Research design:
- all procedures for carrying out a study: design, population/samples, data, analysis
- by identifying major constructs, variables, population research clarifies the focus and logic of the problem
- guides study—the better and more well-designed study, fewer problems in carrying it out

Elements of research in general
1. Identify problem
2. Review literature
3. Specify Theory and hypotheses
4. Select design and rationale
5. Identify population, select sample/case
6. Decide what data to collect and how to collect it
7. Describe procedures for analyzing data
8. Discuss problems/limitations of the design
9. Collect data
10. Analyze data
11. Interpret findings and state conclusions
12. Draw implications
13. Suggest solutions

Units of analysis—what or who to study—populations
General issues
- entities under study—what is to analyzed or described
- examples of units of analysis
  - people
  - organizations
  - institutions
  - states

Why Unit of Analysis Matters
- When done right, prevents us from making inappropriate inferences from “wrong units”—using aggregate data to make inferences about individuals (ecological fallacy)
- tells us what theory to use, what data to collect, types of analysis
GENERAL ISSUES, CONT’D

- Multi-level: multiple units of analysis
  - qualitative and/or quantitative

- Samples or cases
  - subset of population
  - often wish to make inferences or draw conclusions based on a small sample of cases
  - can do this in a number of ways

Constructs, variables, and observations

- Constructs: generalized notions or concepts often derived from theory, not directly measurable, observable

- Variables: expression of a construct in measurable form

- Types of variables—measurement:
  - qualitative or categorical
    - e.g., gender, neighborhoods, types of schools
  - quantitative or numerical variables
    - e.g., test scores, voter turnout, income

- Observations: actual instruments used to measure variable
  - e.g., survey data

VARIABLES—WHAT TO STUDY

- How to decide:
  A. should be based on existing theory/research
  B. make use of deductive logic: coming from theory

TYPES OF VARIABLES

- dependent variables
  - They are the thing you want to explain: they depend upon...
    - Example: Test Scores

- independent variables
  - They are the things that affect the thing you want to explain
    - Example: buying the textbook, going to class

- intervening variables
  - Things that occur between the independent and dependent variables. Often caused by the independent variable and a cause of the dependent variable.
    - Example: reading the textbook, taking notes in class

RELATIONSHIP AMONG VARIABLES

1. Types of relationships
   a. Between categorical variables: strength of relationship
   b. Between quantitative variables: direction and linearity

2. Nature of causal relationships
   a. often interested in determining causal relationships between dependent and independent variables in order to draw implications for policy or practice
      (e.g., Voter Registration Efforts → Higher Voter Turnout; promote GOTV)
   b. Almost impossible to establish direct causality ("direct effect") in social science research, because we don’t experiment.

CAN WE ESTABLISH CAUSALITY DEFINITELY ANY OTHER WAY?

- NO

- Accumulate evidence that strongly suggests causality
- Show association (correlation)
- Establish direction of influence
  - one way: temporal relationships between independent and dependent variables
  - not always possible
  - theory can suggest direction
  - can test hypothesized direction empirically

- Eliminate rival hypotheses (spurious relationships)
  - Always control for other factors that might influence or explain relationship
  - Thus the importance of theory

NEXT BEST THING
Characteristics of survey research designs:
- Participants selected from larger population through systematic sampling procedures
- Standardized procedures for collecting data through interviews or questionnaires
- Data transformed into quantitative measures and analyzed through statistical analysis

Strengths
- Can generalize to a very large population with relatively small sample
- Cost-efficient
- Suitable for a variety of research questions from descriptive to explanatory
- Can be analyzed with sophisticated statistical models
- Questionnaire data can be augmented with other types of data, such as tests and archival data

Weaknesses
- Can't provide the depth of information available from qualitative methods
- Can't establish or prove causality as strongly as experimental methods
- Procedures can be complex and costly to carry out, especially for large-scale studies

Types of survey designs
- Cross-sectional
  - Most simple, least costly type
  - Good for simple analysis of population—e.g., public opinion polls
  - Limited for inferences over time
- Longitudinal
  - More useful for looking at developmental, longitudinal issues
  - Helps untangle life course effects (e.g., aging) vs. cohort effects (although not temporal effects)
  - Types of cohorts: age cohorts, grade cohorts, graduate cohorts
  - Trend studies: repeated cohort studies; e.g., repeated age groups
  - Panel studies: repeated measures of same cohort
- Hierarchical
  - Multiple samples units
  - Research focus on multiple units
  - Can be longitudinal as well

Methods
- Face-to-face interviews
- Telephone interviews
- Group-administered questionnaire
- Mail questionnaire
- Web-based questionnaire

Factors for consideration
- Population size
- Sampling procedures
- Types of questions: structured vs. unstructured (purpose of research)
- Sensitivity of information: self-administered maybe better; telephone may give more biased information
- Response rate
- Timeliness
- Costs
- Staff and facilities required (personnel, training, equipment)

The population includes all objects of interest whereas the sample is only a portion of the population.
- Parameters are associated with populations and statistics with samples.
- We compute statistics, and use them to estimate parameters.
- There are several reasons why we don't work with populations.
  - They are usually large
  - It is often impossible to get data for every object we're studying.
- Sampling comes at a cost, and the more items surveyed, the larger the cost.

Sources of error in the data collection process
- Coverage due to not sampling parts of the population
- Sampling due to variability in sample vs. population (variance) and due to systematic differences in sample vs. population (bias)
- Nonresponse
  - Unit (don't participate in survey) and
  - Item (don't answer certain questions)
- Measurement error not providing accurate or truthful information

Population vs sample
- The population includes all objects of interest whereas the sample is only a portion of the population.
- Parameters are associated with populations and statistics with samples.
- We compute statistics, and use them to estimate parameters.
- There are several reasons why we don't work with populations.
  - They are usually large
  - It is often impossible to get data for every object we're studying.
- Sampling comes at a cost, and the more items surveyed, the larger the cost.
**Population vs. Sample**

- **Population**
  - The term “population” is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study.
  - Examples: TAMU-T students, trees in North America, automobiles with four wheels.

- **Sample**
  - The term “sample” refers to a portion of the population that is representative of the population from which it was selected.
  - Examples assuming the populations to the left:
    - 47 TAMU-T students chosen randomly
    - 8463 trees randomly selected in North America;
    - 20 sample autos from each make (e.g., GM, Ford, Toyota, Honda, etc.).

**Types of Sampling**

- Random sampling is analogous to putting everyone’s name into a hat and drawing out several names. Each element in the population has an equal chance of occurring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling.

- Systematic sampling is easier to do than random sampling. Here the list of elements is “counted off.” That is, every nth element is taken. This is similar to lining everyone up and numbering off “1, 2, 3, 4, 1, 2, 3, 4, etc.” When done numbering, all people numbered 4 should be used.

- Convenience sampling: readily available data is used. That is, the first people the surveyor runs into. Very easy to do, but it’s probably the worst technique to use.

- Cluster sampling is accomplished by dividing the population into groups — usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

- Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into male and female. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

**Discrete vs. Continuous Variables**

- Discrete variables are usually obtained by counting. There are a finite or countable number of choices available with discrete data. You can’t have 2.63 people in the room.

- Continuous variables are usually obtained by measuring. Length, weight, and time are all examples of continuous variables. Since continuous variables are real numbers, we usually round them. This implies a boundary depending on the number of decimal places. For example: 64 is really anything 63.5 <= x < 64.5. Likewise, if there are two decimal places, then 64.03 is really anything 63.025 <= x < 63.035. Boundaries always have one more decimal place than the data and end in a 5.

**Levels of Measurement**

- There are four levels of measurement: Nominal, Ordinal, Interval, and Ratio. These go from lowest level to highest level. Data is classified according to the highest level which it fits. Each additional level adds something the previous level didn’t have.
  - Nominal is the lowest level. Only names are meaningful here.
  - Ordinal adds an order to the names.
  - Interval adds meaningful differences
  - Ratio adds a zero so that ratios are meaningful.

  You might also think of these as ‘types of variables.’

**Measures of Central Tendency**

- The term "Average" is vague.

- Average could mean one of four things. The arithmetic mean, the median, midrange, or mode. For this reason, it is better to specify which average you’re talking about.
This is what people usually intend when they say "average"

- Population Mean: \( \mu = \frac{\sum x}{N} \)
- Sample Mean: \( \bar{x} = \frac{\sum x}{n} \)
- Frequency Distribution: \( \bar{x} = \frac{\sum x f}{\sum f} \)

The mean of a frequency distribution is also the weighted mean.

The data must be ranked (sorted in ascending order) first. The median is the number in the middle.

- To find the depth of the median, there are several formulas that could be used, the one that we will use is: Depth of median = 0.5 \( * \) (n + 1)

- Raw Data

The median is the number in the "depth of the median" position. If the sample size is even, the depth of the median will be a decimal -- you need to find the midpoint between the numbers on either side of the depth of the median.

- Ungrouped Frequency Distribution

Find the cumulative frequencies for the data. The first value with a cumulative frequency greater than depth of the median is the median. If the depth of the median is exactly 0.5 more than the cumulative frequency of the previous class, then the median is the midpoint between the two classes.

- Grouped Frequency Distribution

This is the tough one. Since the data is grouped, you have lost all original information. Some textbooks have you simply take the midpoint of the class. This is an over-simplification which isn’t the true value (but much easier to do). The correct process is to interpolate.

Find out what proportion of the distance into the median class the median by dividing the sample size by 2, subtracting the cumulative frequency of the previous class, and then dividing all that by the frequency of the median class.

Multiply this proportion by the class width and add it to the lower boundary of the median class.

The mode is the most frequent data value. There may be no mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

For grouped frequency distributions, the modal class is the class with the largest frequency.

The midrange is simply the midpoint between the highest and lowest values.

The Mean is used in computing other statistics (such as the variance) and does not exist for open ended grouped frequency distributions (1). It is often not appropriate for skewed distributions such as salary information.

The Median is the center number and is good for skewed distributions because it is resistant to change.

The Mode is used to describe the most typical case. The mode can be used with nominal data whereas the others can’t. The mode may or may not exist and there may be more than one value for the mode (2).

The Midrange is not used very often. It is a very rough estimate of the average and is greatly affected by extreme values (even more so than the mean).

<table>
<thead>
<tr>
<th>Property</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Midrange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always Exists</td>
<td>No (1)</td>
<td>Yes</td>
<td>No (2)</td>
<td>Yes</td>
</tr>
<tr>
<td>Uses all data values</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Affected by extreme values</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
MEASURES OF VARIATION

Because it matters why there is a difference

RANGE
- The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value.
- Since the range only uses the largest and smallest values, it is greatly affected by extreme values; that is, it is not resistant to change.

RANGE = MAXIMUM - MINIMUM

VARIANCE

*Average Deviation*
- The range only involves the smallest and largest numbers, and it would be desirable to have a statistic which involved all of the data values.
- The first attempt one might make at this is something they might call the average deviation from the mean and define it as:
  
  $\text{Ave. Dev} = \frac{\sum (x - \mu)}{N}$

- The problem is that this summation is always zero. So, the average deviation will always be zero. That is why the average deviation is never used.

VARIANCE

Population Variance
- So, to keep it from being zero, the deviation from the mean is squared and called the "squared deviation from the mean." This "average squared deviation from the mean" is called the variance.

$$\text{Population Variance} = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Unbiased Estimate of the Population Variance
- One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$\text{Sample Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

STANDARD DEVIATION
- There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$\text{Population Standard Deviation} = \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$
$$\text{Sample Standard Deviation} = s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

- The sample standard deviation is not the unbiased estimator for the population standard deviation.
The proportion of the values that fall within $k$ standard deviations of the mean will be at least $1 - \frac{1}{k^2}$, where $k$ is a number greater than 1.

"Within $k$ standard deviations" interprets as the interval: $\overline{x} - ks$ to $\overline{x} + ks$

Chebyshev’s Theorem is true for any sample set, not matter what the distribution.

The empirical rule is only valid for bell-shaped (normal) distributions. The following statements are true.

- ~68% of the data values fall within one standard deviation of the mean.
- ~95% of the data values fall within two standard deviations of the mean.
- ~99.7% of the data values fall within three standard deviations of the mean.

A sample space is the set of all possible outcomes. However, some sample spaces are better than others.

Consider the experiment of flipping two coins. It is possible to get 0 heads, 1 head, or 2 heads. Thus, the sample space could be {0, 1, 2}. Another way to look at it is flip { HH, HT, TH, TT }. The second way is better because each event is as equally likely to occur as any other.

When writing the sample space, it is highly desirable to have events which are equally likely.

Another example is rolling two dice. The sums are { 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 }. However, each of these aren’t equally likely. The only way to get a sum 2 is to roll a 1 on both dice, but you can get a sum of 4 by rolling a 1-3, 2-2, or 3-1. The following table illustrates a better sample space for the sum obtain when rolling two dice.

<table>
<thead>
<tr>
<th>First Die</th>
<th>Second Die</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1/36</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2/36</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3/36</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4/36</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5/36</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6/36</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>5/36</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>4/36</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>3/36</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>2/36</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1/36</td>
</tr>
</tbody>
</table>

The above table lends itself to describing data another way -- using a probability distribution. Let’s consider the frequency distribution for the above sums.
If just the first and last columns were written, we would have a probability distribution. The relative frequency of a frequency distribution is the probability of the event occurring. This is only true, however, if the events are equally likely. This gives us the formula for classical probability. The probability of an event occurring is the number in the event divided by the number in the sample space. Again, this is only true when the events are equally likely. A classical probability is the relative frequency of each event in the sample space when each event is equally likely.

\[ P(E) = \frac{n(E)}{n(S)} \]

Empirical probability is based on observation. The empirical probability of an event is the relative frequency of a frequency distribution based upon observation.

\[ P(E) = \frac{f}{n} \]

A commonly used measurement device used in political science research is the Likert scale. The Likert scale, developed in 1932 by Rensis Likert, is an important tool for measuring attitudes in public opinion research. Although it can be a true scale, meaning it is the sum of several individual items (or what is often referred to as an index variable). It is most often used as a single variable or item measuring the degree of agreement with a proposed statement. Another important definitional standard of the Likert scale, or Likert items, is the presence of approval and disapproval values with an obvious middle or undecided point. An example will help illustrate what this means:

Instructions: Please note your agreement or disagreement with the following statement:

**Item:** "Ronald Reagan was the greatest American president."

**Response Options:**
1 = Strongly Agree
2 = Agree
3 = Neither agree nor disagree
4 = Disagree
5 = Strongly Disagree

Note that there are typically an odd number of response options (5 or 7 point scale) and an easily identified middle point. These are the basic characteristics of this important tool for measuring public attitudes on just about any question or subject.

There are two rules which are very important.

- All probabilities are between 0 and 1 inclusive
- The sum of all the probabilities in the sample space is 1

There are some other rules which are also important.

- The probability of an event which cannot occur is 0. The probability of any event which is not in the sample space is zero.
- The probability of an event which must occur is 1. The probability of the sample space is 1.
- The probability of an event not occurring is one minus the probability of it occurring. \[ P(E') = 1 - P(E) \]
A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in a class. It is a way of showing unorganized data e.g. to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc. Some of the graphs that can be used with frequency distributions are histograms, line graphs, bar charts and pie charts.

**Frequency Distribution Charts**
- Have many different shapes. We are concerned mainly with some of the more common configurations here.
  - Unimodal
  - Bimodal
  - Multimodal

**Unimodal Distributions**
- Have one peak or "mode." The one you are most familiar with is the normal distribution.
  - The normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function or informally as the bell curve.

**Unimodal - Asymmetric**
- Asymmetric distributions are positively skewed or negatively skewed. A positively skewed distribution is one in which the right (positive) tail of the distribution is the long one. A negatively skewed distribution is one in which the left (negative) tail of the distribution is the long one.

The distribution of personal income is an example of positive skewness. Income as a variable of education level is typically negatively skewed.
BIMODALITY

Bimodal distributions have two modes, as seen below. Examples include public opinion on same-sex marriage, or abortion on demand.

MULTIMODAL

Frequency distributions with three or more local maxima are called multimodal distributions. Examples include local political preferences, and of course Proportional Representation preference models.

WHAT IS PUBLIC OPINION?

- An aggregation of individual opinions
- Reflection of the majority
  - produces conformity on controversial issues
- Pluralism, the clash of organized interests
  - intensity of opinion
- Media and elite opinion
- A myth, it is rhetoric
  - If a myth or elite driven, despotism may result.

WHICH WAY IS BEST?

- Depends on availability of data
- Type of government
- Available technology

THE GREEKS & ROMANS

- Plato
  - envisioned an ideal citizen and public opinion as the basis of law
  - Believed that the masses were generally uneducated
  - Thus, the best ruler would be a philosopher-king
- Aristotle
  - held a more positive view of the people
  - Saw public opinion as the norms of society
- Romans
  - Saw public opinion as an affirmation of leaders (retrospective voting)
NICCOLÒ MACHIAVELLI
- Agreed with Plato that people are “thankless, fickle, false, studious to avoid danger, greedy of gain...and in your hour of need they turn against you.”
- Also believed in a benevolent dictator
- Began articulating the idea of Pluralism

HOBBES & LOCKE
- Hobbes
  - People live in a state of competition with one another
  - People form a “social contract” with the state, The Leviathan, to protect them from each other
- Locke
  - Rediscovered Aristotle’s more positive view of the people
  - Articulated the idea of inalienable rights that are protected by the state

JEAN-JACQUES ROUSSEAU
- The Social Contract – people are generally honest and want government to serve the general will of the community.
- Jacques Necker – first coined the word “public opinion” to describe the emerging public political discourse among the French bourgeoisie.

JÜRGEN HABERMAS
- Defined the “public sphere” as conversations that occur between people in public.
- At the time of the 19th century, he did not consider women to be a part of public opinion because they were a small part of the public sphere.

SOCIAL CHOICE
- Jeremy Bentham and other “utilitarians” became interested in maximizing happiness (social welfare).
- John Stuart Mill and others became interested in what sort of electoral system best worked for democracy, and rule of the majority

ALEXIS DE TOCQUEVILLE
- Democracy in America – further observed rule of the majority and how it constrained public opinion in America.
- Equality breeds engagement (similar to a reflection of the majority).
- Karl Marx – elites wish to perpetuate inequality so that those on the bottom become passive.
THE MODERN ERA

- James Bryce – observed the emerging role of the media in shaping public opinion.
- Gabriel Tarde – argued that public opinion flowed solely from the media.
- Charles Tilly – technological innovation moved the notion of public sphere from the community to the nation.

MODES OF TRANSMISSION

- Rhetoric
- Demonstrations and Riots
- Printing Press
- Public forums (coffee houses, salons)
- Elections
- Polling
- Blogging? Facebook? Twitter?

PUBLIC: A GROUP THAT HAS SOMETHING IN COMMON

- Types of publics:
  - Everyone
  - People connected to their government
    - Citizens
    - Citizens of voting age
    - People registered to vote
    - People likely to vote
  - Attentive publics
  - Issue publics

THE BIRTH OF POLLING: THE STRAW POLL

- The “straw poll”: first conducted by the Harrisburg Pennsylvanian in 1824.
- Mail out ballots and tally returned votes.
- Also used as a marketing ploy

RELIABILITY OF STRAW POLLS

- Depend on people to return mail-in cards.
  - Pierre du Pont straw poll concerning Prohibition was only returned by people who favored repealing it.
  - People polled can be unrepresentative (haphazard sample)
  - 1936 Literary Digest poll predicted Alf Landon (57%) would be elected president over FDR (43%).

THE BIRTH OF THE MODERN POLL

- Gallup in 1936 predicts FDR wins (55.7% even though FDR won 60.8%).
- Used scientific “quota sampling” of only about 1,200 people compared to the 2 million in the Literary Digest straw poll.
SAMPLING

The Sample

Unknown

The Population

SIMPLE RANDOM SAMPLING ERROR

- The Sample will “likely” look like the Population
- But, by random chance it is unlikely that the Sample will be exactly like the Population

SIMPLE RANDOM SAMPLING ERROR

\[
\text{Standard Error} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}
\]

Where, 
- \( \hat{p} \) is the observed percentage 
- \( N \) is the number of people in the sample

SIMPLE RANDOM SAMPLING ERROR

- 95% of the time, the true Population mean will be within +/- the Sampling Error
- 5% of the time, it won’t
- Sampling Error is dependent on the size of the Sample (but not the size of the Population!)
- Sample Size = Sampling Error
  - \( 600 = 4\% \)
  - \( 1,067 = 3\% \)
  - \( 2,430 = 2\% \)

SIMPLE RANDOM SAMPLING ERROR

- 95% of the time, the true population value is +/- 1.96 S.E.

SIMPLE RANDOM SAMPLING ERROR

- Quota Samples – use the census to find a certain number of people in different groups to force sample to be representative of population
  - The method failed in 1948:
**TYPES OF SAMPLING**

- Simple Random Sampling – everyone has an equal chance of being selected
- Multistage Cluster Sampling – a combination of the two approaches
- Telephone surveys & Random Digit Dialing

**THE PROBLEM OF NON-RESPONSE**

- 56% of people contacted responded to the 2000 National Election Survey.
- 5% of households don’t have phones
- Solution is to weight the surveys to match the census, but...
  - Census is not entirely accurate
  - People who choose not to respond may hold different opinions than those that do, even within the same demographic category

**CELL PHONE COVERAGE**

Cell Phone Only Population
National Health Interview Survey

**CELL PHONE COVERAGE**

- 56% of people contacted responded to the 2000 National Election Survey.
- 5% of households don’t have phones
- Solution is to weight the surveys to match the census, but...
  - Census is not entirely accurate
  - People who choose not to respond may hold different opinions than those that do, even within the same demographic category
OTHER TYPES OF "POLLS"

- Automated Dialing Polls
  - Who answers the phone?
- Internet Polls
  - Are they representative?
- Push Polls
  - Campaigns masquerading as pollsters
- Letter writing campaigns

BIAS

- Bias refers to anything that causes the estimate from the survey to differ from the true population
  - Sampling: how representative the poll is
    - Sampling error
    - Non-response error
  - Survey design:
    - Question wording
    - Item ordering
    - Question ordering
  - Interviewer and response:
    - Social desirability

QUESTION WORDING: MULTIPLE STIMULI

"Suppose an admitted Communist wanted to make a speech in your community. Should he be allowed to speak or not?

- Combines attitudes towards communism with attitudes towards free speech

QUESTION WORDING


"Do you think legal immigrants mostly help the economy by providing low cost labor, or mostly hurt the economy by driving wages down for many Americans?" Options rotated. N=1023. MoE 5

- Mostly Help 42%
- Mostly Hurt 52%
- Neither (vol.) 3%
- Both (vol.) 2%
- Unsure 1%


"Overall, do you think illegal immigrants hurt or help the U.S. economy?" Options rotated

- Help 26%
- Hurt 64%
- Depends (vol.) 5%
- Unsure 5%

ITEM ORDERING: TAKE THE FIRST CHOICE


"...If the congressional election were held today, would you vote for the Republican candidate in your district or the Democratic candidate in your district?" If undecided: "Well, if you had to vote, which way would you lean?"

Rep: 39%   Dem: 36%   Other/Not Sure: 25%


"If the elections for Congress were being held today, which party's candidate would you vote for in your congressional district? (rotate) the Democratic Party's candidate or the Republican Party's candidate?" If undecided: "As of today, do you lean more toward (rotate) the Democratic Party's candidate or the Republican Party's candidate?"

Rep: 42%   Dem: 50%   Other (vol.)/Undecided: 8%

FRAMING

1980

"The U.S. should let Communist newspaper reporters from other countries come here and send back to their papers the news as they see it"

55% “Yes”

When preceded by a question about U.S. reporters sent to Communist countries

75% “Yes”

Problem: We often don’t know all of the questions asked and in what order
CHECK THE HORSE’S MOUTH

<table>
<thead>
<tr>
<th>Hamilton Beattie &amp; Staff for the Democratic Senatorial Campaign Committee. Sept. 4-8, 2003, N=500 likely voters statewide. MoE 4.4:</th>
<th>Brad Carson (D)</th>
<th>Kirk Humphreys (R)</th>
<th>Und.</th>
</tr>
</thead>
<tbody>
<tr>
<td>45%</td>
<td>42%</td>
<td>13%</td>
<td></td>
</tr>
</tbody>
</table>

| Tulsa World/KOTV Oklahoma Poll conducted by Consumer Logic. July 8-12, 2004, Statewide: | 42% | 39% | 19% |


OTHER PROBLEMS WITH POLLS

- Do they capture true feelings on sensitive issues such as race?
- Interviewer bias
- Multiple stimuli versus balanced arguments
- Non-attitudes and response acquiesces
- The surprise poll draws attention, but is it representative of the population?

CRITICS OF MODERN POLLING

- Blumer: general public vs. issue publics
- Rogers: can polls really measure what they claim they measure?
  - “Instead of feeling the pulse of American democracy, Dr. Gallup listens to it’s baby talk.”
- Almond: public moods may shift quickly
- Ginsberg: Pollsters get to decide which questions to ask, not the public.

DEFINITIONS

- **Statistics**
  - Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.
  - **Variable**
    - Characteristic or attribute that can assume different values
  - **Random Variable**
    - A variable whose values are determined by chance.
  - **Population**
    - All subjects possessing a common characteristic that is being studied.
  - **Sample**
    - A subgroup or subset of the population.
  - **Parameter**
    - Characteristic or measure obtained from a population.
  - **Statistic** (not to be confused with Statistics)
    - Characteristic or measure obtained from a sample.

- **Descriptive Statistics**
  - Collection, organization, summarization, and presentation of data.
- **Inferential Statistics**
  - Generalizing from samples to populations using probabilities. Performing hypothesis testing, determining relationships between variables, and making predictions.
- **Qualitative Variables**
  - Variables which assume non-numerical values.
- **Quantitative Variables**
  - Variables which assume numerical values.
- **Discrete Variables**
  - Variables which assume a finite or countable number of possible values. Usually obtained by counting.
- **Continuous Variables**
  - Variables which assume an infinite number of possible values. Usually obtained by measurement.
- **Nominal Level**
  - Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.
- **Interval Level**
  - Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
- **Ratio Level**
  - Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.
- **Random Sampling**
  - Sampling in which the data is collected using chance methods or random numbers.
- **Systematic Sampling**
  - Sampling in which data is obtained by selecting every kth object.
- **Convenience Sampling**
  - Sampling in which data is obtained by selecting the most convenient object.
- **Cluster Sampling**
  - Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.
- **Stratified Sampling**
  - Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.