"You can't fix by analysis what you bungled by design."

–Light, Singer and Willett

Or, not as catchy but perhaps more accurate:

Fancy analysis can't fix a poorly designed study.

Producing Data

The Role of Statistics in Producing and Analyzing Data

A **study** is an examination of a subject for the purpose of advancing knowledge.

Many studies require analysis of data. Data used in a study can arise in a number of ways. **Available data** are data that were obtained prior to the study for purposes other than those of the study.

Statistically designed studies obtain data using a pre-specified plan that ensures that specific questions of interest are answered in a statistically valid way.

The Role of Statistics in Producing and Analyzing Data

Drawing definitive conclusions from available data is often questionable. However, **exploratory analysis** of available data is a vital part of scientific and technological progress.

As the name suggests, in an exploratory analysis investigators explore the data to find patterns that might suggest specific questions about the phenomena under study.

Exploratory analysis cannot establish scientific validity, however. For this, statistically designed studies are needed. A good strategy is to use exploratory data analysis to identify specific questions that can then be answered by targeted statistically designed studies.

The Role of Statistics in Producing and Analyzing Data

For example, a manufacturing company might use exploratory analysis of historical data (that is, available data) to identify process conditions associated with the production of faulty components. Once they have narrowed the search to a few suspects, they might design and conduct a study to test or confirm the results of the initial exploratory analysis.

Characteristics and Conduct of Designed Studies

The two main types of designed studies are

- Controlled Experiments
- Observational Studies

Both use **observational units**: entities on which measurements or observations can be made.

These observatonal units are often selected from a larger population of such units. When discussing their selection, we call them **sampling units**.

Characteristics and Conduct of Designed Studies

Proper selection of the sampling units is essential to the success of designed studies, because it will enable us to

- Get a sample that with high probability is representative of the population.
- Quantify how far from the population results our sample results are likely to be.

Selecting Sampling Units

- **Target Population:** A collection of sampling units about which we want to draw conclusions
- **Frame:** A list of all sampling units in the target population
- **Sample:** A subset of the target population from which conclusions about the target population will be drawn

Characteristics and Conduct of Designed Studies

- **Sampling Design:** A pattern, arrangement or method used for selecting a sample of sampling units from the target population
- **Sampling Plan:** The operational plan, including the sampling design, for actually obtaining or accessing the sampling units for the study

Example 1

Suppose, prior to an election, we want to estimate the outcome in the "population" of WPI students. To do so, we select 20 WPI students and interview them. For this study:

- Sampling units: Individual students
- Target population: All WPI students
- Frame: Campus directory
- Sample: The 20 selected students

Characteristics and Conduct of Designed Studies

- Sampling Design: There are a large number of choices, as we will see below
- Sampling Plan: Operational plan for deciding whom to interview, how to get them to do the interview, what to do if they can't be found, or won't talk, etc.

Reasons to Sample

- Lower Cost
- Shorter Time
- Little Loss of Precision
- The Only Choice (e.g. Destructive Testing)

Probability Sampling Methods

- Simple Random Sampling Each possible sample has the same chance of selection. Good if units are homogeneous and easily accessed.
- Stratified Random Sampling Sampling units are divided into distinct strata, and a simple random sample taken separately in each. Good if units in each stratum are much more homogeneous than the population as a whole or if we want to include certain subgroups in the sample.
- Cluster Sampling Units in close proximity are grouped in clusters, and clusters are sampled. Good if there are large costs due to units being widely dispersed.

Characteristics and Conduct of Designed Studies

In addition, many samples are taken in stages, using any of the above methods at any stage. Such sampling is called Multistage Sampling.

One large study conducted monthly by the US Census bureau is the Current Population Survey, which estimates such things as unemployment and schooling by taking a multistage cluster sample of 100,000 people in 60,000 households.

Example 2

Suppose you want to estimate the average amount spent by first term sophomores at WPI for textbooks, and that you can interview 10 students for your study. How would you choose the 10 students if:

Characteristics and Conduct of Designed Studies

(a) You believe the distribution of the amounts spent for textbooks is pretty consistent across all students. Simple random sample is best.

(b) You believe that textbook expenses for engineering students are substantially higher than for other majors. Stratify into two strata: engineers and others. Then take a stratified random sample.

(c) You want to be certain to obtain an estimate for humanities majors, as well as other majors. Stratify so that one stratum is humanities. Then take a stratified random sample.

Errors in Selecting Sampling Units

- Sampling Error Error inherent in the sampling process. Sampling error occurs normally (i.e., is not really an error), is always present and results from the fact that the sample is not the same as the population.
- Nonsampling Errors These are really errors and are the result of (i) being unable to sample from the entire population, (ii) being unable to get measurements from some selected units, or (iii) getting misleading or false measurements from some selected units.

One type of nonsampling error is **Selection Bias**:

In statistics, a bias is a distortion of the results of a statistical procedure.

Selection bias is a bias introduced to the results because the sampling method to at least some extent misses certain segments of the population.

Some Possible Errors in Selecting Sampling Units for the WPI Election Survey

- Sampling Error: The extent to which the sample is unrepresentative of the population, by chance alone.
- Selection Bias: Selection bias might occur if the survey were conducted in class. In that case, the survey would miss those who don't come to class. This would result in a bias if the views of those who come to class and of those who do not come to class differ.

Note:

If the sampling units are selected by some non-probability method (convenience, for example), the results of the study are, strictly speaking, only applicable to the sampling units in the study. In order to have results of the study apply to the target population, sampling units must be selected from that target population using an appropriate probability sampling design.

Types of Designed Studies

- Controlled Experiment
- Observational Study

Controlled Experiments

Before we can discuss controlled experiments, we need some terminology:

- Experimental Unit: A sampling unit selected for use in a controlled experiment.
- Response: A measurement or observation of interest that is made on an experimental unit.

Controlled Experiments

- Factor: A quantity that is thought to influence the response.
  - Experimental Factor: A factor that is purposely varied by the experimenter.
  - Nuisance Factor: A factor that cannot be controlled by the experimenter. Nuisance factors may or may not be known to the experimenter.

Controlled Experiments

- Level: A value assumed by a factor in an experiment.
- Treatments: The combinations of levels of experimental factors for which the response will be observed.

Controlled Experiments

We are now ready to define **Controlled Experiment**:

A **Controlled Experiment** is a study in which treatments are imposed on experimental units in order to observe a response.

Example 3

A printing company is having trouble with ink overflow on printed documents. After much brainstorming and discussion, they have narrowed the possible causes to two printing machine settings: the pressure plate setting and the ink flow rate setting. They design and run a controlled experiment to evaluate the effect of these settings on the finished product. They decide on three settings, low, medium and high, for the pressure plate and two settings, low and high, for the ink flow rate. The response is the improperly inked area on a test sheet. The data are (in $(\text{mm})^2$):

Controlled Experiments

Ink flow setting	Pressure plate setting		
	Low	Medium	High
Low	25.300 21.600	27.100 24.200	19.700 21.900
High	31.100 29.500	25.600 23.100	26.600 23.900

Controlled Experiments

For this experiment:

- Experimental units: paper sheets.
- Response: improperly inked area.
- Experimental factors: pressure plate setting, ink flow setting.
- Nuisance factors: variation in paper characteristics, environmental factors (temperature, humidity, etc.), ink supplier, etc.
- Factor levels: low, medium, high for pressure plate; low, high for ink flow rate.
- Treatments: pressure plate & ink flow rate combinations.

Controlled Experiments

This is a controlled experiment because treatments (pressure plate & ink flow rate combinations) are imposed on experimental units (paper sheets) in order to observe a response (improperly inked area).

Controlled Experiments

Here are some further quantities and concepts of importance in controlled experiments:

o **Effect:** The change in the average response between two factor levels or between two combinations of factor levels. Here are some effects from example 3:

Controlled Experiments

Ink flow setting	Pressure plate setting			Row Mean
	Low	Medium	High	
Low	25.300 21.600	27.100 24.200	19.700 21.900	
Mean	23.450	25.650	20.800	23.300
High	31.100 29.500	25.600 23.100	26.600 23.900	
Mean	30.300	24.350	25.250	26.633
Column Mean	26.875	25.000	23.025	24.967

The effect of high ink flow over low ink flow is
$26.633 - 23.300 = 3.333$.

Controlled Experiments

Ink flow setting	Pressure plate setting			Row Mean
	Low	Medium	High	
Low	25.300 21.600	27.100 24.200	19.700 21.900	
Mean	23.450	25.650	20.800	23.300
High	31.100 29.500	25.600 23.100	26.600 23.900	
Mean	30.300	24.350	25.250	26.633
Column Mean	26.875	25.000	23.025	24.967

The effect of high pressure plate setting over medium pressure plate setting is $23.025 - 26.875 = -3.850$.

Controlled Experiments

Ink flow setting	Pressure plate setting			Row Mean
	Low	Medium	High	
Low	25.300 21.600	27.100 24.200	19.700 21.900	
Mean	23.450	25.650	20.800	23.300
High	31.100 29.500	25.600 23.100	26.600 23.900	
Mean	30.300	24.350	25.250	26.633
Column Mean	26.875	25.000	23.025	24.967

The effect of low ink flow and high pressure plate settings over high ink flow and medium pressure plate settings is $20.800 - 24.350 = -3.55$.

Controlled Experiments

o **Confounding:** Two or more factors are **confounded** if it is impossible to separate their individual effects.

Here is an example of how confounding might occur in the ink flow experiment:

Suppose that the experimenters ran the trials for the low and high ink flow settings with different kinds of paper, and that the kind of paper used makes a real difference in the response. Then paper and ink flow would be confounded.

Randomized Controlled Experiments

By helping ensure that the experimental units receiving different treatments are similar in all other respects, random assignment of treatments to experimental units protects against unsuspected biases. Random assignment is beneficial in at least two other ways.

Randomized Controlled Experiments

- Random assignment of treatments to experimental units provides the foundation for **statistical inference**, on which many scientific applications of statistics depend. We will study statistical inference in chapters 5 and 6.
- The most compelling reason to conduct a controlled experiment is to establish **causality**: that is, to show that changing the treatment will cause a change in the response. In order to establish causality, treatments must be assigned to experimental units in some random fashion.

Controlled experiments in which treatments are assigned at random to experimental units are called **randomized controlled experiments**.

Assigning Treatments to Experimental Units: Two Commonly-Used Designs

- **Completely Randomized Design (CRD)** Treatments assigned to experimental units completely at random. Works well if units are homogeneous.
- **Randomized Complete Block Design (RCBD)** Experimental units grouped into blocks and all treatments are assigned at random within each block. Effective if units within each block are more homogeneous than all units taken as a whole.

Two Different Ways to Design the Ink Overflow Experiment

Suppose there are two printing machines on which the experiment is to be run.

- For a completely randomized design, assign all 12 treatments to the two machines at random. So, for example, machine 1 might get both treatments at high ink flow setting and high pressure plate setting. Of course the order of the runs will also be randomized. This kind of design makes sense if the machines act very much the same.

Controlled Experiments

- For a randomized complete block design, assign one complete set of six treatments to machine 1 and the other to machine 2. Each machine is a block. Run each set of treatments in random order. This kind of design makes sense if there is substantial machine-to-machine variation.

More on Blocking

Here are two more examples of blocking:

- 25 pleasure boats around the country are available to test two types of marine paint. Make each boat a block by applying both types of paint to each. This (a) reduces boat-to-boat variation and variation due to such things as environment, and (b) makes the results of the study applicable to a wider range of environments and boat types.

Controlled Experiments

- An alloy manufacturer produces aluminum ingots in four furnaces. Each furnace is known to have its own unique operating characteristics, so "furnace" will be a nuisance variable for any experiment run in the foundary that involves more than one furnace. In an experiment to assess the effect of stirring rate on the grain size of the product, the four furnaces are used as blocks: each of four stirring rates (the experimental factor) is assigned in random order to each furnace. This allows comparisons to be made within each furnace and the results extrapolated across furnaces.

Principles of Experimental Design

- Block What You Can.
- Randomize What You Cannot Block.
- Replicate as Time and Budget Permit.
  Replication=repetition. Beware of duplication.
- Confirm the Results.

Principles of Experimental Design in the Ink Overflow Experiment

- Block What You Can. The experiment was run as an RCBD with two different machines as blocks and each treatment run on each machine.
- Randomize What You Cannot Block. Order of run was randomized in each machine (block) separately.

Controlled Experiments

- Replicate as Time and Budget Permit. Only 1 rep was done (1 observation for each treatment and each machine) due to the limited availability of the machines.
- Confirm the Results. To verify the results, confirmatory experiments were later run.

Experimenting With Human Subjects

- **Treatment Group** Group of subjects that receives a treatment.
- **Control Group** Group of subjects that receives no treatment or a neutral treatment.
- **Placebo** Neutral "treatment" given to subjects in the control group.
- **Double-Blind** Neither subject nor evaluator(s) know which treatment (if any) was given.

Example: Salk Vaccine Field Trial, p. 108 of the text.

Steps for Planning Experiments

Successful experiments must be carefully planned. Here are some steps that should be followed in planning an experiment.

1. Decide the objectives of the experiment. These should be unbiased, measurable, and of practical consequence.
2. Obtain relevant background on responses and factors. Where does this experiment fit into the study of the process or system?
3. Decide each response variable that will be measured.

Controlled Experiments

4. Select the factors that are to be systematically varied.
5. Decide each factor to be "held constant" in the experiment.
6. Determine which nuisance factor(s) might affect the response.
7. List and label known or suspected interactions among factors.
8. List restrictions on the experiment.
9. Give current design preferences, if any.
10. If possible, propose analysis and presentation techniques.
11. Determine who will be responsible for the coordination of the experiment.
12. Decide whether trial runs will be conducted.

Observational Studies

Outside of controlled experiments, observational studies are the main class of designed studies. We will study three types:

- **Cohort Study (aka Prospective Study)**
- **Case-Referent Study (aka Retrospective Study)**
- **Sample Survey**

Example 1, Continued:

The study mentioned earlier in which we selected 20 WPI students to see how they would vote is a sample survey. It is not a controlled experiment since no treatments are assigned to experimental units.

Example 4

Another example of an observational study is a study published in the **Journal of the American Medical Association**, which assessed the health patterns of 5,000 Canadians, and found that those with the greatest folic acid intake had 68% less fatal coronary disease than those with the lowest intake.

This study is not a controlled experiment since it merely observed the folic acid intake and coronary death outcomes, rather than assigning a folic acid regimen to individuals.

Cohort (aka Prospective) Studies

Also known as quasi-experiments, because they are controlled experiment "wannabees", cohort studies lack the ability to control the assignment of treatments to experimental units (e.g., we cannot assign a human subject a certain number of cigarettes per day).

As a result, cohort studies (and, in fact, all observational studies) can only demonstrate association, not cause-effect, between treatments and responses.

Observational Studies

In a cohort study, "treatment" and "control" groups (known as **cohorts**) are established based on the hypothesized cause. The pattern of response (i.e., the effect) is then compared for the groups.

In a study of the relation between smoking and lung cancer, the cohorts might be smokers and non-smokers, and the response might be whether or not the subject develops lung cancer.

Example 4, Continued:

If the Canadian folic acid study assigned individuals to groups based on their reported folic acid intake, followed them for a period of time to observe the incidence of fatal coronary heart disease, and compared the groups, then it was a cohort study.

Case-Referent (aka Retrospective) Studies

Case-Referent studies are particularly useful when

- The time between the hypothesized cause and observed effect is large, or
- The effect occurs rarely.

Observational Studies

In a case-referent study, groups are formed based on the response (i.e., effect) and and patterns in the hypothesized causes are compared from group to group.

In a study of the relation between smoking and lung cancer, we might form two groups: those who contracted lung cancer and those who did not. We might then compare the percentages of smokers in the two groups.

Example 4, Continued

If the Canadian folic acid study classified individuals into groups based on whether or not they died of coronary disease, and compared the intake of folic acid for the two groups, then it was a case-referent study.

Caution

Some students get the mistaken idea that if the study was done in the past, or done using data taken in the past, it is case-referent. Case-referent refers only to the fact that the groups are formed based on the outcome and then differences in potential causal factors are sought. The case group is the group that gets the positive outcome (e.g., cancer) and the referent group the group that gets the negative outcome (e.g., no cancer).

Observational Studies

While it is true that outcomes must be observed prior to analyzing a case-referent study, studies based on data taken in the past can also be cohort studies if groups are established based on characteristics observed prior to observing the outcomes.

Observational Studies

For example, suppose the researchers conducted the Canadian folic acid study by randomly selecting 5000 patient records which included folic acid intake, and whether they suffered fatal coronary disease by the end of a 5 year period. Even though the data were taken in the past, if they formed comparison groups based on folic acid intake and compared outcomes, the study is a cohort study.

Cause-Effect

Only a properly-designed and conducted controlled experiment can establish a cause-effect relationship between factors and response.

The biggest difference between controlled experiments and observational studies, such as cohort and case-referent studies, that attempt to show cause-effect, is the idea of **control**: the ability of the experimenter to assign treatments to experimental units. It is the control in controlled experiments that validates cause-effect conclusions and the lack of control in observational studies that casts doubt on cause-effect conclusions.

Observational Studies

Observational studies lack the control (the ability to assign treatments to experimental units) of controlled experiments. This means they also lack the ability to reduce variation due to nuisance variables by blocking.

However, they can reduce variation due to nuisance variables by

- Stratifying by nuisance variables.
- Adjusting responses for values of nuisance variables.

(Note that both these methods can be used to lessen the impact of nuisance variables in controlled experiments as well.)

Sample Surveys:

- Use a sample of sampling units obtained from a population to obtain information about the whole population.
- Have as their primary goals description of various aspects of the population from which the sample is obtained, or comparison of subgroups from that population (not establishment of association).

Observational Studies

The study described in Example 1, in which a sample of 20 WPI students was interviewed to estimate the outcome of an election is an example of a survey.

Another is the Current Population Survey, conducted monthly by the US Census Bureau. The CPS questions about 100,000 people in some 60,000 households nationwide, and uses the results to estimate measures of the state of the nation such as income, unemployment, and schooling.

Non-sampling Errors in Studies of Human Populations

In addition to sampling error and selection bias, studies of human populations in which individuals are asked to respond to questions, verbally or in writing, are subject to other nonsampling errors, such as

- **Nonresponse bias:** Bias due to failure to obtain responses from some subjects.
- **Response bias:** Bias due to erroneous responses from some subjects.

Example 4, Continued

Here are some possible non-sampling errors in the Canadian Folic Acid Study:

- **Nonresponse bias** might occur if certain individuals refuse to supply their medical histories.
- **Response bias** might occur if the subject doesn't tell the truth about his or her health history because of concerns that giving information about poor health will adversely affect future insurability.

Steps in Designing Observational Studies

- Determine what information is required.
- Design the sampling plan.
- Decide how the data are to be obtained.

Observational Studies

- Establish procedures to reduce nonsampling errors.
  - Reduce **nonresponse bias** by
    - Keeping questionnaires short
    - Building-in incentives
    - Doing follow-ups

    In addition, monitoring differences in known demographic variables for respondents and nonrespondents will help assess the severity of the nonresponse problem.
  - Reduce **response bias** by careful design, testing and modification, and training.

Observational Studies

- Always, always, always **Do a pilot study.**

Recap:

- The role of statistics in producing and analyzing data
- Selecting sampling units
  - Sampling designs
  - Sampling errors
  - Designing sampling plans

- Controlled experiments
  - Principles of experimental design
  - Experimenting with human subjects
  - Steps for planning an experiment
- Observational studies
  - Types of observational studies: cohort studies, case-referent studies and sample surveys
  - Steps for planning an observational study
- Cause and effect in designed studies