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

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)

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

- **Experimental Study**: Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher)

  Examples: Clinical Trials (Some control), Laboratory Studies (More Control), Agricultural Field Trials (Some Control), Greenhouse Experiments (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

- **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

- **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.

  Example: 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: *e. coli* outbreak in Europe during summer 2011.

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

  Exposure to risk factors
- 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. (In place of disease, we could observe which subjects commit a crime or become PhD's in statistics.)

    Is this study 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.

    Is this study Prospective or Retrospective?

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 in the sample estimates. These are systematic 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

3. Methods of selection sample bias: Random digit-dialing in telephone surveys exclude those households which only have cell phones. Also, many people screen their phone calls and only answer the phone when the call is from a person they know using called ID.

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?

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 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 can not be recorded

   A field of corn may be partially consumed by a herd of deer so that total yield can not 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.

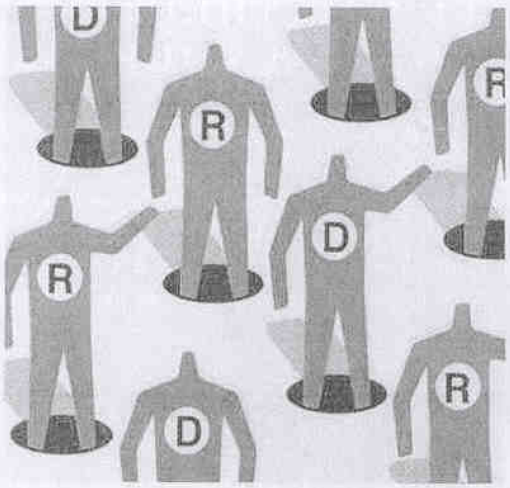
The following two articles from *Newsweek* and the *Houston Chronicle* illustrate possible sources of variation in polling.

News Articles / Polling - Newsweek - October 6, 2008

The Slippery Art of Polling

IF QUANTITY FOSTERED QUALITY, THIS YEAR'S POLLS ASKING "MCCAIN OR Obama?" would be the best in history: polls have proliferated so crazily that Mark Blumenthal, who analyzes them at pollster.com, writes in a typical post, "Another day, another 37 new statewide polls." The profusion isn't surprising. It reflects the tightness of the race, the increased emphasis on state rather than national polls as the Electoral College looms large, and a race so fluid that pollsters are bound and determined to capture every shift. What *is* surprising is that, after a primary season in which polls got several big races wrong (Clinton. New Hampshire. Enough said), pollsters are breaking ranks on some key aspects of methodology and admitting that, this year, old truths about polling are in tatters.

No controversy looms larger than that about "likely" voters. Polls that count them, rather than registered voters, are usually more accurate, but maybe not this year. Pollsters determine who is likely to cast a ballot by asking questions such as whether they voted in 2004, are following the campaign and plan to vote. Respondents get a point for each "right" answer. The pollster then takes a percentage of the sample equal to the percentage of adults who voted (60 percent in '04) and includes only that percentage of scorers. The bottom 40 percent don't count as "likely." This can produce inaccurate results in many ways. For one thing, if 70 percent rather than 60 percent of voters cast a ballot this time, pollsters may be ignoring 14 percent of their sample. For another, because the millions of newly registered voters score zero on the "did you vote last time?" question, they might not count as likely. But if they do turn out on Nov. 4, polls are failing to count a group that could determine the outcome in several states. "The problem with the likely-voter model is that if you have a transformative election, you will have a different turnout," says Nate Silver, who uses polls as input for his forecasting model at fivethirtyeight.com. "My hunch is that there will be more and different voters this time," putting polls on shakier ground.



Cell-phone use might also skew polls. In standard polling methodology, a computer program selects landline numbers at random; laws prohibit random-digit dialing to cell phones. Pollsters have to call cell phones manually, which is time-consuming and expensive. Cell-only voters—now 13 percent of the population, and mostly under 30—are therefore absent from polls. Does that skew results? When ABC News called cell numbers manually, the impact was "negligible," it reported this month: a gain of 0.7 points for Obama and a loss of 0.8 for McCain among likely voters. Those are all smaller than the poll's margin of error. Gallup and the Pew Research Center find that Obama gains no more than 2 to 3 points when they call cell phones.

Why isn't the effect larger? Because of one of the key arts of polling: demographic weighting. If the reported results are, say, 60-40 in favor of McCain, it doesn't mean that 60 percent said they favored the Republican while 40 percent favored Obama. Instead, the pollster counts how many respondents came from each gender, age group, ethnicity, region, even education level. Then the pollster weights the answers. If random dialing yielded 350 men out of 1,000 respondents, and there are 480 men per 1,000 registered voters, then each man's McCain-Obama choice is up-weighted by a factor of 1.37 (480/350). That's how pollsters adjust for any undercount of young cell-only voters: they up-weight the responses of young voters reached on landlines. This works unless young cell-only voters are politically different from young landline users, as a new study from Pew suggests. Of the under-30 cell-only population, 62 percent were Democrats and 28 percent Republicans, Pew finds, and they preferred Obama by 35 points, almost triple the number of under-30 landline users.

Young cell-phone-only users prefer Obama by 35 points, almost triple the spread among landline users.

Few things infuriate critics more than polls that have "too many" respondents of a particular party. The Gallup organization regularly catches flak for supposedly overpolling Republicans, for instance. But few pollsters adjust for party self-identification as they do for gender, age and other fixed categories. The reason is that party ID is surprisingly fluid. In a 2005 study, Larry Hugick of Princeton Survey Research Associates International (who polls for NEWSWEEK) found that 18 percent of voters reported a different party ID in October 2004 than they had the month before. It may seem that a poll with fewer self-identified Democrats than the percentage of total voters in the last election who were Democrats, or with less than the percentage of total voters who are registered as Democrats, has undersampled them. In fact, however, the party that respondents say they belong to may reflect their candidate preference. In this case, Dems moving to McCain change their party ID as well and say they're Republicans. Down-weighting GOP responses because a poll has "too many" Republicans would therefore underestimate McCain's support.

Oh, and something else about polls. Just because they don't agree with the Election Day tally doesn't mean they were incorrect at the time they were done. If the election were held a week later, then news events, shifting mood and voters' most recent exposure to a candidate and his message could easily yield a different result than on Nov. 4.

7

ILLUSTRATION BY HARRY CAMPBELL FOR NEWSWEEK

TUESDAY
SEPTEMBER 12, 2006
HOUSTON CHRONICLE

Polls find Perry loss isn't out of the question

■ But governor's spokesman says methodology used in surveys makes them unreliable

By R.G. RATCLIFFE
AUSTIN BUREAU

AUSTIN — Two polls released Monday found Gov. Rick Perry is vulnerable to defeat, but his campaign is questioning the surveys' accuracy.

Conventional wisdom in the governor's race has been that none of the governor's four opponents would have a chance to beat him if he gets more than 35 percent of the vote on Nov. 7. There is no runoff in the general election, so the top vote-getter wins.

Perry has hovered between 35 percent and 41 percent in public polls for months. But he has fallen into the defeatable zone in polls done by Rasmussen Reports and the Wall Street Journal/Zogby Online.

The Rasmussen poll put Perry's re-election support at 33 percent, and the Zogby poll had his support at 31 percent.

"It's hard to see him losing above 37 percent, but below 35 percent, somebody might get that much out of the remaining 65 percent of the vote," said University of Houston pollster Richard Murray, a Democrat.

Please see **POLLS,** *Page B5*

PERRY AND THE PACK

Two recent polls suggest Gov. Rick Perry's re-election may not be a sure thing:

Wall Street Journal / Zogby Interactive

Gov. Rick Perry (R)	Chris Bell (D)	Kinky Friedman (I)
30.7%	25.3%	22.4%

Carole Keeton Strayhorn (I)	James Werner (L)	Don't know/other
11%	2.6%	8%

Online survey conducted Aug. 29-Sept. 5. Margin of error ± 2.9 percentage points.

Rasmussen Reports

Perry	Bell	Friedman
33%	18%	16%

Strayhorn	Don't know/other
22%	11%

Survey of likely voters conducted Aug. 30. Margin of error ± 4.5 percentage points.

CHRONICLE

CONTINUED FROM PAGE B1

But Perry spokesman Robert Black said the polls paid for by news media companies should not be trusted because their methodology is not sound political science.

"All these media polls that promote the pollster should be taken with a grain of salt," said Black. "If the campaigns, all the campaigns in this race for governor, thought these polls were worth a darn, they wouldn't have their own pollsters."

The polls differed on the challengers to Perry. Rasmussen had independent Comptroller Carole Keeton Strayhorn in second place at 22 percent; Democrat Chris Bell in third place at 18 percent; and independent Kinky Friedman in fourth at 16 percent.

Zogby had Bell in second with 25 percent; Friedman at 22 percent, Strayhorn at 11 percent and Libertarian James Werner at 2.6 percent.

Both polls were conducted days before Perry and Strayhorn began their television advertising last week.

Black said the two polls do not accurately reflect the voting public in Texas. He said Rasmussen uses automated polling methods, while Zogby uses a pool of people who volunteer to be interviewed on an Internet site.

Surveys questioned

Mark Sanders, a spokesman for Strayhorn, touted the results of the Rasmussen poll while challenging the methodology of Zogby.

"Zogby has always been wildly off when it comes to her (Strayhorn)," Sanders said. "The only polls we trust are the ones we take, and they show us in second place."

Bell spokeswoman Heather Guntert said the campaign is not going to be one that "lives or dies" by any poll, but she said Zogby showed growing strength for Bell. She said that also is reflected in a $250,000 donation the campaign received from trial lawyer Harold Nix.

Murray said there are reasons to believe the Rasmussen and Zogby polls are credible and reasons to question their accuracy.

Murray said polls done with robotic interviewers such as those conducted by Rasmussen and SurveyUSA are considered to be "cheap, quick and dirty." But he said they do so much polling that inaccuracies tend to level out over time. He said the hard thing for those polls to reflect is whether they are accurately sampling people who will actually vote.

Murray said Zogby's polling methods called elections very accurately in 2000, but less so in 2004.

May not mean much

Charles Franklin, a political scientist at the University of Wisconsin and a founder of Pollster.com, said the biggest problem with Internet-based polling like Zogby's is that the people who sign up to be surveyed are very interested in politics.

Even if Zogby weights the population of the poll sample to reflect the population of the state, Franklin said a political survey of volunteers is likely one of people who already had strong feelings about politics and the candidates.

Franklin said there can be variability in any poll. So Perry's drop may not mean much, especially if nothing happened to cause it, he said.

"Perry's movement right now isn't large enough for me to be convinced that Perry has dropped significantly," Franklin said.

Friedman not worried

The one campaign not worried by the polls is that of Friedman, who is building his campaign like the one used by former wrestler Jesse Ventura to win the governorship in Minnesota.

"Jesse Ventura had 11 to 15 percent in the polls at this point in the campaign. He didn't break 20 percent until the weekend before the election," said Friedman spokeswoman Laura Stromberg. "This is great news."

r.g.ratcliffe@chron.com

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

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.*

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
- Use all the parts in a single container
- 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
- 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"
- Meat inspector includes in sample meat packages that the inspector "thinks" may be contaminated with *e. coli*

SIMPLE RANDOM SAMPLING (SRS)

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 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
```

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

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

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 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?

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 identifer on each of the woody seedlings under these 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?

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 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 strata 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

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

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

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

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

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

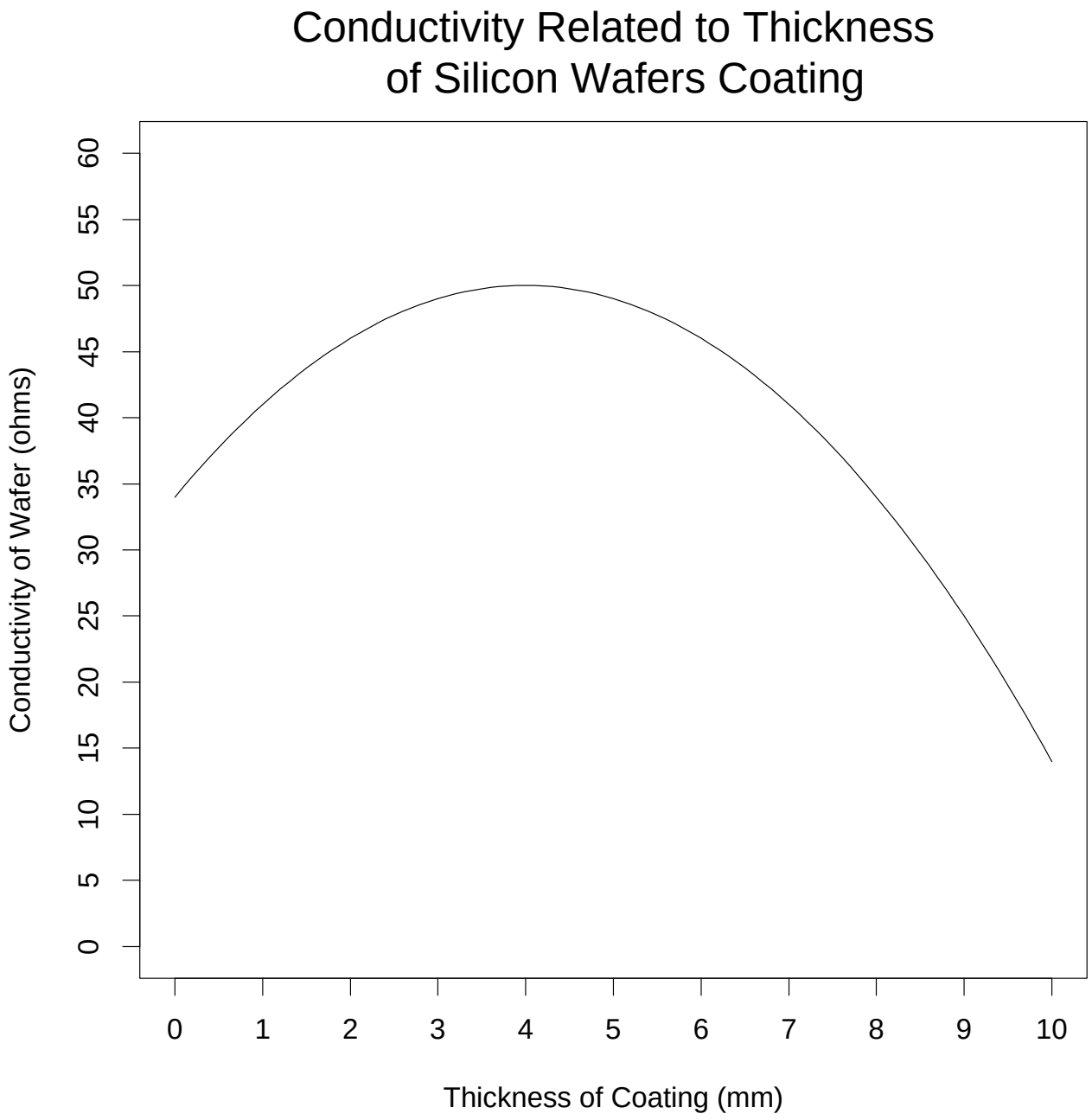
If factors are known to have an effect on the response variable, then these factors should be included in the experiment as either treatment or blocking variables. Failure to carefully consider all factors of importance can greatly compromise the extent to which conclusions can be drawn from the experimental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