....	38
Statistical Packages.....	39
Manipulate the Data .....	40
Read the Raw Data .....	40
Check the Quality of the Raw Data .....	45
Plot the Data (Aggregate Level Data Only).....	49
Create New Variables .....	49
Merge Datasets .....	50
Clean the Data .....	51
<b>V. Suggestions for Data Sources</b>	<b>54</b>
Economic and Tobacco Related Data .....	54
Helpful References for Aggregate-Level Data.....	57
United States.....	57
United Kingdom .....	58
Australia.....	58
Academic Publications and the Development of Tobacco Control Measures.....	58
Subnational Level Data—The United States .....	59
Household Survey Data .....	59
References for Individual-Level Data.....	60
Survey Data Sources.....	60
<b>VI. Conclusion</b>	<b>61</b>
Summary.....	61
Start-Point Data for Tobacco Control Research .....	61
Fundamentals of Data Collection .....	61
Addressing the Technical Aspects of Data Preparation.....	62
A Head Start on Potential Data Sources .....	62
A Few Final Words.....	62
<b>VII. References</b>	<b>63</b>

# List of Tables, Figures, and Boxes

## Tables

Table 2.1: Tobacco Product Categorization .....	24
Table 2.2. Codebook/Index of Column Codes .....	43
Table 2.3. Column Headings for an Excel Worksheet .....	44

## Figures

Figure 2.1. The Importance of Standard Definitions of Tobacco Product Size .....	4
Figure 2.2. Early Economic Theory of Price and Demand of Tobacco Products .....	13
Figure 2.3. Current Microeconomic Theory of Price and Demand of Tobacco Products .....	13
Figure 2.4. Tobacco Employment Resides within Four Industries .....	15
Figure 2.5. The Relationship of Tobacco Product Expenditures to Other Expenditures .....	20
Figure 2.6. An ASCII Data File .....	41
Figure 2.7. Code to Import an Excel File into SAS .....	44
Figure 2.8. Code to Import a Text File into SAS .....	45
Figure 2.9. Code to Generate a Set of Summary Statistics from a Raw SAS Dataset .....	46
Figure 2.10. SAS Window Showing Means Output .....	47
Figure 2.11. Code to Generate a Set of Value Frequencies from a Raw SAS Dataset .....	47
Figure 2.12. SAS Window Showing Frequency Output .....	48
Figure 2.13. Code to Generate Plots of an Aggregate Level Raw Data Set .....	49

## Boxes

Box 2.1. Question on a Consumer Survey Regarding Smoking Behavior .....	10
Box 2.2. Survey of Household Consumption Expenditures and Main Sources of Income .....	19
Box 2.3. Key Determinants of the Tobacco Regulatory Environment .....	26
Box 2.4. Questions Used to Capture Cigarette Consumption .....	30
Box 2.5. Open-Ended and Closed Questions About Consumer Cigarette Price .....	31
Box 2.6. Questions to Help Minimize Biased Estimates .....	31
Box 2.7. Questions About Per Capita Income and Household Income .....	33
Box 2.8. Questions About Proxies of Income .....	35
Box 2.9. Questions About Religion .....	36

# I. Introduction

---

## Purpose of this Tool

This tool provides a general introduction to “the art” of building databases and preparing data for scientific analysis. It addresses a number of issues pertaining to the search, identification, and preparation of data for meaningful economic analysis. It can best be thought of as a reference mechanism that provides support for the occasionally frustrated but endlessly hungry researcher working through the adventures of tobacco control analysis.

---

## Who Should Use this Tool

Anyone can reference this tool. It discusses the basic and fundamental aspects of data and databases—crucial knowledge and guidance for the sociologist, economist, public policy researcher, or other social scientist beginning new research in the field of economic tobacco control. Yet this tool can also be used as a reference for those seeking clarification in already familiar terrain, whether they be skilled economic researchers of the tobacco industry, epidemiologists seeking economic evidence to compliment their findings, or policymakers wanting to produce evidence supporting a political and economic agenda.

Further, this tool shows that similar and universal rules concerning data collection and data analysis exist, regardless of one’s culture, social traditions, or economic system. Following these rules, as outlined in these pages, will help every reader assure the integrity and analyses of their data and subsequent findings.

---

## How to Use this Tool

This tool addresses a variety of data issues as they pertain to the economic analyses presented in the other tools of this toolkit. This tool has been designed to follow a series of step-by-step discussions and examples. The tool follows:

1. A discussion of the different types of available data.
2. Identification of possible data sources.
3. Definitions and examples surrounding key variables required for the aggregate and individual level analyses presented in Tools 3 through 7.
4. Presentation of issues pertaining to data preparation and analysis including:
  - Evaluate and clean raw data.
  - Transport raw data into a statistical package.
  - Quality check the raw data (both missing observations and outliers).
  - Summarize raw data.
  - Plot raw data.
  - Recode survey data into usable form.

## II. Key Information

---

### Defining Tobacco Use and Mortality

#### Smoking Prevalence

*Smoking prevalence is equal to the number of people who smoke as a percentage of the total population.*

Smoking prevalence is the percentage of current smokers in the total population. When discussing prevalence and tobacco use, pay attention to the type of tobacco product that is being addressed with this statistic. The prevalence of smoking is the number of people who report smoking tobacco in the form of cigarettes, *bidis*, cigarillos, cigars, pipes, rolled tobacco, or other methods. A more comprehensive measure of tobacco consumption is prevalence of *all* tobacco use and includes the prevalence of smoking behavior *plus* the percentage of people who chew tobacco or use other forms of smokeless tobacco.

$$\text{Prevalence of Smokers (in \%)} = (\text{Number of smokers in the survey population} \div \text{Total survey population [smokers + non-smokers]}) \times 100$$

Example: A nationally representative survey of smoking behaviors captures the responses of 750 women and 600 men. The survey yielded 250 female and 400 male smokers. We can conclude that:

$$\text{Female Smoking Prevalence is: } (250 \div 750) \times 100 = 33.33\%$$

$$\text{Male Smoking Prevalence is: } (400 \div 600) \times 100 = 66.67\%$$

**Note:** Estimate prevalence rates *separately* for females and males, as the sample must accurately reflect the population's gender distribution. That is, the gender-specific prevalence rates must be weighted to reflect the actual gender composition of a population.

To report a country's total smoking prevalence, calculate the total as the sum of the female and male smokers divided by the sum of the female and male sample populations.

In the above example, the total prevalence is calculated as:

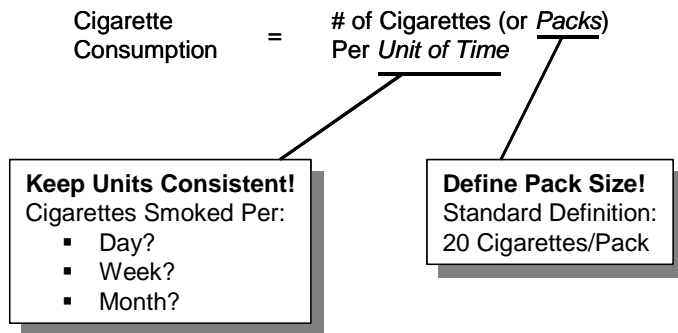
$$((250 + 400) \div (750 + 600)) \times 100 = 48.14\%$$

## Conditional Demand

The conditional demand for cigarettes is the actual number of cigarettes smoked by those consumers who have declared being cigarette smokers.

Researchers may define cigarette consumption as the number of cigarette packs (usually understood to consist of twenty cigarettes) smoked by individuals or households during a given unit of time (e.g., the last month, the last week, daily). Tobacco consumption questions generally ask about the number of packs or number of cigarettes consumed per month (see Figure 2.1).

**Figure 2.1. The Importance of Standard Definitions of Tobacco Product Size**




---

## Assumptions and Requirements

Due to the nature of the topic of this Tool, there are only a few assumptions made within these pages. This Tool assumes that general economic, consumer, and specific tobacco production and marketing data are available to the reader. This Tool also assumes that the reader has either direct access or the minimum financial resources required to help him/her achieve access to basic data management tools, particularly those most popular for use with a personal computer.

To properly and most effectively use this Tool, the reader must have a basic comprehension of data collection and basic knowledge of storage techniques. Practical knowledge of and experience with the data collection and management software referenced in this Tool, while not necessary, is highly beneficial. As a *caveat*, the reader should not strictly rely on this Tool for technical questions and concerns related to the software referenced within. Instead, software user manuals and other client-help documentation (frequently available on-line) should be referenced.

# III. The Basics of Collecting Data

---

## What to Know about Data

### Different Data Types

In the world of numbers, figures can be aggregated or summed to various sub-levels of a society. The highest categorization of data for any country occurs at the national level. This category captures rounded numerical descriptive information about every person (e.g., country's total population), every item traded (e.g., prices, production) and societal or economic mechanisms (e.g., interest rates), and reports it for the country as a whole. The next sub-level category represents a country by region (e.g., north, south, east, west), where the same set of information (e.g., population, prices) is captured to reflect the average for each region (at "the regional level").

It is helpful to think of these levels of data by example. Consider a frequently reported economic figure: earnings. In market economies across the globe, the concept of income or earnings is always of strong interest. The *Wall Street Journal* reports the United State's national income or Gross Domestic Product (GDP) daily. Financial analysts and investment brokers report corporate earnings four times a year. Census Bureaus and Central Statistical Offices regularly track household earnings in homes across their nations. Employers track hourly and monthly wages. Individuals report annual incomes to their respective governments.

Thus there are a number of ways to measure and define income. For example, Mr. Smith's employer, Company A, defines Mr. Smith's income by establishing his hourly wage. Mr. Smith's tax lawyer determines Mr. Smith's income based on the annual sum of his wages earned at Company A plus the sum of his stock, savings, and other investment earnings. A census worker reports Mr. Smith's

income along with Mrs. Smith's income through a household income measure. Finally, the *Wall Street Journal* reports Mr. Smith's income, along with the income of thousands of other individuals and businesses, through a national measure of income—the GDP.

*Be sure to capture comparable and compatible sets of variables that, together, clearly and accurately describe your subject matter.*

Each of these definitions of income is a valid and reflective measure of Mr. Smith's income, even though they present income at different levels (individual, household, national) and magnitudes (hourly, monthly, annual). When conducting economic analyses, it is important to capture comparable and compatible sets of variables which, when combined, tell a clear and cohesive story about the person, community of persons, or nation of people being addressed.

Over the years, economic studies have used various types of data in their analyses, including:

***Aggregate Time Series***

- Consists of several years of aggregate data
- Constructed from stacked annual national estimates

***Aggregate Cross-Sectional***

- Consists of data drawn in one single moment in time
- Based on a nationally representative survey of households

***Pooled Time-Series***

- Consists of several years of individual or household level data
- Pools together several years of aggregate cross-sectional data into a single database

***Longitudinal***

- Consists of several years of individual level data
- Tracks and repeatedly surveys the same sample individuals across time

These economic analyses use a wide variety of statistical and econometric techniques to examine the effects of economic factors and socio-demographic characteristics on issues related to the supply and demand for tobacco in the consumer market.

## **Aggregate Data**

Most national governments in the world today report at least a basic set of national economic and social information. In addition, aggregate or “macro” level data is also largely available at sub-national levels—that is, information reflective of regional, state, provincial, county, or other jurisdictional divisions of a country.

The International Monetary Fund's (IMF) *International Financial Statistics* publication provides a good example of basic economic information that is collected and reported by national governments

each month. This publication lists figures for such economic measures as GDP, money supply, consumer price index (CPI), producer price index (PPI), interest rates, and industrial production. This information provides the researcher with a useful summary of the overall economic status or performance of each country. Such data is especially helpful when trying to account for fluctuations in inflation over time and differences in the cost of living across countries.

In most countries, a similar array of national figures specific to tobacco are reported to central authorities such as the Central Statistical Office, the Ministry of Finance, the Ministry of Commerce, the Ministry of Trade, and others. The reported set of national and sub-national information may include: consumption and sales of tobacco products, retail prices and taxes for tobacco products, export and imports of raw tobacco and finished tobacco products, information on consumer tobacco-related expenditures, and demographic characteristics of consumers.

---

## Where to Begin Looking

Various institutions around the world collect information about people and the societies they live in. Most often and most regularly, our own governments keep close track of our actions, including:

- Who we are (e.g., age, gender, race, religion, education)
- Where we live (e.g., urban or rural, rates of migration)
- Who we live with (e.g., number of children, marital status)
- Where we work
- How we earn our income
- What we buy (e.g., expenditure and consumption of food, energy, luxury goods)

Most governments regularly conduct household surveys in order to map the demographic, socio-economic, expenditure, and employment characteristics of their national societies. These surveys help governments better understand the economic and social conditions that exist, and identify the resources needed to improve national welfare.

Regardless of their political and/or economic systems, most governments maintain large institutions equipped with a number of tools and methods to gather, record, manage, and disseminate such information. These two institutions are typically:

- a centralized data collection agency
- individual ministries or departments of government

## **Central Data Collection Agency**

Governments use different names for their central data collection agencies. Common names include: Central Statistical Office, National Bureau of Statistics, General Statistical Office, or National Institute of Statistics. A national data collection agency generally has two main duties:

- Collect and publish primary data through censuses as well as household and individual surveys
- Gather and report secondary data collected by ministries or departments of government

The collection of data through a national agency ensures that the data gathered represent the entire national society and that there is no influence by interest groups. Furthermore, because it is generally accepted that a government has the authority to collect such data, cooperation in the data collection process is quite strong.

## **Ministries or Departments of Government**

A central data collection agency often relies on other ministries or departments of government for help with collecting national data. These ministries or departments include agriculture, commerce, finance, health, industry, justice, trade, and others that regularly monitor relevant aspects of a society. The following ministries or departments of government are examples of tobacco data sources.

- Ministry of Finance: Directs and records tobacco taxes
- Ministry of Commerce: Tracks all tobacco products, brands, prices, and sales
- Department of Industry: Oversees tobacco production
- Department of Agriculture: Tracks tobacco farming
- Ministry of Department Trade: Monitors tobacco imports and exports; determines trade duties

National data can be obtained from a number of different sources, including: a national central data agency, international data sources (e.g., the United Nations), non-governmental data sources (e.g., Action on Smoking and Health), private data companies (e.g., A.C. Nielsen) as well as select U.S. and other country agencies (e.g., Centers for Disease Control).

Information on national and international data sources and how to access them are discussed in greater detail in the **Suggestions for Data Sources** chapter of this Tool.

---

## Important Data to Collect

### Consumption

Consumption represents product use. Therefore, data on tobacco consumption reflects the amount of tobacco products used by a consumer. Data on tobacco product consumption is required for any economic analysis related to the demand for tobacco. National and sub-national measures of use or consumption of cigarettes and/or other tobacco products are imperative to each of the tools presented in this toolkit, especially **Tool 3. Demand Analysis**, **Tool 4. Design and Administration**, and **Tool 7. Smuggling**.

Tobacco consumption information is obtained through surveys of households and/or individual consumers (see Box 2.1). National population surveys and censuses interview random samples of individuals and/or households in an effort to obtain behavioral and socio-economic information that best describes the characteristics of the nation's current population.

Such surveys generally include a few direct questions about tobacco related behaviors. Such a survey also usually asks if the individual respondent or household uses tobacco, whether cigarettes in particular are smoked regularly and, if so, how much. In this manner, a country's central statistical office or national bureau of statistics is able to gather direct consumption information from individuals and households. Such information is later used to represent current consumption statistics for the national population and to produce estimates of future tobacco consumption behaviors.

Measures of smoking prevalence are often not comparable across countries, as the basic definition of a current smoker tends to vary in such instances. Surveys are often administered to varying age, gender, and social groups. For example, adult daily smokers in country X may range in age from 16 years and over while adult daily smokers in country Y may only include smokers in the range of 21 years and above.

The World Health Organization (WHO) defines a current smoker as one who smokes at the time of the survey and has smoked daily for at least a period of six months (WHO, 1998). Other definitions of smoking prevalence are less restrictive; some groups define a current smoker as one who has smoked one or more cigarettes in the 30 days prior to the survey. Ongoing efforts by the WHO, the Centers for Disease Control (CDC), and others aim to improve the consistency of survey data related to tobacco use across countries.

The CDC's *Global Youth Tobacco Survey* is an international, youth-focused survey conducted in a large and continuously growing number of countries since 1999. This survey contains a standard set of questions universally administered across all participating countries. Such uniformity in survey design and administration

**Box 2.1. Question on a Consumer Survey Regarding Smoking Behavior**

- Q. How many cigarettes a day do you smoke on average (one pack equals 20 cigarettes)?
- A. None
  - B. Less than one cigarette
  - C. Less than half a pack
  - D. About half a pack
  - E. More than half a pack, but less than a pack
  - F. A pack
  - G. More than a pack

Individual answers to this question are aggregated to reflect national consumption measures, which are then reported in two distinct formats:

1. *The Prevalence of Tobacco Use*

Individuals who report smoking “none” are defined as non-smokers, while those who answer smoking less than one cigarette per day or more (responses B through G) are defined as smokers. The percentage of defined smokers relative to the total number of respondents (smokers plus non-smokers) reveals the prevalence rate of tobacco use within a national sample of respondents.

2. *Conditional Demand for Tobacco*

A quasi-continuous measure of daily cigarette consumption can be constructed, so long as the respondent is a smoker. Using the format above, the conditional demand for cigarettes is equal to a value of:

- 0.5 if on average a respondent smokes less than one cigarette per day
- 5 if on average a respondent smokes less than 10 cigarettes per day
- 10 if on average a respondent smokes approximately 10 cigarettes per day
- 15 if on average a respondent smokes 10–19 cigarettes per day
- 20 if on average a respondent smokes a pack of 20 cigarettes per day
- 30 if on average a respondent smokes a pack or more of cigarettes per day

In order to produce an annual estimate that reflects conditional demand of a national population, these individual averages are aggregated to a national level. Once this aggregation is complete, the resulting national average reflects the average daily cigarette consumption of the population.

For recode examples of these consumption measures, refer to the **Data Preparation and Management** chapter of this tool.

ensures feasibility for conducting a standard set of analyses across countries. Contact the CDC (<http://www.cdc.org>) for additional information on this survey.

*Beware of underreporting!*

While survey data provide generally accurate measures of prevalence (depending on the quality of the survey), there is some potential for the individual underreporting of smoking and/or other tobacco use prevalence. This is particularly true in countries and among populations characterized by strong social disapproval of smoking behaviors. In addition, survey data on prevalence may also be biased as a result of the manner in which the survey is conducted. For example, household surveys that are conducted orally by an

interviewer can lead to inaccurately reported measures if the survey is not conducted privately (youth and young adults are less likely to honestly report that they smoke when their parents may overhear their responses). Finally, measures of total consumption derived from survey data are likely to be inaccurate. Past research demonstrates that the level of total cigarette consumption derived from survey data on smoking participation and average cigarette consumption by smokers is significantly lower than cigarette sales. The degree of underreporting is likely to be positively related to the social disapproval of smoking.

*Determine the size of “a pack” of cigarettes*

Survey designers should be sensitive to the fact that standard pack sizes vary from country to country (e.g., 10, 12, 20, or 25 individual cigarettes or “pieces” per pack). Also, in some countries the sale of single cigarettes (cigarette “sticks”) is common. Survey questions that inquire about the number of packs consumed per month should clearly define the size of a cigarette pack in the survey questionnaire.

This confusion over the unit of consumption measure often results in errors in the design and analysis of consumption variables. For example, a researcher may be under the impression that the consumption measure which he/she inquires about in a survey is defined as the number of packs consumed per month (1 pack per day translates into approximately 30 packs per month) while the survey respondent may be reporting the number of single cigarettes (e.g., 20–30 individual cigarettes) smoked per day. In an effort to avoid the miscoding of consumption information, it is worthwhile to check consumption measures by verifying the corresponding price per pack supplied by the respondent of the survey.

## Sales

Cigarette sales information, specifically tax paid sales data, can be used as a proxy (substitute measure) for cigarette consumption in aggregate cigarette demand models. This means, total annual tax paid cigarette sales can be modified to produce per capita proxies of cigarette consumption. Compute per capita cigarette sales by dividing total annual cigarette sales in country X at time Y by the total population of country X in time Y. Similarly, obtain *adult* per capita cigarette sales by dividing total annual cigarette sales by the appropriately defined adult population measure (15 years and older or 18 years and older are commonly used).

As with aggregate estimates of prevalence and consumption obtained from survey data, tax paid sales data can be systematically biased. This is particularly true for countries with a significant black market in tobacco products. In this case, cigarette sales provide an underestimate of total consumption. Smuggling activity can distort consumption estimates. Consumption estimates have been adjusted in those countries within which the extent of cigarette smuggling has been successfully estimated. See **Tool 7. Smuggling** for alternative approaches to estimating the magnitude of the black market for

tobacco products. In addition, cigarette sales may provide misestimates for reasons related to hoarding. Consider:

- Although a consumer purchases a pack of twenty cigarettes in time Y, one cannot be certain that this individual consumes all twenty cigarettes in time Y.
- Cigarettes may be purchased in large quantities in time Y as a safeguard against higher taxes in time Z. Such quantities often go unsold and unused in time Y or time Z, and are discarded after expiration.

Therefore, cigarette sales, although an appropriate proxy for tobacco consumption, are likely to provide a distorted estimate of cigarette consumption and, by definition, should be clearly distinguished from consumption data.

## Tobacco Price

Cigarette and other tobacco price information are critical to each of the tools of this toolbox. Price plays a critical role in tobacco demand estimates and is a key factor in most, if not all, economic issues related to tobacco—including smuggling and taxation.

*As cigarette prices increase, the quantity demanded of cigarettes can decrease.*

For many years, economists believed that because of their addictive nature, cigarettes and other tobacco products were not normal goods. As a result, it was believed that the consumption patterns of a tobacco consumer would not be responsive to changes in price (see Figure 2.2). However, today, through improved econometric techniques and sophisticated statistical programs, many studies show that the demand for tobacco is in fact sensitive to changes in the price of tobacco. Many studies conclude that by altering the price of cigarettes (through tobacco taxation) governments can change tobacco use. Microeconomic theory dictates that, as the price of a normal good rises, the quantity of the good that is demanded by a consumer falls (see Figure 2.3).

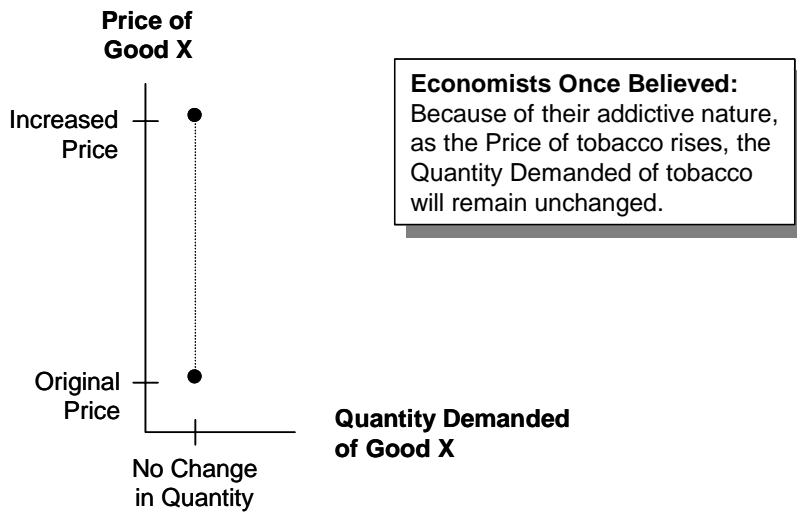
### **Price Elasticity of Demand**

Tobacco demand's sensitivity to changes in tobacco price is the price elasticity of demand, defined as the percentage change in consumption that results from a 1% change in the price of a good.

$$\text{Price Elasticity of Demand} = \frac{\% \text{ Change in Cigarette Consumption}}{\% \text{ Change in the Price of Cigarettes}}$$

In order to understand how price changes may influence smoking decisions, measure this ratio within the population at hand. The relationship between price and consumer consumption carries very strong policy implications and helps determine which taxes, and in what magnitude, need to be altered to achieve a planned reduction in consumption. This in turn also provides estimates of how much government revenue will increase as a result of higher taxes and decreased consumption.

Figure 2.2. Early Economic Theory of Price and Demand of Tobacco Products

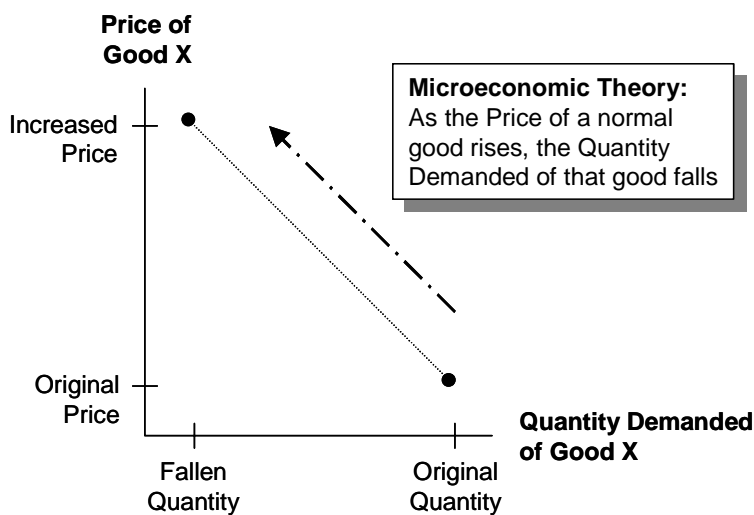


An increase in cigarette taxes and cigarette prices will affect smokers' decisions about their smoking behavior through a number of mechanisms. For the addicted cigarette smoker, higher taxes and prices on cigarettes:

- have a negative effect on the number of cigarettes consumed
- often stimulate the decision to switch to smoking cheaper brands of cigarettes
- enhance the decision to quit or begin to think about quitting the smoking habit

By the same token, higher tobacco prices also have a discouraging effect on the consumer decisions of those who do not smoke. That is, non-smokers, when faced with rising cigarette prices, may think twice before initiating smoking behaviors.

Figure 2.3. Current Microeconomic Theory of Price and Demand of Tobacco Products



The monetary price of a pack of cigarettes that consumers encounter when purchasing their cigarettes consists of several individual and variable components of price. It includes the retail price of a pack of cigarettes plus any combination of tobacco taxes, including:

- percentage sales tax
- flat excise tax
- *ad valorem* tax
- Value Added Tax (VAT)

Use a variety of tobacco price data in demand analysis, including the prices of various categories and types of tobacco products. For example, it is useful to include the prices of alternative tobacco products in the demand analysis in order to understand the potential for substitution among tobacco products in response to relative price changes.

### ***Real versus Nominal Cigarette Prices***

The actual price paid by an individual at a particular moment in time is called the nominal price. However, in many econometric analyses of cigarette demand, a set of nominal prices should not be used. Instead, it is correct to use the real value of price or a deflated price measure. Here, the price variable is adjusted for inflation. The common method for converting nominal prices into real prices is to divide the nominal price by the CPI level and multiply by 100 (see **Tool 3. Demand Analysis** and **Tool 6. Poverty** for further details).

When price data are unavailable, tobacco product excise tax data are often a good proxy for price. Similar to prices, tax levels on tobacco products tend to vary depending on the type, origin, and size of the tobacco products. Research from developed countries shows that tobacco prices are very highly correlated with tobacco taxes and that increases in taxes are generally fully passed on to consumers.

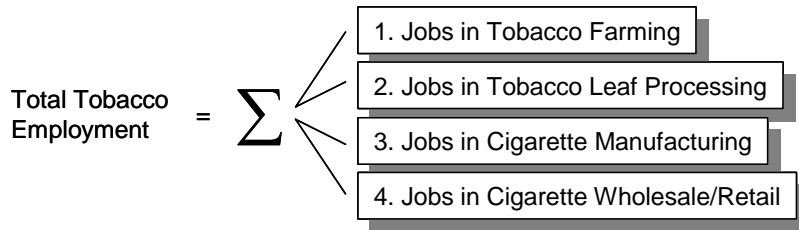
Use tax per pack in a demand equation to get an estimate of tax elasticity. Elasticities estimated from demand models that use tobacco tax rather than tobacco price must be converted to price elasticities.

## **Employment**

To count the total number of jobs (employment) directly related to tobacco, it is necessary to gather data from four types of tobacco employment (see Figure 2.4):

- tobacco farming
- tobacco leaf processing and marketing
- cigarette manufacturing
- cigarette wholesaling and retailing

This information is generally available from government statistical offices, publication agencies, and other employment-related data

**Figure 2.4. Tobacco Employment Resides within Four Industries**

sources. For example, in the United Kingdom this information is available from the Department of Employment. In most countries of Central and Eastern Europe, such information is available from the Central Statistical Office. In the U.S., the Bureau of Labor Statistics and the Department of Commerce publish most information on employment.

### ***Tobacco Farming***

Information on the number of jobs held in tobacco farming is generally not available from statistical sources. As a result, there are two methods to derive or estimate this information. Both methods require *two* steps: (1) estimate the total hours of labor used; (2) convert hours into full-time employment units.

#### **Method 1: Estimate the total hours of labor used as:**

Amount of tobacco produced  $\times$  Hours of labor required per unit of production

#### **Method 2: Estimate the total hours of labor used as:**

Acres of tobacco planted  $\times$  Hours of labor required per acre

### ***Tobacco Leaf Processing***

Tobacco leaf processing can be broken down into two specific components necessary for the preparation of raw tobacco leaves for use in production:

- auctioning and warehousing of raw tobacco leaves
- stemming and redrying of raw tobacco

In the U.S., employment associated with tobacco leaf processing and marketing is available in various publications produced by the Bureau of the Census. The *Census of Wholesale Trade* contains information on auction warehousing and, specifically, the number of auctioning establishments and corresponding employment statistics. The *Census of Manufacturers* provides information on the number of stemming and redrying establishments and corresponding employment.

The organization of tobacco production and, therefore, the organization of tobacco leaf processing both vary from country to country. As a result, in many countries tobacco leaf auction

warehousing may not be regarded as a separate production activity. Similarly, in many countries the stemming and redrying of tobacco leaves may be considered as a part of the cigarette manufacturing industry. In such cases, do not estimate employment associated with these activities, since they are already counted in tobacco farming and manufacturing.

### ***Cigarette Manufacturing***

Data on employment in cigarette manufacturing are commonly available in government statistical offices. This information is usually categorized according to market sectors or industries of the national economy.

In the United States, the Bureau of the Census' *Census of Manufacturers* contains information on the number of tobacco producing establishments and tobacco manufacturing jobs. For similar information in other countries, data is often published by international organizations. The UN *International Development Origination Database* is a good place to start a search.

### ***Cigarette Wholesaling and Retailing***

Cigarette wholesaling is performed by different entities in different countries. In some, cigarette manufacturing and sales are monopolized. That is, the wholesale of cigarettes is part of the cigarette manufacturing industry. Centralized manufacturers have regional depots and transport facilities for the distribution of tobacco products (a function otherwise performed by wholesalers). In these countries, the number of jobs related to wholesaling is included in statistics measuring total employment in cigarette manufacturing.

Elsewhere, wholesaling is a distinct function in a competitive and open market. Wholesalers begin to handle tobacco products immediately after they leave the manufacturer. In such situations, there are several methods to find employment data. Using the United States as an example:

- Employment associated with wholesaling can be obtained from the Bureau of the Census' *Census of Wholesale Trade*.
- Estimates of jobs associated with cigarette retailing can be imputed from information on the number of distribution outlets, the total number employed in each outlet, and the share of tobacco product sales.
- Information on the number of distribution outlets and the total number employed can be found in *Employment and Earnings* published by the Bureau of Labor Statistics of the U.S. Department of Labor.
- Tobacco's share of total retail sales by individual retail outlet can be obtained from the *Census of Retail Trade*.

Note: In most countries, statistics on the distribution channel of tobacco products and tobacco share of the total sales is poor. In such cases, a retailer survey is required to capture such information.

### **Tobacco-Related Employment versus Total Employment**

Macro-employment measures are needed to estimate the proportion of tobacco-related employment to total employment by sector. The relevant sectors of employment include: agricultural production, agricultural marketing, manufacturing, wholesaling, and retail trade. These macro measures are used to create four ratios:

- $(\# \text{ employed in tobacco farming})/(\# \text{ employed in total agricultural production})$
- $(\# \text{ employed in tobacco leaf marketing and processing})/(\text{total } \# \text{ employed in agricultural marketing})$
- $(\# \text{ employed in cigarette manufacturing})/(\text{total } \# \text{ employed in total manufacturing})$
- $(\# \text{ employed in tobacco wholesaling and retailing})/(\text{total } \# \text{ employed in wholesaling and retail trade})$

In many countries, information on employment by sector is available from government statistical data on employment. Be sure to check other specific sources and publications, both nationally and internationally, as well.

For example, in the United States, the Bureau of Labor Statistics and the Department of Commerce publish information on employment. In the United Kingdom, this data is available from the Department of Employments. In Poland and other former centrally planned CEE countries, employment information by industry and sector is available through the Central Statistical Office.

Be sure to consider other data relevant to tobacco employment, other than those figures directly related to employment. For instance, data on consumer expenditures on finished tobacco products are needed to examine the impacts of tobacco control policies on national and/or regional tobacco employment. Other information relevant to the impact of tobacco control policies on employment and production include the amount of labor input required for the production of a unit of tobacco or an acre of planted (or harvested) tobacco (see below). In the United States, this information is available through the Census Bureau's *Census of Agriculture*.

### **Tobacco Production**

When gathering information on tobacco production, it may also be useful to capture data that reflects the economic importance of tobacco (both the value of raw tobacco and the value of finished tobacco products) to a given economy. Such measures include:

- the monetary value of tobacco leaf grown within a defined area
- the value added by tobacco manufacturing

Figures on the annual, national production of cigarettes (often reported in billions of cigarettes) is available in most countries. Information on tobacco production and acreage used in tobacco farming are frequently available through national agricultural statistics of individual countries.

*Anomalies in production figures may indicate stockpiling or other production or inventory changes for the purpose of taking advantage of tax increases.*

Beware of spikes in cigarette or other tobacco product production. These may reflect stockpiling or the process of increasing wholesale or retail inventories of tobacco goods in anticipation of expected tax increases.

When using per capita production figures, also be aware of the weight of tobacco contained per cigarette. Although cigarettes are often assumed to contain one gram of processed tobacco, this does not always hold true. In the past, cigarettes sold in developed countries often contained more than one gram of tobacco. Recently, however, the cigarettes produced in these countries contain less than one gram of tobacco. (WHO, 1998)

The UN's Food and Agricultural Organization (FAO) *Production Yearbook* and the U.S. Department of Agriculture *World Tobacco Situations* regularly contain data on several sectors of agricultural production, including tobacco.

## Consumer Expenditures

In a manner similar to tobacco consumption, obtain information on consumer spending patterns (either by households or individuals) from published or non-published governmental statistics on consumer expenditure. Tobacco expenditure data is based on information collected through national surveys of random samples of households and/or individual consumers. These surveys generally include a few direct questions concerning general household expenditures, including tobacco product expenditures. National household surveys usually inquire about expenditures on a variety of household items including perishables (e.g., fruits, vegetables, meats), dairy products, rice, potatoes, eggs, tea and coffee, alcoholic beverages, oils and fats, sugar, salt, and tobacco (see Box 2.2).

From such surveys, a central statistical office or national bureau of statistics is able to directly gather expenditure information from individuals and households. This information can later be used to represent total expenditures made by the national population. (Refer to **Tool 4. Design and Administration**, **Tool 5. Tobacco Control**, and **Tool 6. Poverty** to learn how information concerning individual expenditure or household expenditures on tobacco products, as well as household expenditures on other types of goods and services, are important to the economic analyses of tobacco.)

**Box 2.2. Survey of Household Consumption Expenditures and Main Sources of Income**

**Directions:** The following question should be asked of the head of household, spouse of the head of household, or of another adult household member, if both head and spouse are absent.

**Question:** What was the total value of food, beverages, and tobacco consumed in your household during the past week?

Food Item	Value of Consumption (Riels)		
	(A) Purchased	(B) Overproduced, Gifts, etc.	Total Consumption (A) + (B)
<b>Rice</b>			
<b>Sugar, salt</b>			
<b>Fruit</b> (banana, orange, mango, pineapple, lemon, watermelon, papaya, durian, grape, apple, canned and dried fruit, etc.)			
<b>Meat</b> (pork, beef, buffalo, mutton, dried meat, innards—liver, spleen, and other meat)			
<b>Tea, Coffee, Cocoa</b>			
<b>Tobacco Products</b> (cigarettes, mild tobacco, strong tobacco)			

Source: Cambodia Socio-Economic Survey (1999)

### ***The Expenditure Ratio***

A simple comparison of household expenditures made towards *tobacco products* relative to household expenditures for *other consumables* reveals how important tobacco is to the national economy. That is, countries with low expenditures on tobacco products relative to expenditures on other necessary goods (e.g., rice, fruits, vegetables) conceivably have lower smoking prevalence rates and depend less on tobacco sales within their economy.

Calculate such an expenditure ratio to better understand how important tobacco is in the lives of the national population and the

**Figure 2.5. The Relationship of Tobacco Product Expenditures to Other Expenditures**

$$\text{Expenditure Ratio} = \frac{\text{Monthly Expenditures on Tobacco Product}}{\text{Monthly Expenditures on:}}$$

- Food
- Housing
- Energy
- Other Necessities

livelihood of the local economy. Figure 2.5 illustrates the relationship between tobacco expenditures and expenditures on other goods and services.

- If the value of the expenditure ratio is 1 or greater, this indicates that the average monthly amount of money spent on tobacco products is *more* than the total monthly amount spent on other household consumables (e.g., food, beverages) or household necessities (e.g., housing, energy).
- If the value of the expenditure ratio is near 0.5, this indicates that individuals spend about the *same* relative monthly amount on tobacco products as on other household consumables.
- If the value of the expenditure ratio is 0 (zero), this indicates that a household spends no household financial sources on tobacco products.

### ***Expenditures by Type of Tobacco Product***

Household expenditures by type of tobacco product may also be available through the statistical offices of some countries. Depending on the country in question and the mix of legally available are tobacco products, expenditure figures may be reported for one or more of the following categories:

- Manufactured cigarettes—processed tobacco manufactured by a machine.
- *Bidis*—small hand rolled cigarette made of tobacco wrapped in a piece of dried temburni leaf and tied with a small string. Most often used in areas of Southeast Asia.
- Roll-Your-Own (RYO) cigarettes—hand rolled cigarettes made by the smoker from processed tobacco and cigarette papers.
- Manufactured cigars
- Manufactured cigarillos
- Tobacco-filled pipe
- Chewing tobacco—loose tobacco leaves that are chewed instead of smoked.

Compare expenditure values across these categories to determine proxy measures for the market shares of each of these tobacco product categories. (Note: in most countries, cigarettes are the most used tobacco products. As a result, cigarette expenditures generally dominate total tobacco expenditures.)

In the U.S., consumer expenditure information on various tobacco products can be found in tobacco statistics published by the U.S. Department of Agriculture and the U.S. Department of Commerce. In other countries, and particularly in the post Stalinist economies of Central and Eastern Europe, this data is collected by the national Central Statistical Office. Similarly, in the Southeast Asian countries of Vietnam and Cambodia, this data has just recently begun to be collected by each country's National Office of Statistics.

## **Demographic Information**

Use demographic information to partition a sample of individuals into population subgroups or sets of people who share common demographic characteristics, such as age and gender.

Use national socio-demographic information to summarize or define the population sample being examined. Commonly collected national demographic information includes:

1. Figures that provide a statistical breakdown of the population by year and by:
  - age
  - gender
  - education level
  - religious denomination
  - area of residence (rural, urban, etc.)
2. Annual measures of gross household or gross per capita income
3. Annual measures of net household or net per capita income

Such information can be obtained from either the Central Statistical Office or the National Statistical Bureau in most middle-income and developing countries.

## **Economic Indices**

National economic indicators are required for even the simplest descriptive analyses of aggregate data. There are two data measures that are particularly important.

- The Gross Domestic Product (GDP) is a measure of national income. This measure is also often cited as a good indicator of national economic performance. By dividing GDP by national population, one can also obtain a satisfactory proxy

of individual income per capita. Similarly, by dividing GDP by the number of households, one also obtains a proxy for household income.

- The Consumer Price Index (CPI) is an aggregate measure of overall prices, and serves as a popular indicator of the rate of inflation in a given economy. The rate of change in the CPI indicates the rate of inflation. With the CPI, it is possible to measure current prices against an overall price level. In other words, the CPI is an instrument (a deflator) that allows one to deflate monetary measures (e.g., taxes, prices, income) to make them comparable over time. By deflating prices by the CPI, prices can be measured in real rather than nominal terms.

As an example, consider nominal versus real prices of a pack of cigarettes. The *nominal price* is the pack's current monetary value, or absolute price. For instance, in the U.S. the current selling price of a pack of regular Marlboro cigarettes is approximately \$3.25, while ten years ago the same pack sold for \$2.75. Therefore, one can conclude that the nominal price of a pack of Marlboro cigarettes in 1991 was \$2.75 while the nominal price in 2001 measured \$3.25.

The *real price* of a pack of cigarettes is its nominal price relative to the CPI. By dividing each nominal price by its respective CPI measure, one can compare the two prices against one another. So, assuming a 1991 CPI of 1.20 and a 2001 CPI of 1.32, the

$$\text{Real Cigarette Price in 1991} = \$2.75 \div 1.20 = \$2.29$$

while the

$$\text{Real Cigarette Price in 2001} = \$3.25 \div 1.32 = \$2.46$$

Therefore, according to the data in this scenario, both the nominal and real prices of a pack of regular Marlboro cigarettes were higher in 1991 than in 2001.

These economic indices are easily obtained through a number of different sources. The IMF's *International Financial Statistics* provides monthly, up-to-date reports of both GDP and CPI data.

## Tobacco Trade Information

Trade information is an important factor in both **Tool 5. Tobacco Control** and **Tool 7. Smuggling**. Important trade measures include tobacco and/or cigarette:

- exports
- imports
- domestic sales
- export sales

Obtain data related to the tobacco trade through the Central Statistical Office or the National Statistics Bureau of most countries. Information is also likely to be available, although less accessible, through a national Ministry or Department of Trade.

---

## The Market for Tobacco

In addition to understanding the regulatory environment surrounding tobacco, it is also often helpful to gain a detailed and descriptive understanding of both the demand and supply sides of the tobacco market. This includes understanding annual production as well as market share of various tobacco producers, their product brands, sizes, and subcategories.

Begin by determining descriptions of the tobacco market from tobacco-related data available from various government sources (e.g., the Central Statistical Office or Bureau of Statistics, the Ministry of Finance, the Ministry of Commerce). Private institutions also focus on the monitoring and tracking of the activities, practices, and performance of tobacco. Marketfile is an example of a private organization that monitors tobacco in countries worldwide (<http://www.marketfile.com/market/tobacco/>).

### Examples of Market Analyses

Analyzing the market requires a calculation of market shares of various tobacco products. This can be completed in a variety of ways; Table 2.1 shows a popular and well-accepted method to categorize tobacco products.

### Tobacco Price Measures

Retail prices of cigarette packs are available from a number of government and private data sources. Cigarette price information can generally be obtained from the Ministry of Finance and/or Ministry of Commerce. Both track retail prices of tobacco products; the Ministry of Finance is concerned about the tax implications associated with varying tobacco prices, while the Ministry of Commerce monitors retail prices of nearly all goods, including tobacco, sold in the domestic market. Many Central Statistical Offices or government data bureaus also report the retail price of cigarettes and/or smokeless tobacco in annual data yearbooks.

Alternatively, tobacco price data can be purchased from a number of private data collection firms. For example, AC Nielsen collects cigarette price data in a wide range of developed, middle-income, and developing countries. Other international private data collection firms include Information Resources International (IRI), and Sofres, Taylor, Nelson Inc. See the **Suggestions for Data Sources** chapter for additional information regarding these sources.

**Table 2.1. Tobacco Product Categorization**

<b>Type of Tobacco Product</b>	<b>Cigarette Category</b>
<ul style="list-style-type: none"> <li>▪ Cigarettes</li> <li>▪ Cigars</li> <li>▪ Cigarillos</li> <li>▪ Smokeless Tobacco</li> <li>▪ Loose Tobacco</li> <li>▪ Other Tobacco Products</li> </ul>	<ul style="list-style-type: none"> <li>▪ Filtered Cigarettes</li> <li>▪ Unfiltered Cigarettes</li> <li>▪ Menthol Flavored Cigarettes</li> <li>▪ Lights (and ultra light) Cigarettes</li> <li>▪ Other Emerging Categories</li> </ul>
<p><b>Cigarette Size</b></p> <ul style="list-style-type: none"> <li>▪ Under 70mm</li> <li>▪ Regular Size (70mm)</li> <li>▪ King Size</li> <li>▪ Superkings</li> </ul>	<p><b>Cigarette Packaging</b></p> <ul style="list-style-type: none"> <li>▪ Soft Packs</li> <li>▪ Box Packs</li> <li>▪ Cartons</li> </ul>
<p><b>Cigarette Producer</b></p> <ul style="list-style-type: none"> <li>▪ Domestic Producers (varies)</li> <li>▪ International Conglomerates (Phillip Morris, RJ Reynolds, British American Tobacco)</li> </ul>	<p><b>Cigarette Brand</b></p> <ul style="list-style-type: none"> <li>▪ Domestic Brands (varies)</li> <li>▪ International Brands (Marlboro, L&amp;M, Winston, Lucky Strike, Salem)</li> </ul>

## Tobacco Regulatory Environment

In order to properly conduct tobacco related economic analyses in a given country, one must first and foremost understand the regulatory environment surrounding tobacco products in that country. National and subnational information on prevailing tobacco control laws is necessary to understanding the intricate economic aspects of tobacco consumption and production within a country. A complete understanding of a country's tobacco regulatory environment is required for a number of analyses discussed in this toolkit, particularly for the demand analyses presented in **Tool 3. Demand Analysis** and the studies of tobacco smuggling presented in **Tool 7. Smuggling**. When collecting this data, be sure to record both the date on which the regulation is announced to the public and the actual date of enactment.

When gathering information in a country or population of interest, note the existence and extent of one or more of the following:

- Controls over tobacco advertising and sales promotion
- Displays of health warnings and statement of tar and nicotine content
- Legal control of harmful substances contained in tobacco products

- Restrictions on sales to adults and youth
- Smoke-free public places, public transport, and workplaces
- Preventions to keep young people from smoking, including
  - Prohibition of sales to minors
  - Restrictions on sales of cigarettes from vending machines
  - Prohibition of smoking and sales in schools
  - Prohibition of free samples and sales of single cigarettes
  - Restrictions on smokeless tobacco products
  - Restrictions on tobacco advertising and promotion aimed at young people
- Health education and its requirements
- Judicial action(s) for tobacco control
- Economic strategies regarding tobacco production, including
  - Tobacco subsidies
  - Tobacco trade policies
  - Efforts to decreasing tobacco production (crop substitution, alternative off-farm work)
  - Tax and price policies (reasons for increasing taxes, types of taxation)

Box 2.3 presents some key regulatory information to include in the collection process.

---

## **Aggregate Model Results**

Use caution when interpreting aggregate model results. As with any analysis that involves data, a number of caveats that may arise when using a gathered set of aggregated variables. There are at least four issues to be aware of when using aggregate measures in economic analyses, as outlined below.

### **Multicollinearity**

One difficulty encountered in studies using aggregate level time-series data originates from the high correlations existing between price and many other key independent variables. For example, in cigarette demand models, estimated price and income elasticities of demand depend upon the descriptive variables (those which control for the effects of other important determinants of smoking such as advertising, health awareness, etc.) included in the model.

Consequently, estimates of the impacts of price and other factors on the demand for cigarettes are sensitive to variables that are both

**Box 2.3. Key Determinants of the Tobacco Regulatory Environment****Tobacco Taxation**

1. How is tobacco taxed?
  - Taxes on raw tobacco leaves
  - Import duties
  - Excise, sales, *ad valorem*, and sales taxes

**Restrictions on Smoking**

1. Are there legal restrictions on smoking?
2. If so, where is smoking restricted?
3. What is the extent of these restrictions?
  - Is smoking *totally banned* in workplaces, theaters, health care facilities, etc.?
  - Is smoking *partially banned* in workplaces, theaters, health care facilities?
  - To what extent are these smoking restrictions *enforced*?

**Advertising Restrictions**

1. Are there legal restrictions on cigarette advertising?
2. If so, what is the extent of the restriction?
  - Is cigarette advertising *totally banned*?
  - Is cigarette advertising *partially banned*?
  - Are advertising restrictions strictly *enforced* by authorities?

**Restrictions on Youth Access**

1. Is there a minimum age requirement for the legal sale and/or purchase tobacco products?
2. To what extent are youth access laws *enforced*?
3. What are the associated fees, fines, etc. for violations of youth access laws?

**Counter-Advertising**

1. Is information on the consequences of tobacco use (counter-advertising) propagated nationally and locally?
2. If so, how? Which of the following policies are required by government?
  - Government-issued Health Warning labels on cigarette packs
  - Government-issued Health Warning labels on cigarette advertisements
  - Warnings against underage purchase of tobacco products
  - Warnings of penalties for underage purchase of tobacco products
3. What is the industry's policy on counter-advertising?
4. Does the industry post warnings against underage purchase of tobacco products?

**Access to Smoking Cessation Therapies**

1. Are cessation therapies accessible in the market?
2. If so, then which therapies are available?
  - Pharmaceutical treatments:
    - Nicotine replacement therapies (nasal sprays, microtabs, patches, gum, inhalators)
    - Zyban
    - Nicotine analogs (Tabex)
    - Herbal curatives (Tobaccoff, Nicofree)
  - Non-pharmaceutical methods:
    - Hypnosis
    - Acupuncture
    - Behavioral methods (individual, family, and group therapies; self-control)
  - Cessation accessories:
    - Filters
    - Fake cigarettes
    - Lock-boxes
3. Are they available over the counter, or by prescription only?
4. How are they priced?

included in and excluded from the econometric models. Including highly correlated variables may result in multicollinearity and unstable estimates. Simultaneously, excluding potentially significant and important variables to cigarette demand may produce biased estimates for the impact of price on demand.

## **Bias**

A second complication arises when tax paid sales are used as measures of sales and/or consumption. These measures are likely to be understated, particularly when tax paid cigarette sales are used. More specifically, in those countries or regions where cross-border shopping and smuggling are significant, sales are likely to understate consumption in jurisdictions with relatively high tobacco taxes and prices. At the same time, consumption may be overstated in relatively low tax and price jurisdictions. Failing to account for such factors can produce upward-biased estimates of the impact of price and taxes.

## **Simultaneity**

Another problem in the analysis of aggregate data exists when cigarette (or other tobacco product) prices, sales, and consumption are simultaneously determined—that is, all three measures are determined by the simultaneous interaction of both the supply and demand for cigarettes or other tobacco prices. Failing to account for this simultaneity leads to biased estimates on price. Several studies try to theoretically model the supply and demand for cigarettes, while others use data from large natural experiments (e.g., large increases in cigarette taxes) to avoid the simultaneity issue.

## **Limitations from Units of Measure**

Studies using aggregate data are limited to estimating the impact of changes in prices and other factors on aggregate or per capita estimates of cigarette consumption. As a result, such studies cannot provide information on the effects of these factors on specific issues such as the prevalence of tobacco use, initiation, cessation, or quantity and/or type of tobacco product consumed. Further, these studies do not allow one to explore differences in responsiveness to changes in price or other factors among various subgroups of the population that may be of particular interest (e.g., age, gender, race/ethnicity, socioeconomic status, education).

---

## **Using Individual-Level Data**

Increasing numbers of studies use data on individuals derived from large-scale surveys. In demand models, the estimated price elasticities of demand using individual level data are comparable to those estimated using aggregate data. This occurs for several reasons.

- Taking individual data from surveys helps avoid some of the problems that arise with the use of aggregate data. For example, data collected by individual surveys provide measures for smoking prevalence and consumption of cigarettes—thus avoiding some of the difficulties associated with using sales data as a proxy for consumption.
- Because an individual's smoking decisions are too small to affect the market price of cigarettes, potential simultaneity biases are less likely. Similarly, individual-level income data and other key socio-demographic determinants of demand are less correlated with price and policy variables than among comparable aggregate measures. This creates fewer estimation problems and is likely to produce more stable parameter estimates.
- Using individual-level data allows for the exploration of issues that are more difficult to address with aggregate data, including estimating a separate effect of price and other factors on smoking prevalence, frequency and level of use, initiation, cessation, and type of product consumed. Also, each of these can be examined in the context of various population subgroups.

For example, the Living Standards Measurement Surveys (LSMS) are country-level examples of household surveys conducted in collaboration with the World Bank. The LSMS surveys collect information representative of entire households and for individuals residing within a household. Such survey data allow researchers to explore the effects of individual or general population characteristics (such as gender, age, income, marital status, education, religion, social status, and occupation) on smoker responsiveness to changes in tobacco prices, taxes, availability, and access. Individual level survey data, in particular, allow for the estimation of the impacts of prices and tobacco-related policies on smoking prevalence, initiation, and cessation, as well as on the quantity or type of cigarettes purchased and consumed.

A survey questionnaire may contain one of two types of questions: open-ended and close-ended.

**Open-ended:** a respondent is not given a choice of standardized or predetermined responses.

**Close-ended:** a respondent is provided with a limited set of possible responses and must choose answers from this predetermined set.

There are several approaches or methods to gain responses to survey questions. Survey may be self-administered via mail, self-administered at-home, conducted through an at-home face-to-face interview, or conducted through a person-to-person telephone interview.

Note: Remember about **standardization!** It is important to use the same principles when constructing survey questions. Similarly, it is

imperative to use identical techniques to calculate survey responses and related statistics. This procedure ensures results are comparable across populations.

A number of important variables needed to conduct economic analyses related to tobacco and tobacco control are described below. Examples of survey questions used to gather social and economic information from individual respondents are also provided.

## Consumption

Various forms of tobacco consumption data can be obtained from an individual survey respondent. Obtain individual information on smoking participation (Do you presently smoke?) and the nature of smoking behavior (Do you smoke daily, occasional, never, or are you an ex-smoker?) from carefully designed survey questionnaires.

Box 2.4 contains an example of these types of questions. Questions 2, 4, and 5 show that smoking behaviors, as they relate to the age of a respondent, are also important in defining key consumption data. Use the answers to determine average age of initiation, length of use, extent of addiction and average age of successful cessation.

When survey space is limited, the WHO (1998) recommends using one or more key questions regarding smoking behavior.

Allowance for *one* question:

“Do you now smoke daily, occasionally, or not at all?”

Allowance for *two* questions:

1. Have you ever smoked daily, occasionally, or not at all?
  - Daily
  - 100 or more cigarettes but never daily
  - Not at all, or less than 100 cigarettes in your lifetime
2. How many of the following items do you smoke, chew, or apply per day?
  - Manufactured cigarettes
  - Hand-rolled cigarettes
  - *Bidis*
  - Pipefuls of tobacco
  - Snuff

## Tobacco Price

Self-reported price per pack measures provide an alternative measure of tobacco price (see Box 2.5). Aggregate these reported prices to either a city or regional level to reflect the average local or regional price paid per pack of cigarettes, even though the data is highly endogenous. These price measures can provide a good scale of comparison to the price data collected by governments and/or private agencies.

**Box 2.4. Questions Used to Capture Cigarette Consumption**

Q. Have you smoked at least 100 cigarettes during the course of your lifetime?

- A. Yes
- B. No

Q. How old were you when you began smoking regularly?

Age: \_\_\_\_\_

Q. Do you presently smoke tobacco?

- A. Yes
- B. No

Q. Have you ever smoked tobacco daily for a period of 6 months?

- A. Yes
- B. No

Q. How old were you when you quit smoking?

Age: \_\_\_\_\_

Q. During the past 6 months, did you smoke tobacco daily?

- A. Yes
- B. No

Q. During the past six months did you smoke filtered cigarettes?

- A. Yes
- B. No

Q. How many *filtered cigarettes* do you usually smoke?

Answer: \_\_\_\_\_

Q. During the past six months did you smoke unfiltered cigarettes?

- A. Yes
- B. No

Q. How many *unfiltered cigarettes* do you usually smoke?

Answer: \_\_\_\_\_

Source: WHO, 1998

Price information is not entirely independent of respondents' decisions about whether to smoke and how much to smoke. That is, because surveys collect self-reported cigarette price information from those respondents who already smoke, these sets of reported prices can reflect endogenous choices, particularly when it comes to choice of cigarette brands and cigarette quality. As a result, the price variable may be correlated with unobservable differences in preferences, yielding biased estimates in analyses that depend on this price measure. This produces a number of analytical concerns.

- Smokers who smoke heavily may be more likely than other smokers to seek out lower-priced cigarettes.
- Smokers may be more likely to purchase cigarettes in greater quantities to which significant market discount may apply (e.g., by the carton rather than the single pack).
- Heavy smokers in particular may be prone to smoke less expensive cigarette brands, etc.

**Box 2.5. Open-Ended and Closed Questions About Consumer Cigarette Price****An Open-Ended Question**

Q. What is the price per pack (a pack is 20 cigarettes) of the cigarettes you smoke most often?

Answer: \_\_\_\_\_

**A Closed Question**

Q. How much do you usually pay for a pack of your usually smoked cigarettes?

- A. Do not smoke
- B. Less than \$3.00 per pack
- C. \$3.00-3.49 per pack
- D. \$3.50-\$3.99 per pack
- E. \$4.00-\$4.49 per pack
- F. \$4.50-5.00 per pack
- G. Over \$5.00 per pack

Given any of the above rationales, analyses using these self-reported prices may produce biased estimates of the effects of price on smoking behavior. To help reduce biased price estimates include a few additional questions concerning brand and product type in a survey that already asks for self-reported cigarette price. Box 2.6 contains two such questions.

To help avoid biased estimates, test whether or not the price variable used in the relevant econometric model is exogenous. Apply various estimation methods, including:

- A 2-stage least square (2SLS) estimation with an instrumental variable (IV) approach
- Cragg's (1971) two-part model

Additional methods to help solve the endogeneity problem in self-reported price variables are discussed in greater detail in **Tool 3. Demand Analysis**.

**Box 2.6. Questions to Help Minimize Biased Estimates**

Q. Which brand of cigarettes do you smoke most often?

Answer: \_\_\_\_\_

Q. What size cigarettes do you smoke most often?

- A. Less than 70mm
- B. 70mm (Regular Size)
- C. Over 100mm (King Size)
- D. Other: \_\_\_\_\_

Q. What type of cigarettes do you smoke most often? (Mark all that apply)

- A. Lights
- B. Ultra Lights
- C. Filtered
- D. Unfiltered
- E. Menthol

## Tobacco Taxes as a Proxy for Price

There are two basic methods of levying taxes on tobacco products:

- Nominal or specific taxes are based upon an established amount of tax per individual cigarette or per gram of processed tobacco.
- *Ad valorem* taxes are levied as a percentage of the price of a given tobacco product.

In cases where a cigarette price measure is clearly endogenous to an estimating equation, tobacco taxes may be used as a proxy for retail price. Tobacco taxes are good proxies for tobacco prices, particularly because tobacco taxes (either national or local) are generally independent of an individual's decision to smoke and/or how much to smoke. As a result, the most appropriate proxy for the retail price of a pack of cigarettes is the total per pack tobacco tax.

In cigarette demand equations, the use of a tax per pack measure yields an estimate of tax elasticity. Convert the tax elasticity into a price elasticity in the following linear demand model:

$$\text{Consumption} = \alpha + \beta \text{ Tax} + \mu$$

Here, the tax variable is used instead of a price variable to estimate the demand for cigarettes.  $\beta(t)$  is the estimated coefficient of the tax variable;  $\bar{p}$  is the sample mean of the cigarette price;  $\bar{y}$  is the sample mean of per capita cigarette consumption; and  $\partial p / \partial t$  is the change in cigarette prices resulting from a change in excise taxes. This can be estimated by regressing price as a function of tax where the estimated coefficient of tax ( $\gamma$ ) is  $\partial p / \partial t$ . Thus:

$$\text{Price} = \alpha + \gamma \text{ Tax} + \mu$$

In econometric models, variation in data points allows for statistically significant and sound findings. This is also true for price data measures. Cigarette demand studies typically obtain variations in price from tax differences across time and jurisdictions. For example, in the United States, each state has a different level of cigarette tax and a single cross-section of a national survey has considerable variation in tax measures. On the other hand, tax levels in most developing countries—particularly smaller countries—rarely vary within country as local taxes are rarely levied. Here, one, two, or even three years of household or individual level survey data does not provide enough variation in prices or taxes for adequate use in statistical analyses.

In most countries, cigarettes are frequently taxed at different rates based on length, production size, quality, type, manufacturing process (hand-made, machine-made), and origin. Once characteristics of the cigarettes that individuals smoke are identified from the survey data, there may be enough tax variation within a single cross-sectional sample. If there is no information on tobacco product characteristics other than price, then one should find other

**Box 2.7. Questions About Per Capita Income and Household Income****Per Capita Income Question**

Q. What is your household's total net per capita income per month (include all employment, investment, and governmental or non-governmental benefit earnings)?

Answer: \_\_\_\_\_

**Household Income Questions**

Q. What is your net total household income per month (include all employment, investment, and governmental or non-governmental benefit earnings)?

Answer: \_\_\_\_\_

Q. How many people constitute your household? (Mark one reply)

- A. 1
- B. 2
- C. 3
- D. 4
- E. 5
- F. 6
- G. 7
- H. 8
- I. 9

sources showing more detailed price information by type, size, quality, origin, and so on. This information is generally available from commerce departments and/or customs and tax administration departments in a country's Ministry of Finance. Such information lets researchers use price to figure out the types of cigarettes smoked and assign a corresponding tax level. Researchers should be aware of price variations of brands in urban versus rural areas, and across different types of points of sale.

**Measures of Income**

A survey respondent is often asked to provide information on the amount of his or her income (refer to Box 2.7 for sample questions). Common income measures include net total per capita income per month or net household income per month or, similarly, gross total per capita income or gross household income per month. A survey may also ask for net total household income and, *in a separate question*, inquire about the number of persons residing in the household. This format lets a researcher obtain information about household size, household income, and per capita income.

**Using a Proxy for Income**

Surveys often ask individuals or households for information on:

- Education<sup>1</sup>

<sup>1</sup> In addition to income effects, education also has a negative effect on smoking decisions. That is, educated individuals are more likely to have access to information on the adverse health impacts of tobacco use and, therefore, may reduce their tobacco consumption even as income levels rise. To this extent the overall effects of

- Self reported standard of living
- Occupation

These measures are often highly correlated with measures of income. That is, as educational attainment, standard of living, or level of occupation increases, so does the associated level of income earned. As a result, these variables serve as good proxies of per capita income or household income. Refer to Box 2.8 for sample questions about proxies of income.

## Socio-Demographic Information

Other socio-demographic data which can easily be asked in individual surveys include measures for: age, gender, race, ethnicity, religious denomination, religious participation, religiosity, marital status, number of children, household structure, employment status, type of employment, educational attainment, and area of residence. Prior research shows that these socio-economic and demographic factors can be important determinants of tobacco use, expenditures, and other related issues. Consider the following examples of individual level socio-economic and demographic information.

### Age

Surveys can ask a respondent his or her age or inquire about the respondent's date of birth. Once the ages of respondents are known, define age groups (e.g., 16–25, 26–40, 41–55, etc.) and construct dummy indicators of each age range variable. Assign each person a value of 1 for the age group variable corresponding to his or her current age.

### Religion

Some religions are openly opposed to smoking and other addictive or substance use behaviors. These religions include Mormons in the United States or Muslims in Egypt. Surveys often aim to identify the religious denomination, religious participation, and/or religiosity of respondents. See Box 2.9 for sample questions about religious beliefs and importance.

---

## Use Caution when Interpreting Model Results Based on Individual-Level Data

Like aggregate data, analyses that use individual-level data also face a number of challenges.

---

this income proxy on tobacco consumption is dependent upon which of the two effects—income or information—is stronger.

**Box 2.8. Questions About Proxies of Income**

**Education as a Proxy for Income**

- Q. What is your educational background?
- A. Less than or equal to primary education
  - B. Technical/Vocational school
  - C. Less than high school
  - D. High school
  - E. Some technical schooling beyond high school
  - F. Some college level schooling
  - G. A college degree

In cases where income levels of a child, youth or young adult are needed, question regarding parental education may be used.

- Q. Did your parents (mother, father) attend college?
- A. Neither father nor mother or father attended college
  - B. Father attended college
  - C. Mother attended college
  - D. Both father and mother attended college
- Q. How far did you father (mother) go in school?
- A. Less than high school
  - B. High school
  - C. Some college or technical schooling beyond high school
  - D. Four-year college degree or more
  - E. Don't know
  - F. Not applicable

**Standard of Living as a Proxy for Income**

- Q. How would you best describe your standard of living?
- A. Very good
  - B. Good
  - C. Fair
  - D. Rather poor
  - E. Very poor

**Occupation as a Proxy for Income**

- Q. Which best describes your current occupation?
- A. Management
  - B. Unskilled
  - C. Skilled
  - D. Farmer
  - E. Self-employed
  - F. Student
  - G. Disabled
  - H. Unemployed
  - I. Housewife
  - J. Do not work

**Bias**

Data may be subject to an ecological bias in that omitted variables that do affect tobacco use are correlated with the included variables. Excluding such variables may produce biased estimates for the included variables.

Furthermore, the use of individual-level data is subject to potential reporting biases. A comparison (Warner, 1978) of self-reported consumption with aggregate sales data shows that survey-based, self-reported consumption understates actual sales. Potential underreporting of consumption can cause problems in the interpretation of estimates produced from using individual-level data. In general, studies using individual-level data assume that the extent of underreporting among respondents is proportional to their actual level of use. This assumption implies that the estimated effects of price and other factors will not be systematically biased. However, this assumption has yet to be demonstrated.

Finally, just as with aggregate data, failure to account for differences in cigarette prices across countries or regional borders means

**Box 2.9. Questions About Religion**

Q. What is your religious denomination?

- A. Atheist
- B. Catholic
- C. Jewish
- D. Muslim
- E. Protestant
- F. Other religion: \_\_\_\_\_

Q. How religious are you?

- A. Very religious
- B. Somewhat religious
- C. Little religious
- D. Not religious

Q. Do you participate in religious services or practices?

- A. Yes, a few times per week
- B. Yes, once per week
- C. Yes, once or twice per month
- D. Yes, few times per year
- E. Do not participate in religious services or practices

Q. How would you describe your position in relation to your faith?

- A. Religious and regularly attend services
- B. Religious but irregularly attend services
- C. Religious but do not practice my faith
- D. Atheist

Q. How important is your religion in your decision not to smoke?

- A. Very important
- B. Important
- C. Somewhat important
- D. Not important

elasticity estimates may become biased (towards zero). When using individual level data, one often has information on where an individual resides. Studies that rely on individual-level data often use a number of approaches to control for potential cross-border shopping that result from differing tobacco prices. Some studies limit their samples to those individuals who do not live near lower-price localities (Lewit and Coate, 1982; Wasserman *et al*, 1991; Chaloupka and Grossman, 1996; Chaloupka and Wechsler, 1997). Other analyses include an indicator of a price differential (Lewit *et al*, 1981; Chaloupka and Pacula, 1999). Other studies use a weighted average price based on the price in the own-locality as well as on the price found in nearby localities (Chaloupka, 1991).

### ***Limitations of Available Data***

Another limitation to using individual-level survey data is that data on price, availability, advertising, policies, and other important macro-level determinants of demand are generally not collected in the surveys. As a result, many relevant variables may be omitted from the analysis.

### ***Measure Changes over Time***

Use and refer to repeated surveys to accurately monitor changes in tobacco use, prices, etc. over a period of time. Repeat surveys that use identical data collection techniques in each consecutive survey help ensure the comparability of the data over time.

# IV. Data Preparation and Management: Easy Steps to Building Your Own Database

---

## Choose a Software Package

Today, statistical researchers and analysts can draw upon a number of computer-based tools that facilitate data manipulation, variable construction, and analysis. In general, a successful software package is flexible and easy to use, yet powerful enough to handle large amounts of data in the shortest amount of time possible.

The selection of a software package is most dependent on budgets and desired program performance. The market price of statistical software packages varies from producer to producer, as does the power and sophistication of the software. For the purposes of this tool, a “good” software package should provide the following at an affordable price:

- Ease of data access
- Sufficient capacity to manage and manipulate data
- Availability of moderately advanced statistical tools
- Capability to present analysis results easily and clearly

The statistical software market offers a number of variously equipped packages within a wide range of prices.

### Spreadsheets

In recent years, traditionally easy to use spreadsheet packages have been greatly improved and, as a result, have become quite sophisticated analysis tools. For example, Microsoft Excel and Corel Quattro Pro are two popular spreadsheet programs that are compatible with all Microsoft Windows operating systems.

Spreadsheet programs are easily accessible and are almost always included with the basic software of a new computer. In general, spreadsheet programs, although reasonably priced, offer straightforward methods for data access and manipulation, yet provide only moderate capabilities for statistical data analysis. Because of their limited space and computing power, spreadsheet programs are really only equipped to handle aggregated sets of data.

## **Statistical Packages**

Popular higher-powered statistical programs include SAS, SPSS, and STATA. These packages are equipped to handle much larger bodies of data than spreadsheet programs. All three statistical packages offer a large array of data manipulation and data analysis tools at various prices.

### ***Retail Price***

The statistical packages mentioned above differ in retail price. As a result, the choice in packages used may be determined by a researcher's budget. In terms of retail price, SAS is one of the most expensive statistical programs on the market. In addition to the cost of purchasing the package, SAS requires annual updates to its corporate license. On the other extreme, STATA is quite affordable, with standard packages purchased in bulk costing as little as \$50 a copy. SPSS falls in the middle in terms of price, though it too may be considered too costly for some researchers.

### ***Capacity***

Statistical packages are much better equipped to manipulate and maintain datasets than spreadsheet packages. Statistical packages, particularly SAS, can house very large sets of data (measurable within the gigabyte range)—though the ultimate amount of data storage may be dependent upon the memory of the computer hosting the program. A spreadsheet, on the other hand, has much more limited storage capacity (approximately 300 columns and 65,000 lines of data).

### ***Data Management***

The grid-like nature of spreadsheet programs makes it very easy to view data and use functions or equations to create new variables. However, spreadsheets are limited in the calculations that they can perform. In addition, they do not allow for easy merging with other aggregations and types of data files.

Statistical packages require only a few lines of code to create new variables, and can merge several different datasets as well as sort and aggregate data. Such data manipulations are carried out quickly and efficiently, even with very large sets of information.

### **Statistical Tools**

In recent years, spreadsheet programs have been enhanced to perform a number of sophisticated operations. Most spreadsheet packages today are equipped to calculate statistical summaries of data and produce basic ordinary least squares and logit estimates. As a result, studies using small datasets and requiring only basic regression models can easily make due with Microsoft Excel, Corel Quattro Pro, or comparable spreadsheet programs for the study analysis. However, statistical packages are much more powerful than conventional spreadsheet programs, as they contain advanced modeling tools and various statistical tests to allow for sophisticated econometric estimation.

### **Presentation**

In terms of presentation, spreadsheets are better equipped to create elaborate tables, figures, and graphs. One can quickly and effectively plot or graph data for presentation. In addition, these graphs can be individually saved and imported into other word processing or presentation programs. Although statistical packages have the capability to plot observations, they do so at only the most basic level. SAS offers a graphing package that can be purchased at an additional cost. However, this supplemental package is complicated and requires relatively large amounts of programming.

---

## **Manipulate the Data**

The first task in any analysis requires a careful cleaning of the collected data. A thorough edit requires checking the data for consistency, completeness, and accuracy. A careful edit includes noting inconsistencies in data values and monitoring the number of missing answers.

Practice these steps to eliminate many data inconsistencies (WHO, 1998):

1. Check the values of the data against the survey questionnaire.
2. Enter the data into the computer more than once.

### **Read the Raw Data**

When either aggregate or survey data is received from an institution or data agency, the dataset may require some manipulation and cleaning before it is ready for even the most basic forms of statistical or econometric analysis. That is, raw data must first be converted into a form that can be read by one or more statistical estimation packages (e.g., SAS, SPSS, LIMDEP, RATS, TSP, Microfit, STATA). Most often, raw data files are imported into statistical programs in the form of ASCII or text format data files (files with

.txt or .csv extensions to their names). For additional information and guidance on reading data files into SAS, refer to the second chapter of *The Little SAS Book* (Delwiche and Slaughter, 1998).

### **View an ASCII Data File**

An ASCII file (a file with an .asc extension to the name) is a simple text file containing rows and columns of numerical information. An ASCII file can be easily opened and viewed in a Word, WordPad, or Notepad file. Figure 2.6 shows an ASCII file containing 1994 individual-level survey data from Poland. The name for this file is *data94.asc*.

In order to be used further, the data in Figure 2.6 must be read into either a spreadsheet or directly into a statistical package.<sup>2</sup>

### **Import a Text File into a Spreadsheet**

The largest advantage of importing an ASCII file directly into a spreadsheet is the ease of viewing and manipulating data. Most spreadsheets allow even a beginning researcher with limited experience to quickly and easily view data. In addition, formulas can easily be constructed to quickly capture descriptive statistics of the raw data. Finally, sophisticated graphics can be easily produced using the spreadsheet program. [Strictly for sake of simplicity within this Tool, Microsoft Excel is used as the software example for importing and manipulating data files.]

**Figure 2.6. An ASCII Data File**

```

11340.3.211112112. . . . . .276432.36
3.....222222122. . . . . .258122.25
110...212121129111 .12511151 32 .1541211.2
140...212111111111 .19911202 .2 .2506612.3
110...222121112111 .12111202 .2 .1674211.2
11230.211111111111 .13011 42 .2 .250112.42
110...1.221111222. . . . . .276732.34
110...1.2211112511 .12011202 .2 .128712.12
21230.3.222222212171152. . . . .1631411.3
11230.1.122111222. . . . . .278112.32
110...1.221111112. . . . . .2401414.5
21230.1.111311212. . . . . .276432.36
210...3.1121122211 .11811302 .2 .171122.52
210...3.1121122211 .12011202 .2 .1392211.2
110...1.111111112. . . . . .269142.23
110...1.2213211212221202. . . . .2686411.3
2122..32111...112221 331..26 11 .2221667.9
22021.2.332212 11111 454.1 .2 .155212.33
311...3.222221111355 . . .711 2. .2553669.9
3144..222331. 114431111...1 1. .551553.11

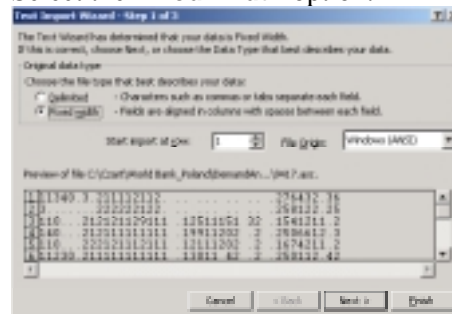
```

<sup>2</sup> Although individual-level data are shown as an example, identical steps apply to ASCII and other text files containing aggregate measures.

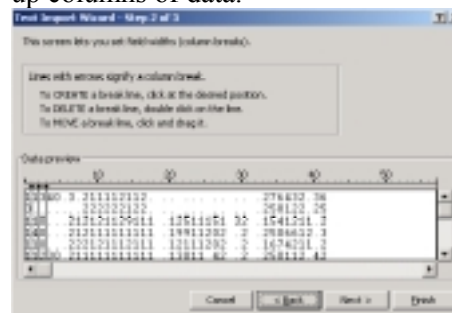
To import the text file presented in Figure 2.6 into Microsoft Excel, follow these steps:

1. Open the Microsoft Excel program.
2. Open the ASCII file by clicking on **File | Open** in the command bar at the top of the screen. (To locate the text file on your computer, in the **Files of Type** list select “All Files.”)
3. Select the desired file and click **Open**. In this example, file *data94.asc* is opened.
4. Use the **Text Import Wizard** to properly open the ASCII file, as outlined in Steps A–C below.

A. Select the **Fixed Width** option.



B. Click on the vertical lines in the **Data Preview** window and place them in such a manner as to break up columns of data.



In order to properly segment columns in this text file, a codebook or index of column codes is needed. Once the data enters this software environment, raw data columns must be assigned with appropriate variable names. Often, an index or codebook containing information for each column of survey data accompanies the raw survey data. Table 2.2 contains an example of a column index.

5. Use the index of column codes to assign titles (in this example, column titles read p45 through m10) to each

column of data in the Microsoft Excel file.<sup>3</sup> Table 2.3 provides an example of how to assign column headings to file *data94.xls*.

6. Save this data file as a Microsoft Excel spreadsheet (for purposes of this example, let's call this file *data94.xls*). Be sure to note the location of the data file and close Excel.

**Table 2.2. Codebook/Index of Column Codes**

Question	Column
p45	1
p4601–4605	2–6
p47	7
p48	8
p49	9
p50	10
p51	11
p52	12
p53	13
p54	14
p55	15
p56	16
p57	17
p58l	18
p58,	19–20
p59l	21
p59m	22–23
p60	24
p6101l	25
p6101m	26–27
p6102l	28
p6102m	29–30
p6103l	31
p6103m	32–33
m1	34
m2	35–36
m4	37
m6	38
m7	39
m8	40
m9	41
m10	42

Source: Polish Data Institution

<sup>3</sup> In this example, a researcher must manually assign and enter titles to each column. Typically in spreadsheet programs, column headings are simply typed into the top cell of each column. In SAS, STATA, or other statistical programs, a program must be written and called to assign a variable name to each column of data.

**Table 2.3. Column Headings for an Excel Worksheet**

	p45	p46_01	p46_02	p46_03	p46_04
	1	1	3	4	0
	3	—	—	—	—
	1	1	0	—	—
	1	4	0	—	—
	1	1	0	—	—

The ASCII file *data94.asc* has been successfully transformed into a Microsoft Excel spreadsheet file called *data94.xls*.

### **Import a Spreadsheet Data File into SAS**

Once a dataset exists in a spreadsheet environment (or, in the case of our example, a Microsoft Excel spreadsheet), it can easily be moved for use into, for instance, SAS statistical software.

In order to import our example file *data94.xls* into SAS, the code presented in Figure 2.7 must be constructed in the **Program Editor** window of version eight of SAS. This import procedure successfully imports Excel file *data94.xls* into SAS to create a permanent SAS dataset with a file called *data94.sd2*.

### **Import a Text File Directly into SAS**

The ASCII and programming codes presented as examples above assume that a researcher would prefer to import a text file into a spreadsheet environment before importing that data into a more sophisticated statistical package. Researchers who are comfortable

**Figure 2.7. Code to Import an Excel File into SAS**

```

/*****
  Set SAS Options
  *****/
  OPTIONS mprint ps=55 ls=100 ERROR=1;
  Libname cc 'c:\Data\SASData\Indata';
  Libname cbos 'c:\Data\SASData\Outdata';
  Libname wb 'c:\Data\Raw';

/*****
  Use the Proc Import Command to read in xls file into SAS
  Create Permanent SAS data file: cbos.data94.sd2
  *****/
PROC IMPORT OUT=cbos.data94
  DATAFILE="C:\Data\SASData\Raw\data94.xls"
  DBMS=EXCEL2000 REPLACE;
  GETNAMES=YES;
RUN;

```

with statistical software and their datasets may instead prefer to move their data directly into SAS to conduct simple analyses, plots, and summary statistics, thus saving themselves a few steps.

The code shown in Figure 2.8 is designed to directly import a .csv (or other type of text file) into SAS and also create a permanent SAS dataset called *cbos.data94.sd2*.

## Check the Quality of the Raw Data

Before analyzing data and reporting the findings of a study, one should begin with a quality check of the data by using a series of basic tests. This helps assure that the observations and variables are in order.

Because some datasets, particularly survey datasets, are quite large, it is almost impossible to look at the entire dataset to find outliers or other odd quirks. As a result, there are two procedures one can use to quickly scan and troubleshoot the data.

1. Generate descriptive statistics for each of the raw variables.
2. Check the frequencies in the values of each variable.
3. In the case of aggregate level data, plot the figures

## Calculate Summary Statistics for Each Raw Variable

The SAS statistical package offers several statistical procedures that can be used to generate summary or descriptive statistics for each raw variable acquired in a dataset. The two most commonly used procedures are a means procedure and a univariate procedure. Although both procedures generate largely similar output, the commands used to produce summary statistics in a means procedure

**Figure 2.8. Code to Import a Text File into SAS**

```

/*****
  Read in Raw Data and Create Temporary SAS file x
  Note: use delimiter=', ' as this raw csv file is comma delimited use
  firstobs=2 as data begins in row #2
  *****/
  Data x;
  infile 'c:\Data\Raw\data94.csv' delimiter=', ' firstobs=2 missover;
  input   p45 p46_1 p46_2 p46_3 p46_4 p46_5 p47 p48 p49 p50 p51 p52
p53 p54 p55 p56 p57 p58l p58m p59l p59m p60 p6101l p6101m
p6102l p6102m p6103l p6103m m1 m2 m4 m6 m7 m8 m9 m10;
  Run;

/*****
  Create Permanent SAS file cbos.svy9412.sd2
  *****/
  Data cbos.data94;
  Set x;
  Run;

```

(known as “Proc Means” in SAS) are the most straightforward.

Figure 2.9 contains code with a Proc Means command to generate a set of descriptive or summary statistics for our raw SAS dataset file *data94.sd2*. This code is entered and run from the **Program Editor** Window of SAS. The Proc Means requests that SAS generate the following seven individual pieces of information about the dataset:

1. N (the number of non-missing observations)
2. Nmiss (the number of missing observations)
3. Mean (a variable’s means)
4. Min (a variable’s minimum value)
5. Max (a variable’s maximum value)
6. Sum (the sum of values contained in a single variable)
7. StDev (the standard deviation in values of a single variable)

The output generated by this procedure is displayed in a SAS **Output** Window (see Figure 2.10), presented as follows:

**Column 1 Variable.** A list of all the variables contained in the SAS version of the raw data (file name: *data94.sd2*).

**Column 2 N.** The number of *non-missing observations* reported for each variable.

**Column 3 N MISS.** The number of *missing observations* reported for each variable. (**Note:** The sum of columns 2 and 3 equals the total number of observations in the dataset.)

**Column 4 MINIMUM.** Across variables, this is the *lowest* value reported for each variable in the dataset.

**Column 5 MAXIMUM.** Across variables, this is the *highest* value reported for each variable in the dataset.

**Column 6 STD DEV.** The average standard deviation between the values reported for each variable in the set.

**Column 7 Sum.** The sum of all values reported for each variable in the dataset.

**Figure 2.9. Code to Generate a Set of Summary Statistics from a Raw SAS Dataset**

```

/*****
  Produce Summary Statistics including: Average, Minimum, Maximum,
  Sum and Standard Deviation values.
  *****/
PROC MEANS data=cbos.data94  N Nmiss mean min max sum stdev;
RUN;

```

Figure 2.10. SAS Window Showing Means Output

Variable	N	N Miss	Minimum	Maximum	Mean	Std Dev	Sum
p45	1841	0	1.0000000	3.0000000	1.5782501	0.6473833	1642.00
p46_1	350	31	1.0000000	5.0000000	1.1915709	0.5977942	1132.00
p46_2	348	33	0	5.0000000	1.9786751	1.2009937	1815.00
p46_3	452	593	0	5.0000000	1.5685841	1.5849924	703.0000000
p46_4	227	814	0	5.0000000	0.9471366	1.6304685	215.0000000
p46_5	54	387	0	5.0000000	0.4253253	1.3402538	23.0000000
p47	950	31	1.0000000	3.0000000	1.7157895	0.8742979	1638.00
p48	148	893	1.0000000	3.0000000	1.3783704	0.5649957	204.0000000
p49	1841	0	1.0000000	2.0000000	1.5869356	0.4926289	1652.00
p50	1841	0	1.0000000	5.0000000	1.6772304	0.7361219	1746.00
p51	1841	0	1.0000000	5.0000000	1.6772304	0.6632898	1746.00
p52	1841	0	1.0000000	3.0000000	1.2853086	0.6302579	1338.00
p53	1841	0	1.0000000	3.0000000	1.4384616	0.8312134	1488.00
p54	1841	0	1.0000000	3.0000000	1.5792507	0.6261295	1644.00
p55	1841	0	1.0000000	5.0000000	1.3548511	0.5002553	2835.00
p56	1841	0	1.0000000	5.0000000	1.7875731	1.0392415	1778.00
p57	1841	0	1.0000000	2.0000000	1.4681345	0.4906478	1526.00
p581	523	478	1.0000000	2.0000000	1.2384614	0.4573635	731.0000000
p58m	184	375	5.0000000	99.0000000	66.2188375	22.4838391	7346.00
p591	582	479	1.0000000	2.0000000	1.1848222	0.3069936	621.0000000
p59m	583	538	7.0000000	99.0000000	25.8926441	21.5897596	13824.00
p60	583	538	1.0000000	5.0000000	1.3335560	0.6173853	671.0000000
p61011	367	674	1.0000000	2.0000000	1.1361853	0.3976525	433.0000000
p6101m	235	746	2.0000000	99.0000000	16.7423729	8.3976483	4339.00
p61021	358	873	1.0000000	2.0000000	1.7984783	0.4053584	659.0000000
p6102m	76	365	1.0000000	99.0000000	21.8447360	13.2459785	1645.00
p61031	368	673	1.0000000	2.0000000	1.3864130	0.1159251	731.0000000
p6103m	5	1036	2.0000000	13.0000000	6.9886000	5.1478151	39.0000000
m1	1841	0	1.0000000	5.0000000	1.5485110	0.5969929	1612.00
m2	1841	0	15.0000000	99.0000000	51.5394286	16.2314371	53715.00
m4	1841	0	1.0000000	5.0000000	3.3458213	2.3137358	3483.00
m6	1841	0	1.0000000	2.0000000	2.5388680	1.7337587	2747.00
m7	1841	0	1.0000000	2.0000000	1.5292508	0.4933898	1532.00
m8	497	544	1.0000000	5.0000000	1.8511666	1.4262578	929.0000000
m9	354	487	1.0000000	5.0000000	2.1289386	1.4169646	1175.00
m10	1841	0	1.0000000	5.0000000	3.9236540	1.5589218	3147.00

### Generate Frequencies

A frequency check (called a “Proc Freq” in SAS) on variables contained in a dataset lets a researcher easily and correctly determine the range and frequency of values for any given variable. As discussed below, the output from a frequency check is also needed to correctly reconstruct raw survey observations into informative econometric and statistical variables. Figure 2.11 shows the procedure in SAS to generate information on the frequency of values for each and every variable in dataset file *data94.sd2*.

As in the case of a Proc Means, the output generated by the Proc Freq command is displayed in a SAS **Output** window. Figure 2.12 shows the Proc Freq output for variables p45 p46\_1 p46\_2 from dataset file *data94.sd2*. The results of the Proc Freq for raw variable p45 show that this variable takes on several values ranging from 1–3.

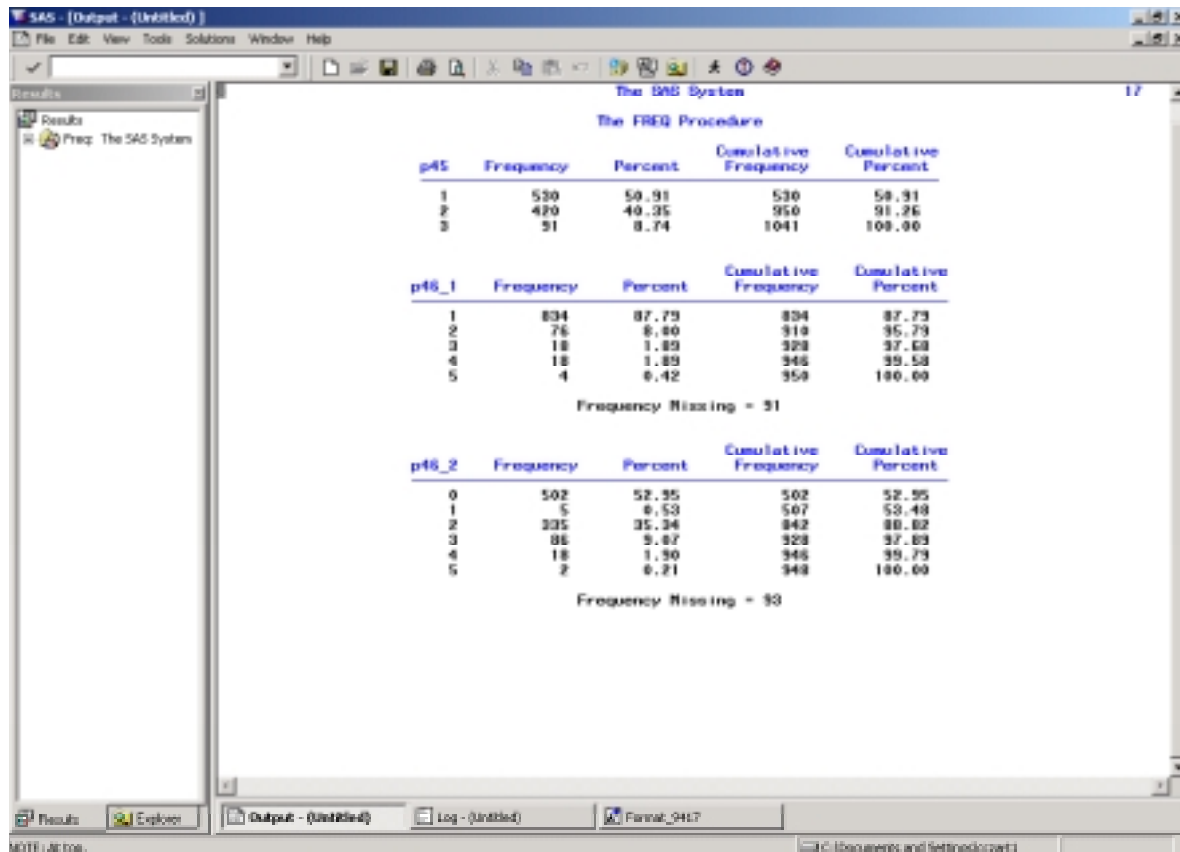
Figure 2.11. Code to Generate a Set of Value Frequencies from a Raw SAS Dataset

```

/*****
  Produce Frequencies of each variable. Save Output as a codebook.
  *****/
PROC FREQ data=cbos.data94;
TABLES p45 p46_1 p46_2 p46_3 p46_4 p46_5 p47 p48 p49 p50 p51 p52
       p53 p54 p55 p56 p57 p581 p58m p591 p59m p60 p61011 p6101mp61021
       p6102m p61031 p6103m m1 m2 m4 m6 m7 m8
       m9 m10/list;
RUN;

```

Figure 2.12. SAS Window Showing Frequency Output



Among the 1,041 respondents captured by variable p45:

- 50.9% of respondents (530 persons) reported a value of 1
- 40.35% (420 people) answered 2
- 8.74% (91 respondents) reported a value of 3

This range of values captures responses for the full sample of 1,041 persons surveyed. As a result, there are no missing observations for variable p45.

The results for the frequency on variable p46\_1 reveal that the range in values for this variable stretches from 1–5. Here, from among the 1,041 total possible respondents: 87.79% (834 respondents) reported a value of 1; 8.00% (76 persons) reported a value of 2; and so on. The frequency on this variable captures values reported by only 950 respondents. A remaining 91 respondents did not provide a value (answer) for variable 46\_1. As a result, the *Frequency Missing* field reports a value of 91.

Once raw items of information are successfully imported into a statistical program and all visually apparent peculiarities are identified and fixed, additional steps can be taken to prepare a strong set of variables for statistical analysis. These steps include:

- Constructing new variables

- Importing other sources of data
- Cleaning the final dataset

## Plot the Data (Aggregate Level Data Only)

In the case of aggregate figures, it is a good idea to plot the data in order to take a look at the trends that appear in the number over time. Figure 2.13 shows the correct SAS procedures to use for a simple data plot.

Note: Large or excessive year-to-year variations are generally unusual and, therefore, should be double-checked and viewed with caution. Minor fluctuations in the data can be removed using a statistical technique called smoothing or moving averages. See the **Create Moving Averages—Smoothing** section for further details.

**Figure 2.13. Code to Generate Plots of an Aggregate Level Raw Data Set**

```

/*****
  Plot Aggregate of Per Capita Consumption Variable
  *****/
PROC PLOT data=cbos.aggregate;
Plot cigs*year;
RUN;

/*****
  Plot Aggregate of Smoking Participation Variable
  *****/
PROC PLOT data=cbos.aggregate;
Plot smoke*year;
RUN;

```

## Create New Variables

Many survey questions have answers that, taken at face value, are difficult if not impossible to tabulate and quantify. Such raw variables may require recoding before they can be used for statistical modeling.

- Dichotomous variables are those whose survey questions have simple “yes” or “no” answers. Recoding the variable so that each answer has a unique, numerical value is necessary. For example, the variable SMOKE is constructed for the question “Do you currently smoke?” and has two possible answers, “Yes” or “No”. Assign the variable a value of 1 if the respondent is a smoker and has a value of 0 if the respondent is not a smoker.
- Categorical variables are those that take on a range of values, depending on possible responses to survey questions. For example, possible answers to the question “What is your

level of education?” include: “No education”, “Completed elementary level schooling”, “Completed high school”, and “Obtained a college degree”. Here, the variable EDUCATION takes on some whole number value between 1 and 4 for each respondent where the variable is coded 1 for respondents with no education, 2 for respondents with elementary schooling, 3 for those who have completed secondary or high school education, and 4 for completion of a college degree.

- Continuous variables are for those questions (e.g., age, income) whose answer will already have a unique, numerical value. For instance, the variable AGE is set equal to the actual age of the respondent.

Depending on the purpose of the research and the interest of the researcher, each of the variables can be recoded into additional variables. For example, transform the variable AGE from a continuous variable with values of 18, 19, 20, and so on into a categorical variable with a value of 1 for all respondents under the age of 18, a value of 2 for respondents between 18 and 21, a value of 3 for respondents between 21 and 31, and so on. Or, construct a group of variables to reflect individuals from different income strata. That is, create different income group variables (e.g., poor, middle income, high income) where the variable POOR equals a value of 1 for all respondents earning less than \$2,000 per month; the variable MIDDLE equals a value of 1 for all respondents earning over \$2,000 per month but less than \$4,000 per month, and the variable HIGH equals a value of 1 for all respondents earning \$4,000 per month or more.

## **Merge Datasets**

Often, researchers want to include specific sets of information in their analyses. This may require the use of additional data sources that contain information on specific variables. When using more than one dataset, merge them together. This is a common and often necessary practice for a study. For example, when a household survey and a survey of individuals both contain tobacco-related information, it will enrich and simplify analyses to merge two datasets together. Household size can be gathered from household survey data while employment status, education level, and marital status for an individual member of a household are obtained from individual-level survey data.

Merge two or more datasets together only if they have one or more fields or variables in common. For example, every household in a survey is usually assigned two identification numbers. One is the household identifier and the other is an identifier for the household's location (city, region, state). When individuals in households are surveyed, they are assigned an individual identification number as well as the same household identification number used in the household survey. As a result, household level survey data can be

merged with individual level survey data to create a larger and more comprehensive dataset.

In another situation, household survey data and individual-level survey data may lack information on particular variables important to the analysis. In this case, try to find another dataset to merge and use with the household and individual level data. For example, if income information is not contained in either the household or individual level data, try turning to state, city, or regional data for an average income measure. The national or central statistical offices of most countries usually collect such information.

*Remember the collection level of any merged data you analyze, and use your best judgment when interpreting such data.*

Use your judgment when deciding on income measures that are most suitable for a household-level or individual-level analysis. State, province, or city identification numbers may make it possible to merge household data with government income data. However, after datasets are merged and results are obtained, it is important to remember the level of information when interpreting results. For example, in cigarette demand analyses, if the demand for cigarettes is estimated at the household level, but the income variable is an average city income measure, interpret the result as changes in household consumption (by packs, pieces, etc.) in response to changes in the average city income.

## Clean the Data

Once datasets are merged and necessary variables are created, then the next step is to filter or clean the data of inconsistencies in coding. The most important filtering approach is to deal with missing or miscoded information.

First, once a dataset is constructed, examine the descriptive statistics (including the standard deviation, mean, minimum, maximum, and the number of observations) for each variable of interest that is contained in the dataset. Look for missing values, outliers, and miscoded information. If a variable is missing information for one or more observations, then the total number of observations will differ from the total number of observations for other variables contained in the dataset. When checking minimum and maximum values, identify outliers or miscoded variables. For example, gender variables are generally coded where a male takes a value of 1 and females take a value of 0 (or *vice versa*). If there is “2” in the maximum value for the gender variable, it is clear that one or more observations are miscoded or incorrectly imported into the working dataset. Similarly, when checking income values, a 0 value or an extremely high value flags a potential problem with the data. If it is not possible to clean problematic variables from odd values, it is typical to delete or drop observations with missing or miscoded information from the estimation.

Second, it is important to look at frequency tables and the distribution of the variables as a method for checking the data and deciding on an appropriate model form or data transformation. For

example, the distribution of cigarette consumption is usually skewed. Given this, consider using the log of consumption in the demand model and choose an appropriate estimation method (linear, log linear, two-part model, etc.) for the analysis. If, for example, the data show that 80 percent of individuals are male and only 20 percent are female, then there appears to be selection bias in the survey data, and it may be a good idea to either weight the data or stratify the sample by gender.

### ***Input Missing Values***

Most statistical packages automatically drop observations with missing values from the estimation. When the remaining number of observations is not large enough, a researcher may not want to drop observations with missing values and instead try to input values. For example, if the income variable is missing, a researcher might input income based on the income levels of other households with the same characteristics. An alternative approach is to run regression for income as a function of other characteristics (e.g., age, education, occupation) and use the regression result to predict or estimate income for observations with missing values for income. Before running the regression, assign all missing variables a value (0 or 1) so that observations are not dropped from the regression.

The following is a regression equation used to input income where the level of income is a function of age, gender, either neighborhood, size, type, or location of house, education attained, marital status, occupation, and assets:

$$\text{Income} = f(\text{age, gender, house, education, marital status, occupation, assets})$$

This regression technique estimates income for each observation. Use the regression coefficients to estimate income for observations with missing income values, where the income of individual  $i$  is estimated as (the constant term + age)  $\times$  (coefficient for age + sex)  $\times$  (coefficient for sex + each other variable in the equation)  $\times$  the variable coefficient.

### ***Adjust for Reporting Bias in Cigarette Consumption***

As mentioned earlier in this Tool, individuals often underreport their actual levels of consumption of tobacco and alcohol. Therefore, be sure to adjust the survey data to agree with aggregate sales data (and carefully explain how and why this is done). In making this adjustment, know what fraction of the total population is covered by the survey and how representative the sample is of the total population. At the very least, estimate the degree of underestimation and acknowledge the reporting bias in your results.

### **Create Moving Averages—Smoothing**

Annual fluctuations in time series data such as annual cigarette consumption levels may be smoothed over time with the use of a statistical procedure called moving averages.

For example, assume cigarette smoking participation rates are available for years 1980-2000. Assume that rates for years 1980–1985 equal 36%, 37%, 41%, 36%, 35% and 32% respectively. To create a series of average rates based on three consecutive calendar year rates, one must:

1. Calculate the average of the rates for 1980, 1981 and 1982
2. Next, calculate the average of the rates for 1981, 1982, 1983
3. Next, calculate the average rates for 1982, 1983 and 1984 and so on.

The series of average rates that is constructed is based on three consecutive calendar year rates and results in a much smoother trend than the original annual data. As five consecutive years are averaged each time, the new series is called a three-year moving average.

# V. Suggestions for Data Sources

---

## Economic and Tobacco Related Data

### *International Organizations*

Organization: United Nations (UN)

Web site: [http://www.un.org/depts/unsd/sd\\_databases.htm](http://www.un.org/depts/unsd/sd_databases.htm) for listings and links to UNSD statistical databases available on-line

For Helpful Links to National Data Sources See:

[http://www.un.org/depts/unsd/sd\\_nat\\_data.htm](http://www.un.org/depts/unsd/sd_nat_data.htm), and

[http://www.un.org/Depts/unsd/g\\_s\\_natstats.htm](http://www.un.org/Depts/unsd/g_s_natstats.htm)

UN Organization: Food and Agriculture Organization (FAO)

Web site: <http://www.fao.org>

Data on tobacco production and harvest area, producer prices for tobacco leaves, tobacco leaves and cigarette trade (export-import) by volume and value.

UN Organization: International Labor Organization (ILO)

Web site:

<http://www.ilo.org/public/English/bureau/stat/staff/index.htm>

UN Organization: World Health Organization (WHO)

Web site: <http://www.who.int> and specifically,

<http://www.who.int/whosis> for health and health related statistical information from the WHO Global Programme on Evidence for Health Policy

For Helpful Links to Other Resources See:

<http://www.who.int/library/reference/desk/statistics/index.en.shtm>

UN Organization: World Health Organization-Tobacco Free Initiative

Web site: <http://tobacco.who.int/>

For Helpful Links to Other Resources See:

<http://tobacco.who.int/en/research/index.html>

World Health Organization. 1997. Tobacco or Health: a Global Status Report. Geneva, Switzerland. Country-level data on smoking prevalence, cigarette consumption, tobacco production, trade, industry, health impact and tobacco control legislation. Available online at <http://www.cdc.gov/tobacco/who/whofirst.htm>

UN Organization: International Monetary Fund

Web site: <http://www.imf.org>

For Helpful Links to Other Resources See:

<http://www.imf.org/external/pubs/ft/fandd/1999/12/jha.htm>

Data on total revenues, revenues from excise taxes and all taxes.

UN Organization: United Nations Industrial Development Organization (UNIDO)

Web site: <http://www.unido.org>

Tobacco manufacturing employment.

UN Organization: World Bank

Web site: <http://www1.worldbank.org/tobacco/>

World Bank. 1998. World Bank Economic Survey of Tobacco Use: <http://www.worldbank.org/tobacco>. Data on average retail price for most popular domestic and foreign cigarettes, cigarette excise tax, tobacco tax revenue.

World Bank. 1998. World Development Indicators. Washington D.C. General socioeconomic, population, and health indicators for 148 countries and 14 country groups. Select pieces of the database are available at <http://www.worldbank.org/data/wdi/home.html>.

Organization: Organization for Economic Co-Operation and Development (OECD)

Web site: <http://www.oecd.org/oecd/p.../o,3380,EN-statistics-0-nodirectorate-no-no-no-0,FF.htm> for links to free statistical data

### ***Non-Governmental Organizations***

Organization: The International Tobacco Control Network

Web site: <http://www.globalink.org/>

Organization: Research on Nicotine and Tobacco

Web site: <http://www.srnt.org/>

Organization: Research for International Tobacco Control

Web site: <http://www.idrc.ca/tobacco/en/index.html>

Organization: PATH Canada

Web site: <http://www.pathcanada.org/english/tobacco.html> (vietnam)

Organization: Research for International Tobacco Control

Web site: <http://www.idrc.ca/tobacco/en/index.html>

Organization: PATH Canada

Web site: <http://www.pathcanada.org/english/tobacco.html>

Organization: Tobacco Pedia

Web site: <http://www.tobacopedia.org/>

Organization:  
Web site: <http://tobacco.org/>

### **Universities**

Organization: OFFSTATS—Official Statistics on the Web  
University of Auckland Library (New Zealand) lists web sites that offer free access to social, economic and general data from official sources

Web site:  
<http://www2.Auckland.ac.nz/lbr/stats/offstats/OFFSTATSmain.htm>

Organization: Latin American Statistical Sources  
Cornell University Library (United States) lists links to the home pages of National Statistical Bureau

Web site: <http://libl.library.cornell.edu/colldev/lastatistics.html>

Organization: Document Center  
University of Michigan (United States) lists links to agencies that provide access to economic, demographic, agricultural, health, labor, education, transportation and environmental data.

Web site: <http://www.lib.umich.edu/govdocs/statsnew.html>

### **Private Data Agencies**

Organization: Market File  
Web site: <http://www.marketfile.com/market/tobacco/>  
A commercial online tobacco database. Data on cigarette consumption, production, price and tobacco control measures. Subscription is required to obtain access to these data.

Organization: The Tobacco Manufacturers' Association (TMA)  
Web site: <http://www.the-tma.org.uk/miscellaneous/main.htm>

Organization: The Retail Tobacco Dealers of America, Inc.  
Web site: <http://www.rtda.org/>

Organization: AC Nielsen Inc.  
Web site: <http://www.acnielsen.com>

### **US Agencies with International Interests**

Organization: Centers for Disease Control and Prevention  
Web site: <http://www.cdc.gov/tobacco/>  
Current and historical state-level data on the prevalence of tobacco use, the health impact and costs associated with tobacco use, tobacco agriculture and manufacturing, and tobacco control laws in the United States.

Also reference the NATIONS database from web site:  
<http://apps.nccd.cdc.gov/nations>

Organization: United States Department of Agriculture (USDA), Economic Research Service (ERS)  
Web site: <http://www.econ.ag.gov/briefing/tobacco/>  
Data on cigarette sales, cigarette and tobacco leaves production.

Organization: U.S. Department of Health and Human Services

Web site: <http://www.dhhs.gov/> or

<http://www.dhhs.gov/topics/smoking.html>

The US Department of Health and Human Services provides information on US statutes restricting smoking in general categories of public places.

### **Industry Reports**

Use a search engine to identify and search the home pages of cigarette and other tobacco product producers.

A few examples:

<http://www.philipmorrisusa.com>

<http://www.rjrt.com/index.asp>

<http://www.bat.com>

<http://www.rothmansinc.ca/English/default.htm>

---

## **Helpful References for Aggregate-Level Data**

### **United States**

Tobacco Situation and Outlook Reports of the U.S. Department of Agriculture, Economic Research Service. Provides data on total U.S. expenditure on cigarettes, per capita cigarette consumption, U.S. cigarette production, exports, wholesale prices, and the market share of filter cigarettes.

The Tobacco Institute, various publications and years, provides data on total U.S. cigarette prices and sales.

U.S. Federal Trade Commission's reports to Congress pursuant to the Federal Cigarette Labeling and Advertising Act provides data on U.S. market share of low-tar cigarettes, U.S. tobacco advertising expenditure and nicotine delivery per cigarette (Barnett, Keeler and Hu, 1995; Harris, 1994)

Publications and reports by J.C. Maxwell in Business Week Magazine, Advertising Age, and Tobacco Reporter provide information on U.S. cigarette brand sales and the Herfindahl index of industry concentration for the tobacco sector.

The United States Department of Commerce, Bureau of the Census conducts annual surveys of industry groups and industries, and census reports on wholesale trade and retail trade, and publishes employment and wage data for U.S. tobacco manufacturers, wholesalers and retailers.

The United States Department of Labor, Bureau of Labor Statistics produces publications and provides information on US tobacco industry capital stock estimates.

The United States Department of Commerce, Bureau of the Census conducts annual surveys of industry groups and industries, and provides U.S. data on inventories and capital expenditures.

Information on U.S. laws that restrict smoking in public areas is available from state legislative records.

## United Kingdom

*National Income and Expenditure Yearbook* produced by the Central Statistical Office (CSO). Provides economic data for the United Kingdom including expenditure on tobacco and other consumer nondurables. *Monthly Digest of Statistics* is also published by the UK's CSO and provides population data.

*The Advertising Statistics Yearbook* of the UK Advertising Association and *Quarterly Digest of Advertising Expenditure*. Provides data on UK tobacco and other advertising expenditure.

## Australia

*National Accounts* published by the Australian Bureau of Statistics provide data on nominal expenditures on cigarette and tobacco products, an implicit price deflator for tobacco products, and measures of nominal household disposable income.

The Bureau of Statistics' *All Groups: Capital Cities Consumer Price Index* produces Australia's Consumer Price Index

*Commonwealth of Australia Year Book* produced by the Australian Government Publishing Service provides Australian population data

Commercial Economic Advisory Service of Australia publishes data on Australian advertising expenditure by cigarette companies.

## Academic Publications and the Development of Tobacco Control Measures

Wasserman et al. (1991) and Ohsfeldt, Boyle and Capilouto (1999). Provide updated indices of smoking regulations in the United States

The United States Surgeon-General's Report (1989) contains a four-step scale of restrictiveness of smoking laws.

Alternative measures of anti-smoking regulation (based on major local smoking ordinances published by Americans for Non-smokers' Rights and weighted by local population) in Sung, Hu and Keeler, 1994.

Hu, Sung and Keeler (1995) quantified total pages of cigarette advertising in *Life* magazine distributed in California as a representative sample of industry media presence in the state. They point out that comprehensive and systematic data on tobacco industry advertising and promotion activities are very difficult to

obtain, especially at sub-national level. They suggest therefore that the frequency of advertisements in newspapers and magazines may be the best proxy for the industry's countervailing behavior in response to tobacco control policies.

Gruber (2000) categorizes clean indoor air laws according to a Youth Access Index based on an index developed by the National Cancer Institute to evaluate state laws limiting youth access to cigarettes.

## **Subnational Level Data—The United States**

The Tobacco Institute provides state level cigarette sales, prices and tax rates for the United States (see Wasserman *et al.*, 1991; Sung, Hu and Keeler, 1994; Chaloupka and Grossman, 1996; Chaloupka and Pacula, 1999; Gruber, 2000)

The Tobacco Tax Council provides state level cigarette sales, prices and tax rates for the United States.

The California Board of Equalization reports monthly cigarette sales data for California.

The Population Research Unit of the California Department of Finance provides Californian metropolitan area population data.

The Bureau for Economic Analysis of the US Department of Commerce reports per capita income data for California as calculated from estimates of total personal income (See: Hu, Sung and Keeler, 1995).

## **Household Survey Data**

### ***United States***

U.S. Current Population Survey files contain economic and demographic data on large sample sizes of individuals and households. Provides data on cigarette prevalence and intensity, and the prevalence of smokeless tobacco use. The large size of this data source allows for age cohort analyses. State and metropolitan area identifiers allow for the analysis of applicable tobacco tax rates and restrictions. Note, proxy responses are often given for tobacco use, (particularly for teenagers), which tends to underestimate tobacco use even more substantially than the systematic under-reporting generally associated with surveys (Ohsfeldt, Boyle and Capilouto, 1999).

### ***United Kingdom***

The Tobacco Research Council can provide survey data on UK cigarette consumption by social class. This data has been inflated to agree with cigarette sales figures and to correct for the problem of survey under-reporting of tobacco use.

Income data for selected occupational groups can be obtained from the Family Expenditure Survey.

Smoking prevalence rates by sex, age and socio-economic group are available from the UK General Household Survey.

---

## References for Individual-Level Data

### Survey Data Sources

#### ***United States***

The National Health Interview Survey captures information on the percentage of US adults currently smoking cigarettes and US cigarette consumption. Survey results are provided by the Office of Smoking and Health at the US Centers for Disease Control (CDC).

The Monitoring the Future surveys conducted by the Institute for Social Research of the University of Michigan capture data on youth smoking prevalence. These data provide information on a variety of independent variables including: age, income, gender, ethnicity, marital status, parental educational level, family structure, mother's work status during respondent's childhood, presence of siblings, average weekly working hours, rural vs. urban location, and religious observation.

#### ***Australia***

The Australian Bureau of Statistics administer a survey which captures data on smoking propensity by different demographic groups in the state of New South Wales, Australia.

# VI. Conclusion

---

## Summary

### **Start-Point Data for Tobacco Control Research**

The underlying information needed for successful tobacco and tobacco control research includes measuring smoking prevalence and the conditional demand for cigarettes. Without this, the magnitude of the smoking epidemic and the economic significance of the tobacco market cannot be measured.

### **Fundamentals of Data Collection**

**The Basics of Collecting Data** Section of this Tool describes the basic elements of data collection, outlines the concept of aggregate data, identifies potential sources of such data, and defines a number of aggregate data variables necessary for the analyses explained further in the remainder of this Tool. Some of the many limitations or caveats associated with the use of aggregate data are also identified. This Section also outlines the concept of individual level data, defines a number of individual level variables, and provides a number of example survey questions that are frequently used to capture them. Similarly to the case of aggregate data, the limitations associated with using individual level data for econometric analysis are also highlighted.

Many developing countries around the world will be limited to aggregate level data sources, as the abundance of individual survey data is likely to be quite limited. When gathering aggregate data in your country, try to cross-validate the economic information that has been extracted from local sources with other, globally prominent sources of international aggregate data, such as The International Financial Statistics published by the International Monetary Fund.

## **Addressing the Technical Aspects of Data Preparation**

When choosing a software package, select the one that best accommodates the type of data you have gathered and the analyses that you plan to undertake. The **Data Preparation and Management: Easy Steps to Building Your Own Database** Section of this tool provides step-by-step instructions on how to manipulate data using SAS Software. The instructions include specifics on importing, reading and viewing raw data, cleaning and checking the quality the data and recoding the raw data, and merging one or more datasets together in order to maximize the usefulness of available price, policy, and general economic information from all levels of government and society. Packages similar to SAS, though quite comparable to the functions provided by SAS, are likely to be even more user-friendly and easy to use.

## **A Head Start on Potential Data Sources**

The final Section of tool is compromised of a series of lists that merely suggest various sources that may provide helpful leads or simple examples of data collection worldwide. Many of the sources also include addresses to informative websites or direct access to relevant data. Also, many of the websites listed provide additional links to similar data resources or data collection agencies. Begin your search by clicking away!

---

## **A Few Final Words**

The process of data collection may vary from country to country. Underlying differences are likely to exist in the prevalence of smoking use, the types of tobacco products available, approaches to tobacco marketing, the design of tobacco taxes and tariffs, cigarette pricing, tobacco employment and tobacco relevant public policy. The research techniques highlighted in this Tool should be adapted to the data collection process depending on the extent and array of available data.

Data collection in developing countries may be an especially difficult task to undertake due to limited extraction of such information from government sources in the past. Much of the information may not be regularly published in Central Statistical Agency Yearbooks. In some countries, government agencies may hesitate or require convincing before releasing key information to researchers. Although the need for intense data digging may turn into a difficult task for many, once completed, its significance will be large. The availability of such information for an increasing number of countries will enable all researchers to engage in new and, often, collaborative studies—many of which will improve both national and regional approaches and strategies to tobacco control policy making.

## VII. References

- Barnett, Keeler and Hu “Oligopoly Structure and the Incidence of Cigarette Excise Taxes.” *Journal of Public Economics* 57(3): 457–470, 1995
- Becker, G.S., Grossman, M., Murphy, K.M., “An Empirical Analysis of Cigarette Addiction.” *American Economic Review*, 84: 396–418, 1994.
- Chaloupka, F.J., “Rational Addictive Behavior and Cigarette Smoking.” *Journal of Political Economy*, 99, 722–742, 1991.
- Chaloupka, F.J., Grossman, M., “Price, Tobacco Control Policies, and Youth Smoking. Working Paper 5740, National Bureau of Economic Research, September 1996.
- Chaloupka, F.J., Wechsler, H., “Price, Tobacco Control Policies and Smoking among Young Adults. *Journal of Health Economics*, 16, 359–373, 1997.
- Chaloupka, F.J., Pacula, R.L., “An Examination of Gender and Race Differences in Youth Smoking Responsiveness to Price and Tobacco Control Policies.” *Tobacco Control*, 8, 373–377, 1999.
- Cragg, John G. “Some Statistical Models for Limited Dependent Variable with Application to the Demand for Durable Goods.” *Econometrica* 39:5, September 1971, 829–844.
- Delwiche Lora D. and Susan J. Slaughter, “The Little SAS Book: A Primer”. SAS Institute Inc., 1998.
- Gruber, J., Koszegi, B., “Is Addiction ‘Rational’? Theory and Evidence.” Working Paper 7507, National Bureau of Economic Research, April 2000.
- Harris, J.E., “A Working Model for Predicting the Consumption and Revenue Impacts of Large Increases in the U.S. Federal Cigarette Excise Tax.” Working Paper No. 4803 (National Bureau of Economic Research, Cambridge, MA.), 1994.
- Hu, Sung and Keeler, T.E., “Reducing Cigarette Consumption in California: Tobacco Taxes vs. An Antismoking Media Campaign,” *American Journal of Public Health* 85(9): 1218–1222, 1995.
- Lewit, E.M., Coate, D., Grossman, M., “The Effects of Government Regulation on Teenage Smoking.” *Journal of Law and Economics*, 24: 545–69, 1981.
- Lewit, E.M., Coate, D., “The Potential for Using Excise Taxes to Reduce Smoking.” *Journal of Health Economics*, 1: 121–145, 1982.
- Ohsfeldt, R.L., Boyle, R.G., Capilouto, E.I., “Tobacco Taxes, Smoking Restrictions, and Tobacco Use.” *The Economic Analysis of Substance Use and Abuse: An Integration of Econometric and Behavioral Economic Research*. The University of Chicago Press, September 1999.

- Sung, H., Hu, T., Keeler, T.E., “Cigarette Taxation and Demand: An Empirical Model.” *Contemporary Economic Policy*, 12 n3: 91–100, July 1994.
- Warner K.E., “Possible Increases in the Underreporting of Cigarette Consumption.” *Journal of the American Statistical Association* 73(362):314–318, 1978.
- Wasserman, J., Manning, W.G., Newhouse, J.P., Winkler, J.D., “The Effects of Excise Taxes and Regulations on Cigarette Smoking.” *Journal of Health Economics*, 10: 43–64, 1991.
- World Health Organization, “Guidelines for Controlling and Monitoring the Tobacco Epidemic.” Geneva, Switzerland, 1998.