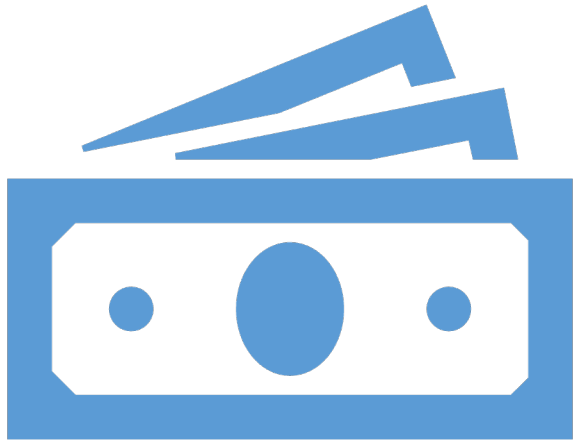
Succeeding in Academia – Finding and Collecting Data

COLLEGE of AGRICULTURE and APPLIED SCIENCES
UtahStateUniversity

Today

- Data in applied research
- Data source types and uses
- Primary data collection methods
  - Qualitative vs. quantitative
- Survey design and question formats
- Secondary data resources



Data in Applied Research

- Data is essential applied research, examples…
  - Prove a current or proposed theory
  - Illustrate or understand an issue
  - Compare policy over time
  - Demonstrate impacts
- Data availability may dictate the research undertaken or methods used
  - Understanding what data is available and in what forms, structures, time periods, etc. is critical
  - There are many data structures, types, and resources to consider

Data Source Types

- Secondary Data
  - Desk or background research
  - Data collected for another purpose
    - Company reports, government data, newspaper articles, etc.
  - Limited use, highly variable formats
- Primary Data
  - Data collected specifically for the research in question
    - Surveys, focus groups, tests….
  - More control, specific information

Secondary Data Sources

- External secondary data
  - Government stats
    - Expenditures, investments, consumption
  - Published reports
    - Trade associations, banks, government
  - Market research organizations
    - Consumer panels, retail audits
- Internal secondary data (business, industry, etc.)
  - Sales records
  - Prices or quotes
  - Promotions
  - Advertising strategies
  - Public relations

Primary Data - Qualitative

- Understand behavior
- Evaluate reactions
- Focuses on image, product usage, and associations with name
- Types
  - Focus groups
  - Individual in-depth interviews

Interviews

- Provide **qualitative** data by allowing respondents to expand on an answer
  - May provide more detail than surveys
  - Advantages: Obtain stories, memories and experiences that help explain behavior, data gathered through focus groups can be used to design full surveys
  - Disadvantages: Time consuming
- Observations consist of observing consumers and taking note of their behavior
  - Allow access to subtle visitor behavioral information
  - Example: Watching families interact as they choose a food booth at a festival

Primary Data - Quantitative

- Experiments
  - Sensory or taste testing
  - User panels
    - Trained panels
      - Trained individuals who can scrutinize
    - Un-trained panels
      - Sample of the population
      - Example of consumers in general
- Surveys
  - Most popular method
  - Sample a section of the population
    - Various survey methods

Quantitative Data

- Data, usually in numerical form, that define the characteristics or properties of a subject
- May be collected through surveys
  - Telephone surveys
    - Contact a random sample by phone from a telephone directory or number generator
  - In-person (face-to-face) surveys
    - Interviewing homeowners or in public areas
  - Internet surveys
    - Email or mail invitation to take the survey on-line, similar to mail survey
  - Dot surveys
  - Social media polls
    - Very short with 2-4 questions/options for current followers or targeted segment

Quantitative Data

- Choice of which survey type to use depends on several factors
  - Number of responses desired
  - Time frame to complete
  - Characteristics of sample
    - Access to phone or internet, language ability, literacy, etc.
  - Budget
    - Sample list charges, interviewer salaries, phone charges, copy fees, etc.

Telephone Surveys

- Include calling a random sample by phone to collect survey responses
- Advantages:
  - Immediate feedback
  - Caller can encourage person to take the survey
  - Target specific population
- Disadvantages:
  - Costly
  - Require access to large random sample of contact numbers
  - Laws against solicitation calls, use of unlisted cell numbers

Face-to-Face Surveys

- Conducted in-person
- Advantages:
  - Immediate feedback
  - Easier to obtain fully completed surveys
  - Targeted to very specific population
- Disadvantages:
  - Costly (interviewer payment)
  - Some individuals may be put off by approach or unwilling to reveal information about themselves
  - Requires permission at survey site

Internet Surveys

- Use online survey software and invite participants to respond
- Advantages:
  - Quick and immediately available
  - Complete responses
  - Very affordable (unless purchasing a sample list)
  - Easy to distribute invitations through email and social media
- Disadvantages:
  - Requires Internet access
  - Some people may be skeptical about providing sensitive information over the Internet
  - Purchasing email lists may be expensive
  - Motivating people to complete the survey may require incentives

Dot Surveys/Posters

- Focus on only a few questions
- Participants indicate response using colorful round stickers in the columns that represent their responses
- Advantages:
  - Higher response rates
- Disadvantages:
  - Respondents can see others' responses and may be swayed
  - Limited number of questions
  - No open-ended questions

Social Media Polls

- Twitter, Instagram, Facebook, etc…..
- Instagram polls use interactive stickers with two options that you can drag-and-drop on visual content
- Twitter has easy to create interactive, customized polls with four options
- Facebook offers two response fields/options, but you can include images and gifs
  - Consider trying more robust third-party apps like **Polls for Pages**

Survey Questionnaire Design

- Preliminary decisions
  - What information is required?
    - What problem or questions do we have?
  - Who are the target respondents?
    - Homeowners, ethnic population, farmers' market shoppers
  - What method of communication will be used to reach them?
    - Language issues, response issues, internet access…

Question Content

- Do the questions provide the needed information?
- Can the respondent answer the question correctly?
- Will the respondent answer the question correctly?
- Is there any bias in question, interviewer or external event?

Question Phrasing

- Do the words used mean the same to all the respondents?
- Are the words misleading or loaded?
- Does the question imply the choice of one of the alternatives?
- Can the question be asked as an open ended, multiple choice, or yes/no format?

Other

- Are the questions organized in a logical manner?
- Are the questions designed to avoid confusion & minimize recording error?
- Has a pre-test been conducted on a sample of respondents
- Has a focus group been conducted to formalize all potential answers

Question Formats

- Open-ended
- Multiple choice
- Yes/no
- Likert rating scales
- Semantic differential scales

Open Ended

- Identified range of answers that may be used later
- Best for focus groups or individual interviews
- Not very reliable for surveys due to variety & coding issues
- Example: What do you feel are the biggest issues with eating habits today?

Multiple Choice

- Provides a range of potential answers
  - Answer choices must be based on research and/or focus groups
- The respondent chooses the one that is most applicable

1. What is your primary motive for attending the farmers' market? (Check only one)

☐ Purchase produce
☐ Purchase packaged foods
☐ Purchase arts/crafts
☐ Social interaction
☐ Events/activities
☐ Concerts/music
☐ Purchase ready-to-eat food (food vendors)

Dichotomous Choice (yes/no)

- The questions only has a yes/no or true/false answer
- Limit of choices cuts down on survey design & respondent uncertainty

1. Are you the primary food purchaser for your household?

☐ Yes ☐ No

2. Have you attended a farmers' market prior to today?

☐ Yes
☐ No

Likert Rating Scales

- The respondent indicates his/her level of agreement
- A 5 or 7 point scale is most common
- Use in attitudinal questions

28. Please specify if you agree or disagree with each of the following statements.

Statement	Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree
I am concerned about the safety of my food	1	2	3	4	5
I have little time to prepare meals	1	2	3	4	5
I am concerned about my health/diet	1	2	3	4	5
I buy products with low environmental impact	1	2	3	4	5
I eat out frequently	1	2	3	4	5
Physical activity is an important part of my routine	1	2	3	4	5
Eating out is an event in my family	1	2	3	4	5
Supporting local farmers is important to me	1	2	3	4	5
Agricultural open space is important to me	1	2	3	4	5
I am concerned about the origin of my food	1	2	3	4	5
I am a vegetarian or vegan	1	2	3	4	5

Semantic Differential Scales

- Used to rate product attributes or range of attitudes
- 5, 7, and 9 point scales used

14. How important are the following farmers' market attributes/features?

Farmers' Market Attributes	Not important	Slightly important	Somewhat important	Very important	Extremely important
Concerts/Music	1	2	3	4	5
Free parking	1	2	3	4	5
Hours of operation	1	2	3	4	5
Convenient location	1	2	3	4	5
Number of vendors	1	2	3	4	5
Child/Family activities	1	2	3	4	5
Cultural events	1	2	3	4	5
Educational events	1	2	3	4	5
Certified farmers' market	1	2	3	4	5
Product variety	1	2	3	4	5
Food/beverage vendors	1	2	3	4	5

Secondary Data Resources

- USDA ERS: Data on consumption and production of agricultural goods, international markets and trade, consumer food choice and health, rural economic factors, etc.: https://www.ers.usda.gov/
- USDA AMS: Detailed data/information on marketing conditions for hundreds of agricultural commodities at major domestic and international wholesale markets, production areas, and ports of entry: https://www.ams.usda.gov/market-news
- US Bureau of Labor Stats: Detailed data on inflation, salaries, labor force, cost of living, economic trends, etc.: https://www.bls.gov
- US Census: Data collected in US residents and households, socio demographics by area. https://www.census.gov
- FAOSTAT: Aggregate data on various dimensions of food and fiber production, food security, and trade (among others):https://www.fao.org/faostat/en/#home

Secondary Data Resources

- USDA NASS “Quick Stats”: Aggregate data from the USDA/NASS universe (e.g. US Ag Census, ARMS, etc.): https://www.nass.usda.gov/Quick_Stats/
- USDA Surveys (of various kinds, especially USDA-ARMS): the main location to look up data from the universe of USDA-related surveys, farm-level and other. For students, it might be useful to note the distinction between the Ag. Census (every 5 years, mandatory for all farmers, limited scope) versus the USDA-ARMS (repeated cross-section, focused on specific crops each year, smaller samples): https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Ag_Resource_Management/
- Living Standards Measurement Study – Integrated Surveys on Agriculture data (LSMS-ISA data): Panel data of households in 9 African countries, often with has GPS data for location of households: https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA
- Google Earth Engine (GEE) data catalog: A universe of geo-spatial and remote sensing data for various types of research: https://developers.google.com/earth-engine/datasets
- World Bank country databases: Data on aggregate country-level statistics, UN sustainable development goals, etc.: https://data.worldbank.org/country
- Rural Household Multi-Indicator Survey (RHoMIS): another resource for household level, LSMS-style surveys around the world: https://www.rhomis.org

Next

- June 19: Managing the Tenure Process
  - Last webinar

Questions?

Common Data Structures

- Cross-sectional data
- Time series data
- Pooled cross-section time-series data
- Panel or longitudinal data

Cross-Sectional Data

- A cross-sectional data set consists of a sample of observations on individual economic agents or other units taken at a single point or periods in time
- Common units of observation in economic cross-sectional data sets include:
  - Individual economic agents such as:
    - Individual persons, households or families
    - Individual firms
  - Geographical units such as:
    - Countries, cities, metropolitan areas or urban areas, provinces, states, or regions
  - Economic units such as:
    - Occupations -- groups of individual workers categorized by the type of work they do or the nature of the skills they require to perform their jobs
    - Industries -- groups of firms categorized by the types of product outputs they produce or sell

Cross-Sectional Data Cont.

- A cross-sectional data set has individual observations with no natural ordering
  - The observations in a cross-sectional data set can be ordered or sorted in any way without altering or corrupting the nature of the sample information
- Cross-sectional data sets are constructed by a random sampling from an underlying population
  - Random sampling means that the observations can be assumed to be statistically independent
  - Random sampling is sufficient to satisfy the statistical assumption of zero error covariances, or non-autoregressive errors, in regression models

Cross-Sectional Data Example

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
SFO	1/1/2015	70	35%	21 mph
Boston	1/1/2015	34	39%	16 mph
Chicago	1/1/2015	29	15%	54 mph

Time Series Data

- A time-series data set consists of a sample of observations on one or more variables over successive periods or intervals of time
- An important characteristic of economic time-series data is the frequency at which the observations are collected, reported or analyzed
- The most common data frequencies for economic time-series are:
  - Daily, weekly, monthly, quarterly, or annual

Time Series Data Cont.

- Time-series data set observations have a natural ordering -- specifically a chronological ordering
  - Time is an inherently important variable in time-series data sets
  - The chronological ordering of time-series observations conveys important sample information because past events can influence future events and because lags in behavior are very common in economics due to habit persistence and adjustment costs
  - Time-series observations cannot be re-ordered in a non-chronological way without corrupting the nature of the sample information
- Time-series data sets are generally not generated by random sampling
  - Time series observations are not usually statistically independent
  - The observations in an economic time-series data set almost always violate the statistical assumption of zero error covariances, or non-autoregressive errors, in regression models
  - Econometric analysis requires a variety of special statistical procedures to account for the statistical dependence of time-series observations
    - Economic time series exhibit time trends and seasonal variations

Time Series Data Example

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2012	35	56%	3 mph
NYC	1/1/2013	47	65%	21 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2015	55	45%	4 mph

Pooled Cross-Section Time-Series Data

- A pooled cross-section time-series data set consists of two or more different samples of cross-sectional observations from the same population taken at two or more points in time
  - Observations in a pooled cross-section time-series data set have both an individual identifier and a time or period identifier
  - The cross-sectional units of observation may be individual economic agents (e.g., individual persons, households, or firms), economic aggregates (e.g., industries and occupations), or geographic aggregates (e.g., urban areas, provinces/states and countries)

Pooled Cross-Section Time-Series Data Cont.

- Pooled cross-section time-series data sets have cross-sectional samples for different periods of time that contain different cross-sectional observations -- different persons, different households, different firms
  - Observations in a pooled cross-section time-series data set can be assumed to be statistically independent across time
- Advantages of pooled cross-section time-series data sets:
  - Provide larger sample sizes
  - Permit investigation of whether economic relationships have changed over time

Pooled Cross-Section Time-Series Data Example

Pooled Cross Sections: Two Years of Housing Prices

	obsno	year	hprice	proptax	sqrft	bdrms	bthrms
Year 1993	1	1993	85500	42	1600	3	2.0
	2	1993	67300	36	1440	3	2.5
	3	1993	134000	38	2000	4	2.5
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	250	1993	243600	41	2600	4	3.0
Year 1995	251	1995	65000	16	1250	2	1.0
	252	1995	182400	20	2200	4	2.0
	253	1995	97500	15	1540	3	2.0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	520	1995	57200	16	1100	2	1.5

Panel or Longitudinal Data

- A panel or longitudinal data set consists of two or more sets of observations on the same sample of cross-sectional members at two or more points in time
  - A panel data set consists of repeated observations over time on the same set of cross- sectional units
  - A panel data set therefore provides time series observations for each cross-sectional member in the data set
    - It follows the same cross-sectional units over time
  - The cross-sectional units of observation may be either individual economic agents (such as individual persons, households, or firms), geographical units (such as cities or provinces), or other entities (such as occupations or industries)

Panel Data Cont.

- A panel data set is that it provides observations at different points in time on the same sample of cross-sectional units
  - The variables in a panel data set have both an individual identifier and a time or period identifier
  - There are three different types of variables in a panel data set.
    - Individual- and time-varying variables that vary both over cross-sectional units and over time
    - Individual-constant time-varying variables that vary over time but take the same value for all cross-sectional units of observation in any one time period
    - Individual-varying time-constant variables that vary over cross-sectional units but take the same value in all time periods for any one cross-sectional unit
- Advantages of panel data sets:
  - Having multiple observations on the same cross-sectional units allows us to control for certain unobserved characteristics of individuals, firms, provinces, etc.
    - For example, a panel data set on the earnings, education and other variables of individual workers allows us to control for unobserved differences among individuals in innate ability
  - Panel data sets allow us to investigate lags in the behavior of individual economic units
    - For example, a panel data set on the earnings, education and other variables of individual workers allows us to investigate the extent to which earnings in one year depend upon earnings in previous years

Panel Data Example

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2013	47	65%	21 mph
SFO	1/1/2015	70	35%	21 mph
SFO	1/1/2014	75	23%	2 mph
SFO	1/1/2013	71	39%	13 mph
Boston	1/1/2015	34	39%	16 mph
Boston	1/1/2014	26	17%	27 mph
Boston	1/1/2013	45	46%	18 mph