HANDOUT #2 - TYPES OF STATISTICAL STUDIES

TOPICS

1. Observational vs Experimental Studies
2. Retrospective vs Prospective Studies
3. Sampling Principles:
   (a) Probability Sampling: SRS, Systematic, Stratified, Cluster
   (b) Estimation of population parameters
4. Experimental Design Principles
5. Common Problems in Designed Experiments
6. Selecting an Appropriate Design

Supplemental Reading:

- Chapter 3 in Tamhane/Dunlop book

Sampling From a Population

The basic goal of most studies is to use a subset of a population to make a statement about the whole population. These types of situations were illustrated in Handout 1 with our examples of market surveys, polling, estimating ozone levels, determining side-effects of drugs, etc.

Two basic types of studies: Observational and Experimental

- **Observational Study**: Records information about subjects without applying any treatments to subjects (passive participation of researcher). The purpose is to describe a group or situation.

  Examples: Challenger Data, Political Polls, Market Surveys, Industrial Production Records, Traffic Accident Studies, Epidemiological Studies

- Observational studies can only show correlation, not causation. Vegetarians, for example, have lower rates of heart disease than the general public. Is this due to their meatless diet? Or because they smoke less and exercise more regularly than people who eat large amounts of meat? Observational studies cannot cannot sort out these kinds of issues.

- **Experimental Study**: Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher). The purpose is to study whether a treatment causes a change in a response.

  Examples:

  - Clinical Trials (Some control),

  - Laboratory Studies (More Control),

  - Agricultural Field Trials (Some Control),

  - Greenhouse Experiments (More Control),

  - Pilot Plants in Industry (More Control)

Observational studies are of four basic types:

- **Sample Survey:** Provides information about a population based on a sample from the population at a specific time point.

  Political Polls, Market Surveys, Customer Satisfaction Questionnaires, some Epidemiological studies

- **Prospective Study:** Observes population in the present by using a sample survey and proceeds to follow the sample forward in time in order to record the occurrence of specific outcomes.

  Example: Academic success of two groups: Head Start vs No Head Start

  Example: Subjects quit smoking then record weight gain over 1 year post smoking

- **Retrospective Study:** Observes population in the present by using a sample survey and collects information from the sample about the occurrence of specific outcomes that have already taken place.

  Examples:

  - Is incidence of colon cancer related to Diet? Collect information about the diets of two groups of people, those with and those without colon cancer.

  - Epidemiological studies: 2015 Listeria bacteria outbreak in Texas. What was the source of the listeria.

  - Tuskegee Public Health Service Study (see details in next few pages)

The Study Begins

In 1932, the Public Health Service, working with the Tuskegee Institute, began a study to record the natural history of syphilis in hopes of justifying treatment programs for blacks. It was called the "Tuskegee Study of Untreated Syphilis in the Negro Male."

The study initially involved 600 black men – 399 with syphilis, 201 who did not have the disease. The study was conducted without the benefit of patients' informed consent. Researchers told the men they were being treated for "bad blood," a local term used to describe several ailments, including syphilis, anemia, and fatigue. In truth, they did not receive the proper treatment needed to cure their illness. In exchange for taking part in the study, the men received free medical exams, free meals, and burial insurance. Although originally projected to last 6 months, the study actually went on for 40 years.

What Went Wrong?

In July 1972, an Associated Press story about the Tuskegee Study caused a public outcry that led the Assistant Secretary for Health and Scientific Affairs to appoint an Ad Hoc Advisory Panel to review the study. The panel had nine members from the fields of medicine, law, religion, labor, education, health administration, and public affairs.

The panel found that the men had agreed freely to be examined and treated. However, there was no evidence that researchers had informed them of the study or its real purpose. In fact, the men had been misled and had not been given all the facts required to provide informed consent.

The men were never given adequate treatment for their disease. Even when penicillin became the drug of choice for syphilis in 1947, researchers did not offer it to the subjects. The advisory panel found nothing to show that subjects were ever given the choice of quitting the study, even when this new, highly effective treatment became widely used.

The Study Ends and Reparation Begins

The advisory panel concluded that the Tuskegee Study was "ethically unjustified"–the knowledge gained was sparse when compared with the risks the study posed for its subjects. In October 1972, the panel advised stopping the study at once. A month later, the Assistant Secretary for Health and Scientific Affairs announced the end of the Tuskegee Study.

INVOLVING HUMAN SUBJECTS WMA DECLARATION OF HELSINKI – ETHICAL PRINCIPLES FOR MEDICAL RESEARCH

Adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964
and amended by the:
29th WMA General Assembly, Tokyo, Japan, October 1975
35th WMA General Assembly, Venice, Italy, October 1983
41st WMA General Assembly, Hong Kong, September 1989
48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996
52nd WMA General Assembly, Edinburgh, Scotland, October 2000
53rd WMA General Assembly, Washington DC, USA, October 2002 (Note of Clarification added)
55th WMA General Assembly, Tokyo, Japan, October 2004 (Note of Clarification added)
59th WMA General Assembly, Seoul, Republic of Korea, October 2008
64th WMA General Assembly, Fortaleza, Brazil, October 2013

Preamble

1. The World Medical Association (WMA) has developed the Declaration of Helsinki as a statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data.

The Declaration is intended to be read as a whole and each of its constituent paragraphs should be applied with consideration of all other relevant paragraphs.

2. Consistent with the mandate of the WMA, the Declaration is addressed primarily to physicians. The WMA encourages others who are involved in medical research involving human subjects to adopt these principles.

General Principles

3. The Declaration of Geneva of the WMA binds the physician with the words, “The health of my patient will be my first consideration,” and the International Code of Medical Ethics declares that, “A physician shall act in the patient’s best interest when providing medical care.”

4. It is the duty of the physician to promote and safeguard the health, well-being and rights of patients, including those who are involved in medical research. The physician’s knowledge and conscience are dedicated to the fulfilment of this duty.

5. Medical progress is based on research that ultimately must include studies involving human subjects.

6. The primary purpose of medical research involving human subjects is to understand the causes, development and effects of diseases and improve preventive, diagnostic and therapeutic interventions (methods, procedures and treatments). Even the best proven interventions must be evaluated continually through research for their safety, effectiveness, efficiency, accessibility and quality.

- **Cross-sectional study:** Involves data collected at a specific point in time. This type of study is often used to assess the prevalence of acute or chronic conditions, or to answer questions about the causes of disease or the results of medical intervention.

  Example: Study the effect of oral contraceptives (OC) on heart disease in women aged 40-44 years. Randomly select 5000 users of OC and 10000 nonusers and record the occurrences or nonoccurrence of myocardial infraction for the 15000 women.

Sample Survey Example

The Bureau of Labor Statistics determines the unemployment rate. The Current Population Survey (CPS), or ”Household Survey”, conducts a monthly survey based on a sample of 60,000 households.

The data is also used to calculate 6 unemployment rates (UR) as a percentage of the labor force based on different definitions:

UR1: Percentage of labor force unemployed 15 weeks or longer.

UR2: Percentage of labor force who lost jobs or completed temporary work.

UR3: Official unemployment rate: % of people who are currently not working but are willing and able to work for pay, currently available to work, and have actively searched for work.

UR4: UR3 + ”discouraged workers” (current economic conditions makes them believe that no work is available for them).

UR5: UR4 + other ”marginally attached workers” (would like” and are able to work, but have not looked for work recently).

UR6: UR5 + Part time workers who want to work full time, but can not due to economic reasons.

Comparison of Retrospective and Prospective Studies

- Retrospective studies are generally cheaper and can be completed more rapidly than prospective studies.

- Retrospective studies have problems due to inaccuracies in data due to recall errors.

  Dietary Study: What did you eat during the past three days?

  Customer Survey: Was your shopping experience at our store enjoyable?

- Retrospective studies have no control over variables which may affect disease occurrence. Dietary Study: There are many other factors other than diet that may impact onset of Colon Cancer - Genetics, Occupation, Environment

- In prospective studies subjects can keep careful records of their daily activities

  Diet Diary, Check Ups, Record weight at 8am every day

- In prospective studies subjects can be instructed to avoid certain activities which may bias the study

  Exposure to risk factors - environmental toxins,work and personal stress factors, etc.

- Although prospective studies reduce some of the problems of retrospective studies they are still observational studies and hence the potential influences of confounding variables may not be completely controlled. It is possible to somewhat reduce the influence of the confounding variables by restricting the study to matched subgroups of subjects.

  Group subjects according to similar occupations, ethnicity, location of residency

- Both prospective and retrospective studies are often comparative in nature. Two specific types of such studies are **cohort studies** and **case-control studies**.

  - Cohort Studies: Follow a group of subjects forward in time to observe the differences in characteristics of subjects who develop a disease with those who do not. Prospective or Retrospective?

  - Case-Control Studies: Identify two groups of subjects, one with the disease and one without the disease. Next, gather information about the subjects from their past concerning risk factors which are associated with the disease. Prospective or Retrospective?

  - Case-Control studies are an improvement over just taking a random sample. Why?

    If a disease is very rare, then a random sample from the population may have only a very small number of individuals with the disease.

Sampling versus Non-sampling Errors:

A sample provides only an estimate of the whole population because we only observe a fraction of the units contained in the population. The difference between the information contained in the sample and the information contained in the population is called **Sampling Error**. In theory, the sampling error can always be eliminated by simply increasing the sample size until we have observed the whole population. However, even when we attempt to observe the whole population, called a census, errors may still exist. These are called **non-sampling errors**.

Non-sampling errors may cause biases/systematic errors in the sample estimates. These are consistent deviations of the sample estimates from the true population values. These are truly problematic because even if we greatly increase the sample size, the biases will persist. Several of these of errors are listed below:

1. Measurement bias: a measuring device which always records the value for the sampling unit either smaller or larger than the actual value. Improperly worded questionnaires or unclear questions in a survey can result in measurement bias. The interviewer's body language can result in the respondent giving answers which do not truly reflect their position on an issue.

2. Self-Selection bias: The people who choose to participate in a survey may be a totally different subset of the population from those people who choose not to participate:

   Younger people participate at a lower rate than older persons

   Politically active persons participate at a higher rate than those who are not politically active

   Higher income and lower income persons participate at a lower rate than middle income persons

   Persons who return survey may have a strong opinion about issue whereas persons who do not return survey have no opinion - end of the semester student evaluation of course/instructor.

3. Methods of selection sample bias: Random digit-dialing in telephone surveys are problematic in that many people screen their phone calls and only answer the phone when the call is from a person they know using caller ID or they use their answering machine to screen calls.

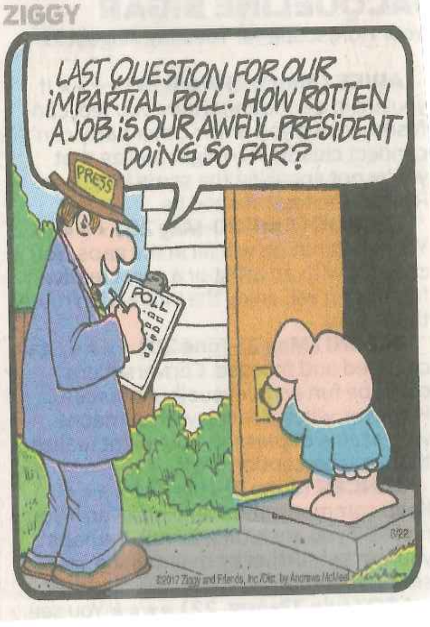
4. Response bias: Untruthful responses can occur due to the asking of very personal questions or questions which require the recall of events from the distant past.

   Did you inhale?

   Do you use illegal drugs?

   Have you ever cheated on your income tax form?

ZIGGY


LAST QUESTION FOR OUR IMPARTIAL POLL: HOW ROTTEN A JOB IS OUR AWFUL PRESIDENT DOING SO FAR?
PRESS
POLL
8/22

©2017 Ziggy and Friends, Inc./Dist. by Andrews McMeel

5. Timing of Poll: How close to the election was a political poll taken?

   If poll is too far from election, voters may receive new information about candidates that may change their mind.

6. Non-response: Selected person/experimental unit does not respond

   Person refuses to answer telephone, does not send survey back, refused request to answer questions

   In agriculture or wildlife surveys, we would refer to non-response as "Missing Data".

   A predator may raid a bird's nest and consume the eggs so the number of eggs laid cannot be recorded

   A field of corn may be partially consumed by a herd of deer so that total yield cannot be recorded

7. Possible ways to deal with non-response:

   a. Design survey so that non-response is low

      Follow up phone calls to non-respondents

      Offer payment/donation to charity if survey is returned

   b. Randomly select a subset of non-respondents and use subset to make inferences about other non-respondents

   c. Use a statistical model to predict the responses for the non-respondents.

   d. Ignore the non-response - VERY bad idea but often occurs in practice.

8. The following two articles from *The Atlantic* and *Politico* discuss possible problems in polls.

Went Wrong With the 2016 PollsP1.pdf

What Went Wrong With the 2016 Polls?

The Atlantic November 9, 2016

Donald Trump's surprise victory poses the question: How did we get this thing *this* wrong? From the myriad polls and poll aggregators, to the vaunted oracles at Nate Silver's FiveThirtyEight and the *New York Times*'s shiny forecasting interface, most serious predictors completely misjudged Trump's chances of victory.

Though election night had the appearance of an unlikely come-from-behind victory by Trump, that narrative only exists because virtually all predictions—perhaps even from the Trump camp—started with the assumption that Trump was an underdog. In reality, when viewed with proper perspective, Trump sailed to a rather easy victory, challenged Clinton in several stronghold states, and realistically wrapped the election well before midnight. That kind of result doesn't come out of nowhere, but few pre-election polls even began to pick up such large effects.

So what happened? *Caveat emptor*: If pollsters don't really know the answer, we probably won't really know it for some time. Also, as of the time of this writing, Hillary Clinton is ahead in the popular vote totals, meaning that polls showing her ahead by a few points in head-to-head matchups with Trump were wrong in magnitude, but not directionality.

National polls don't usually show Electoral College vote counts, and don't often maintain the granularity to make the kind of state-by-state predictions to make those projections, so their usefulness even in aggregate to forecast elections is limited. Given that electors are determined by congressional representation, that representation is only reapportioned every 10 years, and that the overall number has not increased in over 50 years, there is an increasing discrepancy between the popular vote and the actual outcome of elections, one that will make national overall polls that simulate the popular vote less relevant to predictions over time.*

Forecasting sites and models have keyed into this discrepancy and had success over the past few election cycles by aggregating smaller state and county-level polls, and then forecasting actual Electoral College votes from those aggregates. That approach has obvious advantages, but suffers sometimes from lack of available and reliable data. As a rule, many state and local polls are newer and more volatile than national polls, and several rely necessarily on unorthodox methods to achieve enough proper sample sizes, which are also often much lower than national polls. Also, the baseline statistics from Census products and other large surveys used for "weighting" state and local results become less reliable as they drill down.

Long story short: Statistical power is important, and any misrepresentation of the population in the sample or weights can lead to unusable results.

The problem with finding accurate and random samples of voters to poll has plagued polling since cell phones came into wide use. Prior to that technological development, the ubiquity of landline telephones made finding reasonably-random and representative samples easy, as pollsters could just pick random names out of phone books, call potential voters, and talk them

Went Wrong With the 2016 PollsP2.pdf

through interviews, which supplied the kinds of rich context and human understanding necessary for properly analyzing their responses. That method also ensured reasonably high response rates and helped control nonresponse bias, by which the polls themselves become skewed by the *kinds* of people who tend to answer.

But the rise of cell phones and the demographic differences of their adoption meant that random samples of landlines became increasingly inadequate in finding good samples. The problem with moving to cell phones or even attempting a hybrid approach is that cell phones are not usually publicly-listed, making it harder and harder to find representative samples. Various online survey methods have been used to supplement or supplant more expensive and less expansive phone methods, but they often also suffer from bias and are generally considered of lower quality than other polls.

Did we all believe Clinton would win because of bad data, or did we ignore bad data because we believed Clinton would win?

The difficulties in polls are illustrated by FiveThirtyEight's final forecast model of Pennsylvania, where only three of the model's polls from the week before the election were rated by the site as an "A-" or above. The poll with the most weight in that model is the Remington Research Poll, a robo-call-powered poll run by former Ted Cruz manager Jeff Roe that does not appear to publish its sampling or weighting methodology, and thus has not been given a rating by FiveThirtyEight.

The most recent poll in that model came from the mixed landline and online Gravis Marketing poll, and featured results with a whopping 3 percentage-point margin of error and a sample that was weighted not to Pennsylvania demographics, but to national demographics. One other poll in the aggregate is the SurveyMonkey poll, which is likely limited by its reliance on a largely skewed group of voters—people who respond to SurveyMonkey polls. Each of these showed Clinton leads in the state that Donald Trump eventually won.

New forecasting models of aggregation like FiveThirtyEight's are marvels in increasing predictive power, and work well in smoothing out the kinks of individual state polls by increasing their statistical power in groups, but when those polls suffer similar problems, those models might theoretically amplify their discrepancies.

Namely, if polls tend to weight Democratic or Republican likely-voters and demographics based on 2012 elections patterns or older demographic distributions, they will naturally miss out on big shifts in the composition of likely-voters or where they live. If high numbers of the wealthy, white, educated pieces of the Obama coalition turned out for Trump, and he also picked up unprecedented turnout from rural voters, models that weight data to recent past elections might understate those effects. Many of these polls might be ill-suited to understanding sudden changes in the electorate or the way the electorate votes.

There are some solutions to this "likely-voter" problem in polls, but many of them involve methods that might make several cheap and accessible polls less so. Utilizing advanced statistics, analyzing previous similar election events, using machine-learning, and creating "kitchen-sink" models based on voter rolls are established ways to improve the underlying assumptions of polls.

But those methods might be a bit too costly and time-intensive for polls that use online surveys and publicly-available annual Census data precisely because they tend to be cheaper than deep research.

Bad models happen, and the very nature of what appears to be the Trump constituency probably made most models worse. Forecasts are best at telling us what old data tells us about new data, and the thing about using existing data is that large deviations in the underlying assumptions of those data may go unnoticed. Those deviations are especially dangerous when they bolster existing confirmation bias among analysts and journalists, but the directionality of that bias is often unclear. Did we all believe Clinton would win because of bad data, or did we ignore bad data because we believed Clinton would win? There's the question for the ages.

Perhaps the lesson here about the Trump presidency is that it was truly unpredictable. Good models often fail to accommodate events outside of the bounds of their sensitivity, and sounding the alarm on their flaws would necessarily involve knowing or suspecting *more* about elections than the data we fed the polls.

For many unfortunate Cassandras like Silver himself, caution was roundly ridiculed from this lack of perspective. But if this is the new normal, pollsters will have to adapt in order to maintain relevance.

Pollsters tackle what went wrong – Politico November 11, 2016

They know they screwed up. Pollsters have a few ideas why.

It’s possible Donald Trump’s upset victory this week was powered by a surge of late deciders. Or the mysterious group often referred as “shy Trump" voters somehow escaped their radar. Many in the polling industry are also second-guessing their turnout modeling, trying to discern whether there’s a serious flaw that went unnoticed.

No matter the root cause, an industry already reeling from a series of misses in the United States and overseas is engaging is a round of serious introspection. While the data streams required to evaluate whether they modeled the electorate incorrectly — or whether Trump voters disproportionately wouldn’t respond to polls — won’t be available for months, already the nation’s leading professional organization of pollsters is admitting it “clearly got it wrong this time” and pledging to study the causes of the errors.

“It seems like the catastrophic polling error that we’ve been fearing for decades,” said Jon Cohen, the vice president of survey research at SurveyMonkey and a former pollster for The Washington Post and the Pew Research Center. “But it may prove to be less than that.”

The polls underestimated Trump — most acutely in a number of battleground states viewed as leaning in Hillary Clinton’s direction — much as they systematically underestimated Republicans in the 2014 midterm elections. But the polls were off in the other direction in 2012, with national surveys understating President Barack Obama’s margin of victory by about 3 points.

Overseas, the polls badly missed the 2015 election in Great Britain. The polls were closer in Britain’s vote to leave the European Union this year, though they are often blamed to a greater degree because the result stunned so many observers.

The pre-election polls in this year’s presidential election could actually end up closer to the actual result than the polls four years ago, at least on the national level. But there are no laurels for that. Trump won the Electoral College and not Clinton, so it’s viewed as a far more significant polling malfunction.

Even if pollsters don’t yet know precisely what went wrong, they are asking the questions — and cautioning against downplaying the extent of the breakdown this week.

The pollsters who are out there saying the polls are really OK because they all fall within the margin of error, that is just not credible — and not helpful,” said Democratic pollster Jefrey Pollock, whose firm, Global Strategy Group, worked for the pro-Clinton group Priorities USA.

“It is a mistake to go out and say polling is useless,” Pollock added. “It is just as big a mistake to claim the polls are all OK because they all fall within the margin of error.”

The potential causes for error run the gamut from a secret army of Trump voters lying to pollsters, to a late-breaking wave of voters flocking to Trump after most of the polling had concluded.

Here are the questions pollsters are asking this week:

Did Trump surge at the end?

There's some limited evidence that's the case.

National polls certainly closed in the last few weeks of the race — but, in the final days, it appeared Clinton had arrested Trump's momentum.

Indeed, the national exit poll suggests she did: Trump led by 10 points among voters who said they made up their minds in the final month of the campaign, but his lead among voters who decided in the last week was only 5 points. (Clinton led by 5 points among voters who decided before October, which comprised the majority of the electorate.)

In some of the states where Trump won unexpectedly, however, Trump won more late deciders. In Wisconsin, where the latest results show Trump ahead by a point, Trump overwhelmingly won voters who made up their minds in the final week, 59 percent to 30 percent. In Pennsylvania, where Trump is also ahead by a point, he won last-week deciders, 54 percent to 37 percent. In Michigan — which hasn't yet been called with Trump ahead by three-tenths of a percentage point — Trump won those who decided in the final week, 52 percent to 37 percent.

That kind of late movement can thwart pollsters who conduct surveys in the weeks leading up to an election.

"If the accuracy of your final forecast is the most important thing, then there is an incentive to poll right up until the last minute," said Charles Franklin, who conducts the Marquette Law School poll in Wisconsin.

Franklin, a member of the American Association for Public Opinion Research's task force that will examine the election polls, said he concluded polling in Wisconsin on Oct. 31, more than a week before Election Day.

"I made a policy decision when we started this that we would release [the final poll] the Wednesday before the election, in part so we could give the campaigns a chance to react," Franklin said. (That Marquette Law School poll had Clinton ahead by 6 points.)

Clinton aides blame loss on everything but themselves

By Annie Karni

In fact, the final poll in Wisconsin from any nonpartisan outlet was conducted fully a week before Election Day, leaving no instrument to capture if there was last-minute movement.

If that movement did occur late, however, it might make sense that it would be against Clinton, who was running for a third-consecutive Democratic term.

Republican pollster Dan Judy pointed to Clinton's vote shares in a number of states, relative to where she finished in the final polling averages. The result? She only scored a point or so higher in most states on Election Day — suggesting undecided and some voters choosing third-party candidates in the pre-election polls drifted to Trump in larger numbers.

Clinton was at 46 percent in the Wisconsin polls, according to HuffPost Pollster. She won just shy of 47 percent of the vote. She was at 46 percent in the Pennsylvania polls and won less than 48 percent of the vote. In Michigan, she was at 47 percent in the polls and won that percentage on Election Day.

That reflects an old rule of politics, Judy said — one that hasn't always applied in the past: Undecideds vote against the incumbent in the end.

"The swing states and 'Blue Wall' states that Trump won definitely treated Hillary like an incumbent," Judy said. "It's remarkable how close her actual percent was to her final polling average across a lot of the most competitive states. She really was polling like an unpopular incumbent — maybe not surprising given the run for a third term, and Obama's aggressive campaigning for her down the stretch."

That was also evident in a key number from the exit poll: The vote preferences of the roughly 1-in-5 voters who had an unfavorable opinion of both candidates. Among those voters, who made up a historically high 20 percent of the electorate, Trump won by 21 points: 50 percent to 29 percent.

Those voters, who could have been among the last to decide, helped propel Trump despite the fact that more voters viewed Clinton favorably (42 percent) than had a favorable opinion of Trump (39 percent).

"It looks like the biggest chunk in the end chose to take the dive with Donald Trump," said Joe Lenski, the researcher who oversees the exit poll for Edison Research. "That's the only group that puts him over. In any other election, if you just had the two favorable numbers, you would say the person with the highest favorable number would win."

Were there “shy Trump” supporters?

It was a theory POLITICO sought to test in late October, along with the online pollster Morning Consult: Were there people who voted for Trump but wouldn’t admit it to a pollster?

The study found only a slight impact by moving poll respondents from the internet to a phone call with a live interviewer — with larger effects among college-educated white voters.

Perhaps that happened on Election Day: Exit polls — which are an imperfect but immediately available record of who voted, for whom they voted and what they thought about candidates and issues — indicate greater support for Trump among college-educated white voters than the pre-election polls suggested.

“The discussion was all about white non-college men and women,” Cohen said. “But it’s the white college constituency that look dramatically different when you look at the pre-election polls versus the exit polls.”

But it might not be because voters are lying to pollsters — telling them they won’t support Trump but voting for him on Election Day. It might be because they don’t pick up the phone in the first place.

“Differential turnout and participation in surveys is maybe the more worrisome” factor, said Franklin, the Marquette Law School pollster. “If there was some percentage of Trump supporters that refused to do any polling but did go to the polling place, then we’re missing them completely in our samples.”

Who is a likely voter?

Election pollsters are always trying to identify a universe of people that doesn’t yet exist: the future electorate.

Every pollster does that differently: Some allow every voter into their poll that says they are certain to vote. Some make assumptions about who will turn out, including party identification or registration. And most campaign pollsters add what they consider is the most important factor: whether voters have turned out before.

Trump appears to have upended some of those approaches. But it will be months before pollsters know how it happened.

The first clues are trickling in now as the votes are tallied. Once all the votes are tabulated, pollsters will know whether turnout was greater or lower than expected — and, most importantly, where those trends apply.

Was turnout markedly lower in urban centers Clinton needed, like Milwaukee, Philadelphia and Detroit? Early indications, especially in Milwaukee and Detroit, indicate drops in turnout.

Meanwhile, turnout appears higher outside cities and suburbs.

“If you look at these states, and you look at the turnout ratios from four years ago county by county, it’s pretty clear the biggest percentage increases in turnout were in the non-urban, non-suburban areas,” said Lenski, who administered the exit polls. “That’s hard to predict both in a pre-election poll and an exit poll.”

In Pennsylvania, for example, Clinton carried more votes out of Philadelphia and the suburbs than President Barack Obama in 2012 — but polls missed the higher turnout in more rural areas of the state.

"If you just showed me Hillary’s numbers in the Southeast, I would have said she would have won by 2 or 3 points,” said Christopher Borick of Muhlenberg College in Allentown. (Muhlenberg’s final poll had Clinton ahead by 6 points in a head-to-head matchup with Trump, and 4 points in a four-way matchup.)

But a complex analysis of the electorate — beyond just from where the votes came — will take months. Pollsters will be able to look at precisely who voted — whether they were regular voters or less-frequent voters drawn out by Trump’s unique candidacy — and who didn’t.

“That will give us the best evidence about new voters, about previous voters who dropped out,” said Franklin. “That will be incredibly valuable.”

For pollsters trying to figure out what happened this week, those voter files — in addition to next year’s Census Current Population Survey — will be worth the wait. Trump’s candidacy rocked the political system, from the Republican primary through the general election. And a Trump presidency could upend how Americans view and interact with their government in a similar way.

“We’ve reacted well to failure before,” said Cohen, the SurveyMonkey pollster. “Polling is too important to go away. The way that we are going to understand what happened in the election, and the contours of where we sit as a country, is through polling.”

SAMPLING PRINCIPLES

Suppose a researcher wants to make an inference about a specific population. They may choose to inspect a small portion of the population, a **sample**. Alternatively, they could perform a **census**, that is, an inspection of the entire population.

Why select a sample in place of a census?

- Reduced cost
- Less time consuming
- More information per subject - Less effort expended per sampling unit
- Greater accuracy - better training of technicians, more accurate measurements, subjects may be missed in census
- Census may be impossible in a mobile population
- Measurement may require destroying units being inspected

Two important questions need to be answered while either designing a study or reading the results of a study:

What is the population of interest to the researcher?

- All diesel powered VW cars less than 10 years old to check on the exhaust emissions

What is the method by which the sample was selected?

- Examine all VW cars in used car lots in Houston

What is the population of interest to the researcher?

- Impact on undergraduate college students to switching all core courses to online courses

What is the method by which the sample was selected?

- Evaluate the results of the switch to online courses on students' at Texas A&M

Thus, it can then be concluded whether or not the sample is properly selected from the population of interest.

Sampling Frame A complete list of all N units in the population

Note: There is a 1-1 correspondence between the numbers $1, 2, \ldots, N$ and the sampling frame.

Probability Sampling

1. Given a frame, one can define all the possible samples that could be selected from the population. Label the distinct samples $S_1, S_2, \ldots, S_k$.
2. Assign a probability S_i, $P(S_i)$, to each possible sample S_i, $\sum_{i=1}^{k} P(S_i) = 1$.
3. The sample is selected by using a random process in which the sample S_i has probability $P(S_i)$ of being chosen.

Advantage of probability sampling: *allows an objective assessment of the accuracy of inferences made about the population based on the information in the sample.*

Two questions that **must** be answered when viewing a research study or polling results:

- What is the population of interest? and
- How were the observed units selected from the population?

Example of non-probability sampling:

1. **Convenience Sample:** Data selected based on the availability of data.

   Examples of convenience samples:

- Historical data, Medical records, Production records, Student academic records
- Select next 50 people who walk in a store
- Meat inspector inspects just the packages conveniently provided by the meat store

  Problems: Data may yield a sample which is not representative of the population due to many uncontrolled variables which may be confounded with the sampling strategy.

- The next 50 people going in the store may be off the same bus which is carrying people from a particular religious or political organization
- Use Instructor's classroom of 75 undergraduates in instructor's research project

2. **Judgemental Sample:** an expert selects "typical" or "representative" members of the population.

   Problem: This type of process is extremely subjective and does not admit a scientific assessment of accuracy.

- Biased by personal judgement or level of expertise
- Participants in survey are selected according to economic status
- Selected because there are members of "influential organization"

RANDOM SAMPLING

SRS is the most basic method of taking a probability sample. In this method of selecting a sample of n units from a population of N, each of the $\binom{N}{n}$ possible samples has the same chance of being selected. The actual choice of a specific sample can be done using a random number generator on a computer. The following R commands can be used.

```
  The following R  commands generate random permutations of n integers or
    random sample from a population of numbers.

1. Random permutation of integers 1 to n :  "sample(n)"

    EX.       sample(10)

              3  8 10  6  9  5  1  4  7  2


2. Random permutation of elements in a vector x: "sample(x)"

    EX. x<-c(23,45,67,1,-45,21,.9,4,-3,.25)

              sample(x)

        -3.00  45.00  21.00   0.90   0.25  23.00  67.00   4.00 -45.00   1.00



3. Random sample of n items from x without replacement: "sample(x,n)"


    EX.       sample(x,5)


              67.00  21.00  45.00   0.25 -45.00


4. Random sample of n items from x with replacement: "sample(x,n,replace=T)"


    EX.       sample(x,5,replace=T)


              -45.0   4.0  -3.0 -45.0   0.9
```

5. Random sample of n items from x with elements of x having differing probabilities of selection: "sample(x,n,replace=T,p)", where p is a vector of probabilities, one for each element in x.

```
EX.        x<-c(23, 45, 67, 1,-45, 21, .9, 4,-3,.25)

           p<-c(.1, .1, .1, 0,  0,  0,  0, 0, 0, .7)


           sample(x,5,replace=T,p)

           0.25  0.25 45.00  0.25  0.25
```

6. Randomly select n integers from the integers 1 to N, without replacement:

   "sample(N,n)"

```
EX.        sample(1000,10)


           189 182 638 903 112 126 490 928 850 291
```

7. Randomly select n integers from the integers 1 to N, with replacement:

   "sample(N,n,replace=T)"

```
EX.        sample(1000,10,replace=T)


           189 182 638 903 112 182 490 928 850 291
```

For example, suppose you have 500 units and randomly select 10 units for destructive inspection. There are $\binom{500}{10} = 2.458x10^{20}$ distinct samples of size 10 that are possible

SYSTEMATIC RANDOM SAMPLING

Suppose we have a list of the population units or units are produced in a sequential manner. A **1-in-k** systematic sample consists of selecting one unit at random from the first k units and then selecting every kth unit until n units have been collected. In a population containing N units, systematic sampling has a selection probability of $\frac{n}{N}$ for each unit. However, not all $\binom{N}{n}$ possible samples are equally likely, as in SRS.

In essence, we are forming k clusters of n units each:

$$C_1 = \{U_1, U_{k+1}, U_{2k+1}, \ldots, U_{(n-1)k+1}\}$$

$$C_2 = \{U_2, U_{k+2}, U_{2k+2}, \ldots, U_{(n-1)k+2}\}$$

$$\vdots$$

$$C_k = \{U_k, U_{2k}, U_{3k}, \ldots, U_{nk}\}$$

Randomly select 1 of the k clusters

The chance that a particular unit is selected is $\frac{1}{k} = \frac{1}{N/n} = \frac{n}{N}$

Example: Suppose we have $N = 1000$ units, $U_1, U_2, \cdots, U_{1000}$ and we want to sample $n = 10$ of the units. Select $k = \frac{N}{n} = 100$.

Randomly select a number between 1 and 100, say, 23

The Sample then consists of the following units:

$$U_{23},\ U_{123},\ U_{223},\ U_{323},\ U_{423},\ U_{523},\ U_{623},\ U_{723},\ U_{823},\ U_{923}$$

Systematic sampling is often used when a sequential list of sampling units exists or when sampling units become available in a sequential manner. Systematic sampling provides a sample which is representative of the population provided there are no cyclic patterns in the population lists.

Example Parts are inspected on a production line with every 20th part inspected

Example A jury of 50 persons is selected from a list of 50,000 registered voters or driver license holders by randomly selecting a person from first 1000 persons on list, e.g., the 452 person and then including the $1452, 2452, 3452, \ldots, 49452$ persons on the list.

Possible Problem with Systematic Sampling: Suppose the production process produces units such that a set of 1000 consecutively produced units has the following pattern: the first 50 units in any sequence of 100 units are very different from the second set of 50 units. If the number 23 is selected then we would only sample units from the first 50 units whereas, if the number 77 was randomly selected then we would only sample units from the second 50 units in every batch of 100 units. The sample of 100 units would provide a distorted view of the 1000 units.

STRATIFIED RANDOM SAMPLING

Population is divided into L groups or strata. The strata are non-overlapping and contain $N_1, N_2, \ldots, N_L$ units respectively. Note: $N_1 + N_2 + \cdots + N_L = N$. Suppose simple random samples of sizes $n_1, n_2, \ldots, n_L$ are selected independently from the L strata. This sampling procedure is known as *stratified random sampling.*

Reasons for Using a Stratified Random Sample:

- Precise estimates within subpopulations (strata)
- Administrative convenience
- Sampling problems differ according to different parts of the population.
- Possible gain in precision in the overall estimate of population parameters. (This occurs when there are large differences between stratum but there is homogeneity within the L strata.

Example of stratified sampling: Suppose we wanted to determine the percentage of people in Texas who have health insurance.

- Stratify counties by into four strata: rural, mostly small towns, medium size cities, large metropolitan area
- Randomly select n_i people from each of the four strata.

CLUSTER RANDOM SAMPLING

Population consists of N primary sampling units (psu's) or clusters. The N clusters contain $M_1, M_2, \ldots, M_N$ smaller units called secondary sampling units (ssu's) or elements. Population contains a total of

$$\sum_{i=1}^{N} M_i = M^* \quad \text{elements}$$

For example, suppose the research objective is to determine how many bicycles are owned by residents in a community of 10,000 households. A simple random sample of 300 households could be used to address this problem. However, an alternative sampling plan would divide the community into 500 blocks of approximately 20 households each and randomly select 15 blocks from the 500 blocks of households. Each household in the 15 selected blocks would then be surveyed.

Single-Stage Cluster Sample A SRS of n cluster is selected and all elements within each cluster is measured or surveyed.

In the example, the clusters are the blocks of households and the elements are the individual households.

Suppose $M_i = M$ for all i. What advantage is there to taking a cluster sample of nM elements as opposed to a SRS of nM elements from the population? In general, the cluster sample will be less precise than the SRS due to units from the same cluster are more alike than units from different clusters. The main reason for using cluster sampling is administrative difficulties of obtaining a frame for all M^* elements in the population. For example, suppose an element is a household in Houston. Define a cluster as a city block in Houston. Obtaining a frame of all city blocks in Houston is undoubtedly easier than obtaining a frame of all households in Houston.

Multi-Stage Cluster Sample A SRS of n cluster is selected from the population of N clusters. Random samples of elements of size $m_1, m_2, \cdots, m_n$ are selected from the n clusters and each of the selected elements is measured or surveyed.

Stratified Sampling vs Cluster Sampling

Stratified Sampling:

1. Often will yield smaller value for $Var(\hat{\mu})$
2. Guarantees population elements will be selected into the sample from each stratum
3. Allows estimation of means for each stratum.
4. May be more convenient and less expensive to administer
5. Requires a sampling frame for each stratum

Cluster Sampling:

1. Useful when sampling frame for clusters is available but there is not a frame for the individual elements
2. Useful when elements are individuals that need to be interviewed or selected objects that need to be measured
3. Population elements may be widely separated or may occur in natural clusters such as households or schools

Example 1:

The EPA designed a study to determine the impact of chemical discharges on the water quality in lakes. The study involved first randomly selecting 10 states from the 50 states. Next, a random sample of m_i lakes is taken from a list of polluted lakes within each of the selected states. At each of the selected lakes, a determination of the water quality is made at each of the points where there is a chemical discharge into the lake. This example is what type of study/sampling method?

- States are Clusters of Lakes
- Each lake contains 1 or more discharge points
- PSU is a State, randomly selected from 50 states
- MU is a discharge point in lake

This is a Multistage Cluster Random Sample

Example 2:

A study was designed to evaluate the effects of feral pig activity and drought on the native vegetation in rural northern California. The researcher divided northern California into 20 regions. Within each of these regions she randomly selected 10 oak trees and placed an identifier on a random sample of eight seedlings under each of the trees. Two years later she returned and determined the amount of damage to each these woody seedlings. This example is what type of study/sampling method?

- Region is a Stratum with the population of Northern California
- Oak trees are clusters of Seedlings
- PSU is a oak tree, randomly selected from population of oak trees in each region
- MU is a woody seedling

This is a Stratified Multistage Cluster Random Sample

ESTIMATION OF POPULATION MEAN: μ

Consider the estimation of μ under three different sampling Methods

Simple Random Sampling

Let $y_1, y_2, \ldots, y_n$ be the measurements obtained from the SRS of n units from the population. The estimator of the population mean μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

with estimated variance of $\hat{\mu}$ given by

$$\widehat{Var}(\hat{\mu})_{SRS} = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Note, $\widehat{Var}(\hat{\mu}) \approx \frac{s^2}{n}$ provided $\frac{n}{N}$ is very small or $\widehat{Var}(\hat{\mu}) = \frac{s^2}{n}$ if sampling is with replacement

Stratified Random Sampling

Suppose we have independently selected SRS's of size $n_1, n_2, \ldots, n_L$ from the L strata. Let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_L$ be the sample means of the L SRS samples selected from the L strata with the number of units in each stratum - given by $N_1, N_2, \ldots, N_L$. The estimator of the population mean μ is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$$

with estimated variance of $\hat{\mu}$ given by

$$\widehat{Var}(\hat{\mu})_{STRATIFIED} = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right)\frac{s_i^2}{n_i}\right]$$

where $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$.

Note, $\widehat{Var}(\hat{\mu})_{STRATIFIED} << \widehat{Var}(\hat{\mu})_{SRS}$ when $s_i^2 << s^2$

SINGLE STAGE CLUSTER Random Sampling

Let N be the number of clusters in the population;

n be the number of clusters selected in a simple random sample from the population;

m_i be the number of elements in cluster i, $i = 1, 2, \ldots, N$;

$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ be the average cluster size for the sample of n clusters,

$M = \sum_{i=1}^{N} m_i$ be the number of elements in the population,

$\bar{M} = \frac{M}{N}$ be the average cluster size for the population,

$y_i = \sum_{j=1}^{m_i} y_{ij}$ be the total of all measurements of the m_i elements in the ith cluster.

The estimator of the population mean μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{\sum_{i=1}^{n} m_i \bar{y}_i}{\sum_{i=1}^{n} m_i} = \sum_{i=1}^{n} \frac{m_i}{\sum_{i=1}^{n} m_i} \quad \bar{y}_i$$

with estimated variance of $\hat{\mu}$ given by

$$\widehat{Var}(\hat{\mu}) = \left(\frac{N-n}{Nn\bar{M}^2}\right) \frac{\sum_{i=1}^{n}(m_i\bar{y}_i - m_i\hat{\mu})^2}{n-1}$$

Based on the very large difference in the above formulas for $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$, it is crucial that we know what type of sampling procedure was used in obtaining the data from the population. If we always assumed that a simple random sample was used, our computation of $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ could be grossly incorrect, if in fact, some form of stratified or cluster sampling was the method of sampling used in collecting the data.

If you are interested in learning more about sample surveys and these types of estimation procedures, then I would suggest that you take STAT 607.

EXPERIMENTAL DESIGN PRINCIPLES

The rest of the handout will be covered in STAT 642

Two very important comments from noted pioneers of applied statistics:

- "Whenever possible, experiments should be comparative. For example, if you are testing a modification of a process, the modified *and* unmodified processes should be run side by side in the same experiment." (G. Box, S. Hunter, and W. Hunter)

- "It is possible, and indeed it all too frequent, for an experiment to be so conducted that no valid estimate of error is available. In such a case, the experiment cannot be said, strictly speaking, to be capable of proving anything." (R.A. Fisher)

Selected Comments from *Experimental Design* by W. Federer

I. "All fields of research have at least one feature in common:

   The variability of experimental responses."

II. When there is considerable variation from observation to observation on the same experimental material and it is not feasible to run a large number of experiments (which would reduce the variation in the mean response), THEN the experimenter must:

   1. Refine the experimental design in order to obtain a specified degree of precision (Blocking)
   2. In order to attach a probability statement to the observed treatment mean differences (a measure of the degree of confidence in the observed results), it is necessary that proper Randomization and Replication occur.

III. Certain Principles of Scientific Experimentation should alway be followed: (Many are nonstatistical, however, the analysis of the data resulting from improperly designed and conducted experiments may complicate the analysis to the point at which NO analysis of the data can be conducted.)

   P1. Formulation of Questions to be Asked and Research Hypotheses to be Tested:

   Clearly stating and precisely formulating questions and hypotheses prior to the running of the experiments will help to

   1. Minimize the number of replications required
   2. Make sure all necessary measurements are taken.

   P2. A Critical and Logical Analysis of the Stated Research Hypotheses:

   1. Review the relevant literature

2. Evaluate the reasonableness and utility of the aim of the experiment as reflected in the Research Hypotheses. (May need to reformulate the Research Hypotheses.)
3. Forecast the possible outcomes of the experiment in order to determine if the resulting data can be analyzed using the proper statistical methodology: For example,

   Too many 0's,

   Categorical data,

   Too few replications for projected variability,

   Correlated (nonindependent observations)

P3. Selection of Procedures for Conducting Research

1. What Treatments to be included in experiment?
2. What Measurements should be made on the experimental units?
3. How should experimental units be selected?
4. How many experimental units should be used?
5. What sampling or experimental design should be used?
6. What is the effect of adjacent experimental units on each other? How can this effect be controlled? (Competition between experimental units leads to dependent data.)
7. Outline of pertinent summary tables for recording data.
8. Experimental procedures outlined and documented.
9. Statement of costs in terms of materials, personnel, equipment.
10. Consideration of the above items may often result in a restructed experiment, rather than an experiment in which the results are highly incomplete and not very useful.

P4. Selection of suitable Measuring Devices and Elimination of Personal Biases and Favoritisms:

1. Never observe 3 samples and discard "most discrepant" observation
2. Never place "Favorite Treatment" under the best experimental conditions
3. Discard $0's$ or values from abnormal experimental units only after a **critical examination** of the experimental units and a determination of the degree of unsuitablity of the results in reference to standard experimental conditions. **Always** report the data values and explain why they were excluded from the analysis.

P5. Carefully evaluate the statistical tests and the necessary conditions needed to apply these tests with respect to experimental procedures and underlying distributional requirements. (Residual analysis to check that assumptions hold.)

P6. Quality of the Final Report:

1. Include well designed graphics

2. Include description of statistical procedures and data collection methodology so that the reader of the report can determine the validity of your experiment and analysis.
3. Report should be prepared whether or not the research hypotheses have been supported by the data; otherwise Type I errors alone may produce misleading conclusions. Many experiments result in the acceptance of the null hypothesis but no report is written. Thus, even when the research hypothesis is in fact false but many experiments were conducted concerning this hypotheis, there may be a number of these experiments (5% Type I Errors) that support this research hypothesis incorrectly whereas a large number of experiments (95%) in fact find that the research hypothesis is not supported by the data but since report is written the research hypothesis may be incorrectly supported in the literature.
4. It is crucial that the size of the treatment effect, for example an estimate of $\mu_i - \mu_{i'}$, be reported and not just the p-value of the test. Include confidence intervals on the effect size. Thus, a distinction is being made between **Statistically Significant Results** (small p-value) and **Practically Significant Results** (small p-value with large Treatment effect).

IV. Statistically Designed Experiments are

- Economical
- Allow the measurement of the influence of several factors on a response
- Allow the estimation of the magnitude of experimental variability
- Allow the proper application of statistical inference procedures

EXPERIMENTAL DESIGN TERMINOLOGY

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:

a. Completely Randomized Design (CRD)
b. Randomized Complete Block Design (RCBD)
c. Balanced Incomplete Block Design (BIBD)
d. Latin Square Design
e. Crossover Design
f. Split Plot Design
g. Many others

C2. Treatment Structure

a. One Way Classification
b. Factorial
c. Fractional Factorial
d. Fixed, Random, Mixed factor levels

C3. Measurement Structure

a. Single measurement on experimental unit
b. Repeated measurements on experimental unit: Different Treatments
c. Repeated measurements on experimental unit: Longitudinal or Spatial
d. Subsampling of experimental unit

II. Specific Terms Used to Describe Designed Experiment:

1. **Experimental Unit:** Entity to which treatments are randomly assigned
2. **Measurement Unit:** Entity on which measurement or observation is made (often the experimental units and measurement units are identical)
3. **Homogeneous Experimental Unit:** Units that are as uniform as possible on all characteristics that could affect the response
4. **Block:** Group of homogeneous experimental units
5. **Factor:** A controllable experimental variable that is thought to influence the response
6. **Level:** Specific value of a factor
7. **Experimental Region (Factor Space):** All possible factor-level combinations for which experimentation is possible
8. **Treatment:** A specific combination of factor levels
9. **Replication:** Observations on two or more units which have been randomly assigned to the same treatment

10. **Subsampling:** Multiple measurements (either longitudinally or spatially) on the same experimental unit under the same treatment
11. **Response:** Outcome or result of an experiment
12. **Effect:** Change in the average response between two factor-level combination or between two experimental conditions
13. **Interaction:** Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors
14. **Confounding:** One or more effects that cannot unambiguously be attributed to a single factor or interaction
15. **Covariate:** An uncontrollable variable that influences the response but is unaffected by any other experimental factors

EXAMPLE

A semi-conductor manufacturer is having problems with scratching on their silicon wafers. They propose applying a protective coating to the wafers, however, the wafer engineers are concerned about the diminished performance of the wafer. An experiment is designed to evaluate several types and thicknesses of coatings on the conductivity of the wafer. Two types of coatings and three thicknesses of the coating are selected for experimentation. A random sample of 72 wafers are selected for use in the experiment with 12 wafers randomly assigned to each combination of a type of coating (C_1, C_2) and a thickness of coating (T_1, T_2, T_3). Only 24 wafers can be evaluated on a given day. Thus, the engineers each day test 4 wafers under each of the coating types-thicknesses combinations. On each wafer, the conductivity is recorded before and after applying the coating to the wafer. Furthermore, to assess the variability in conductivity across the wafer surface, conductivity readings are taken at five locations on each wafer.

- Designed Experiment Consists of Three Components:

  C1. Method of Randomization:

  C2. Treatment Structure:

  C3. Measurement Structure:

OTHER POSSIBLE WAYS OF CONDUCTING THE WAFER EXPERIMENT

Scenario I: All 72 wafers are evaluated in the same day. Each of the 6 treatments $((C_i, T_j), i = 1, 2; j = 1, 2, 3)$ is randomly assigned to 12 wafers. The conductivity readings are all done in the same lab under essentially identical conditions.

Scenario II: Only 24 wafers are evaluated on the same day (3 days to complete the experiment). On each of the three days, 4 wafers are randomly assigned to each of the 6 treatments $((C_i, T_j), i = 1, 2; j = 1, 2, 3)$. The conductivity readings are all done in the same lab under essentially identical conditions.

Scenario III: Only 6 wafers can be evaluated on the same day. Thus to reduce the time to complete the experiment, 6 different labs are used. Two wafers are randomly assigned to each of the 6 treatments. The randomization is such that each treatment appears in every Day-Lab combination.

Scenario IV: A new machine used to apply the coating to the wafers has recently been purchased. This machine requires a considerable amount of time in order to change from applying coating type C_1 to C_2 but almost no set-up time for changing from one thickness to another thickness. Therefore, the engineers want to apply all three thicknesses of coating C_1 and then apply all three thicknesses of coating C_2 rather than doing the applications in a random fashion. This will save them considerable amount of set-up time. Furthermore, only 24 wafers can be coated in a given day and only 1 lab is available for the experiment. Therefore, the following randomization was conducted. On a given day, 12 wafers were randomly assigned to each of the two coatings. Then, 4 of these 12 wafers were randomly assigned to each of the three thicknesses. The randomization was repeated on each of the three days needed to complete the experiment.

COMMON PROBLEMS IN EXPERIMENTAL DESIGNS

I. Masking of Factor Effects

When the variation in the responses are as large as the differences in the treatment means, the treatment differences will not be detected in the experiment. For example, σ_ϵ is large relative to $\mu_i - \mu_{i'}$ in a completely randomized design. In this situation, the experiment must be redesigned by

1. Increasing the sample sizes to reduce
$$\text{StDev}(\hat{\mu}_i - \hat{\mu}_{i'}) = \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_{i'}^2}{n_{i'}}}$$
2. Blocking the experimental units to reduce the size of σ_i's
3. Using Covariates
4. All the above

II. Uncontrolled Factors

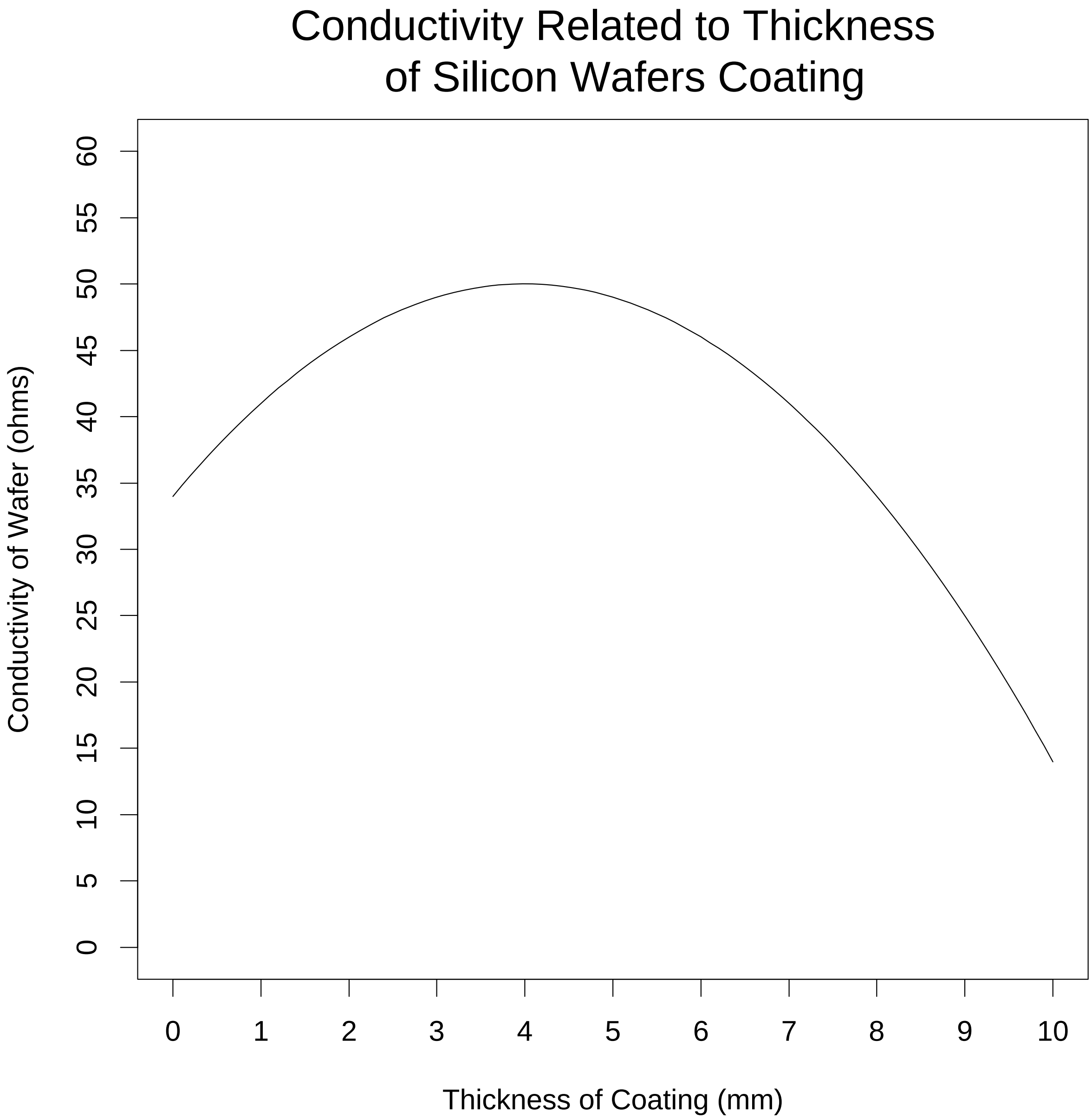
If factors are known to have an effect on the response variable, then these factors should be included in the experiment as either treatment or blocking variables. Failure to carefully consider all factors of importance can greatly compromise the extent to which conclusions can be drawn from the experimental outcomes.

1. Differences between experimental plots in terms of soil fertility, drainage, exposure to sun, exclusion of wildlife, etc.
2. Position of experimental units on greenhouse benches
3. Position of experimental units on trays or in ovens
4. Time of day or week in which experiment is run

III. Erroneous Principles of Efficiency

If the time to run experiments or the cost to run experiments place restrictions on the number of factors and the number of levels of the factors that can be included in the experiment, then the overall goals of the experiment must be reevaluated since

1. Important factors may be ignored or left uncontrolled
2. Non-linear effects may not be determined since the number of levels may be too few or not broad enough to detect higher order effects.


Conductivity Related to Thickness
of Silicon Wafers Coating
Conductivity of Wafer (ohms)
0
5
10
15
20
25
30
35
40
45
50
55
60
0
1
2
3
4
5
6
7
8
9
10
Thickness of Coating (mm)

SELECTING AN APPROPRIATE EXPERIMENTAL DESIGN

I. Consideration of Objectives

1. Nature of anticipated results helps to determine what factors need to be included in the experiment:

   Suppose experiment is designed to determine which of 6 fuel blends used in automobiles produce the lowest CO emissions. The 6 blends include a standard commercial gasoline and 5 different methanol blends. After determining that blend number 5 has the lowest CO emission, the question arises what properties of the blends (distillation temperature, specific gravity, oxygen content, etc.) made the major contributions to the reduced CO level in emissions using the selected blend. A problem that may arise is that the fuel properties may be confounded across the 5 blends and it may not be possible to sort them out with the given experimental runs. This problem could have been avoided if this question was raised prior to running the experiments.

2. Definition of concepts (Can the goals of the experiment be achieved) :

   Suppose we want to study the effects of radiation exposure on the life length of humans

   - Design 1: Subject randomly selected homogeneous groups of humans to various levels of radiation (unethical experiment)
   - Design 2: Use laboratory rats in place of humans (extrapolation problem)
   - Design 3: Use observational or historical data on groups that were exposed to radiation
     (Many uncontrolled factors, genetic differences, amount of exposure, length of exposure, occupational differences, daily habits)

3. Determination of observable variables

   What covariates should be observed? How often? How accurately should they be measured?

II. Factor Effects

1. Inclusion of all relevant factors avoids uncontrolled systematic variation.
2. Need to measure all important covariates to control heterogeneity of experimental units or conditions.
3. Anticipated interrelationships between factor levels helps to determine type of design:

   a. No interactions between factor levels: Use simple screening design

   b. Interactions exist: Need full factorial design

   c. Higher order relationships between factor levels may require a greater number of levels of the factors in order to be able to fit high order polynomials to the responses.

4. Include a broad enough range of the factor levels so as not to miss important factor effects, include lowest and highest feasible values of factor.

III. Precision - Efficiency of Experiment

Degree of variability in response variable determines the number of replications required to obtain desired widths of confidence intervals and power of statistical tests. Determine variability through pilot studies or review literature for results from similar experiments.

IV. Randomization

In order to protect against unknown sources of biases and to be able to conduct valid statistical procedures:

1. The experimental units **MUST** be randomly assigned to the treatments or
2. The experimental units **MUST** be randomly selected from the treatment populations and
3. The time order in which experiments are run and/or spatial positioning of experimental units must be randomly assigned to the various treatments. This avoids the confounding of uncontrolled factor effects with the experimental factors. For example, drifts in instrumental readings, variation across the day in terms of temperature gradients, humidity or sunlight exposure, variation in performance of laboratory technicians (grad students), or various other conditions in the laboratory or field.

DESIGNING FOR QUALITY: INDUSTRIAL PROCESSES

Two Basic Types of Experiments

1. On-Line: Running experiments while process is in full production.

   EVOP - Evolutionary Operation

   Design strategy where 2 or more factors in an on-going production process are varied in order to determine an optimal operation level.

   Problem: Examining very narrow region of the factor space since only small deviations from *normal operations* are allowed by the company.
2. Off-Line: Running experiments in Laboratories or Pilot Plants

Two Basic Goals in Experiments Involving Quality Improvement

1. Bring product On Target

   Average measurement of product characteristic are equal to the target value

2. Uniformity - Consistency

   Measured product characteristics have a small variability about the target value

Combining both of these criterions, we obtain

Minimize MSE = $(Bias)^2 + (StDev)^2$ = (Distance to Target)^2+Variance

Taguchi Approach:

1. Emphasized the importance of using fractional factorial designs
2. His choice of designs were often highly inefficient
3. His analyses of experiments were often incorrect
4. He was successful in convincing engineers at large corporations to use designed experiments. The experiments were very successful even though there were not the best possible experiments that could have been run.