HANDOUT # 1 - INTRODUCTION TO STATISTICS

TOPICS

1. Definiton of Statistics
2. Statistics and the Scientific Method
3. Research Process
4. Why Study Statistics?
5. Some Current Applications of Statistics
6. Preparation of Data
7. Guidelines for a Statistical Analysis and Report
8. Examples of Studies/Experiments

Some of material in Handout 1 is from *An Introduction to Statistical Methods and Data Analysis, 6th Ed.* by R. Lyman Ott and Michael Longnecker.

Statistics and the Scientific Method

Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision-making and scientific discovery when the available information is both limited and variable. That is, statistics is the science of *Learning from Data*. A description of the early impact of statistics on solving problems in science can be found in the book, *The Lady Tasting Tea, How Statistics Revolutionized Science in the Twentieth Century* by David Salsburg. A second book, *The Theory That Would Not Die* by Sharon Betsch McGrayne discusses the major impact that Bayesian Analysis had on solving problems in industry, government, military, and science.

Almost everyone-including corporate presidents, marketing representatives, social scientists, engineers, medical researchers, and consumers- deals with data. These data could be in the form of quarterly sales figures, percent increase in juvenile creme, contamination levels in water samples, survival rates for patients undergoing medical therapy, census figures, failure rates of newly modified production equipment, or information that assists a consumer in selecting which brand of car to purchase.

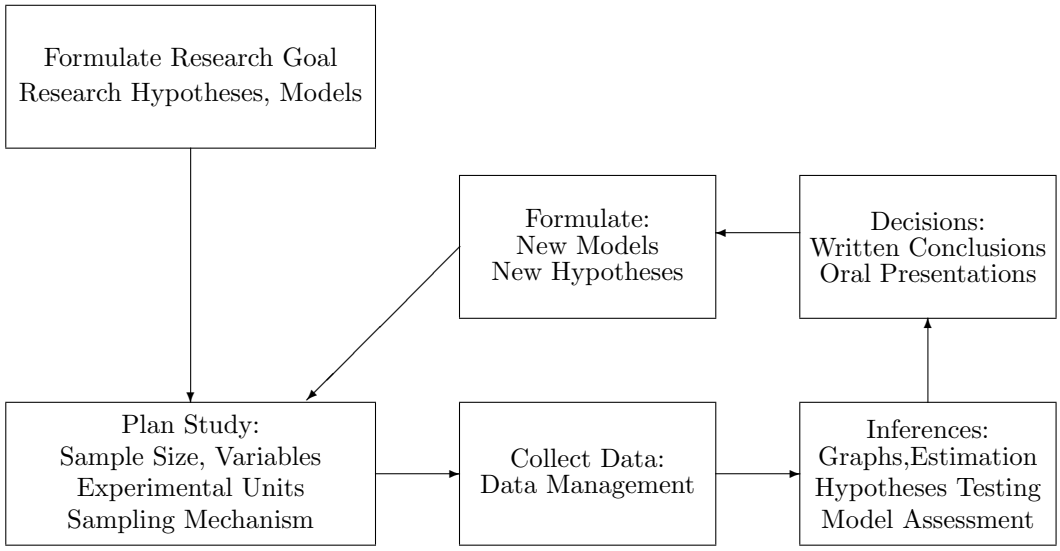
In this course, we will approach the study of statistics by considering a process by which we Learn from Data:

1. Defining the problem
2. Collecting the data
3. Summarizing the data
4. Analyzing/modeling the data
5. Interpreting the analyses/models
6. Communicating the results obtained from the analyses/models

The Learn from Data process described above closely parallels the Scientific Method, which is a set of principles and procedures used by successful scientists in their pursuit of knowledge. The method involves the formulation of research goals, the modeling/analyzing of the data in the context of research goeal, and the testing of hypotheses. The conclusions of these steps is often the formulation of new research goals for another study. These steps are illustrated in the schematic given on the next page.

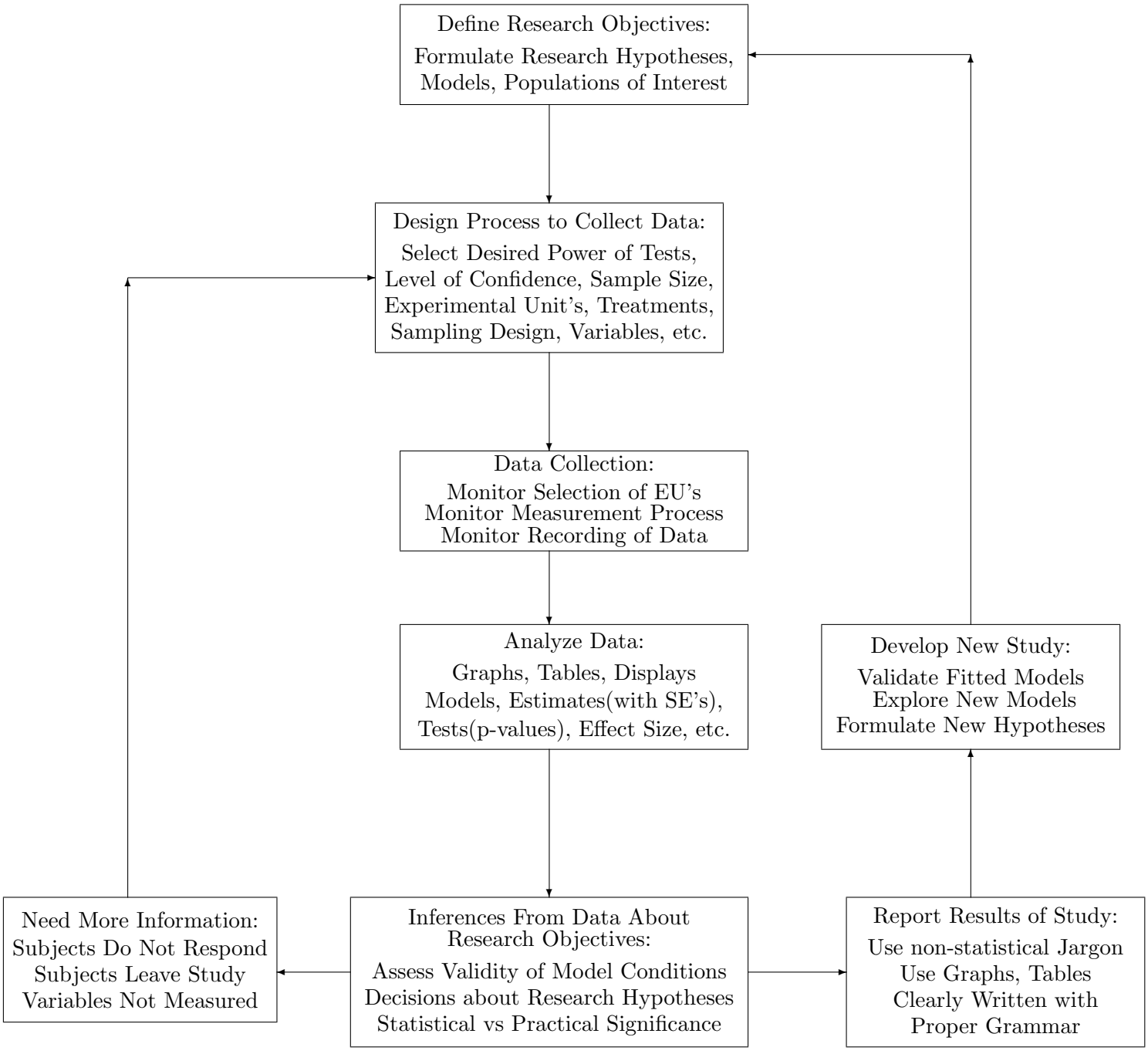
The design of experiments, collection of data, and analysis of data are integral components of the **Scientific Method**. Researchers use the results of studies and experiments to examine the validity of existing theories, to revise these theories, and eventually to formulate new theories. When the observed data contradicts existing theories, the researcher attempts to formulate new theories which explains the discrepancies between the observed data and what the existing theories would have predicted for the data. When the research studies are in a new area without existing theories, the methods of exploratory data analysis often provides the researcher with insights which will enable the researcher to formulate theories to govern the phenomenon under study. A summary of the steps in the scientific method are provided in Figure 1 on the next page.

Figure 1: The Scientific Method



A somewhat more detailed schematic is given on the next page. This depiction of the Research Process illustrates made of the decisions that must be made during a complex research study

Figure 2: Research Process



To illustrate some of the above concepts, we will consider several situations in which the process of Learning from Data could assist in solving a real-world problem.

1. **Problem: Monitoring the ongoing quality of a light bulb manufacturing facility.**

   A light bulb manufacturing produces approximately half a million bulbs per day. The quality assurance department must monitor the defective rate of the bulbs. It could accomplish this task by testing each bulb, but the cost would be substantial and would greatly increase the price per bulb. An alternative approach is to select 1,000 bulbs from the daily production of 500,000 bulbs and test each of the 1,000. The fraction of defective bulbs in the 1,000 tested could be used to estimate the fraction defective in the entire day's production, provided that the 1,000 bulbs were selected in the proper fashion. We will demonstrate in a subsequent handout that the fraction defective in the tested bulbs will with a high probability be quite close to the fraction defective for the entire day's production of 500,000 bulbs.

2. **Problem: Is there a relationship between quitting smoking and gaining weight.**

   To investigate the claim that people who quit smoking often experience a subsequent weight gain, researchers selected a random sample of 400 participants who had successfully participated in programs to quit smoking. The individuals were weighed at the beginning of the program and again one year later. The average change in weight of the participants was an increase of 5 pounds. The investigators concluded that there was evidence that the claim was valid. We will develop techniques in later handouts to assess when changes are truly significant changes and not changes due to random chance.

3. **Problem: What effect does nitrogen fertilizer have on wheat production?**

   For a study of the effects of varying amounts of nitrogen fertilizer on the amount of wheat produced, thirty 10-acre fields were available to the researcher. The same variety of wheat was planted in all 30 fields. She then randomly assigned six fields to each of the five nitrogen rates under investigation. The fields were cultivated in the same fashion until harvest, and the pounds of wheat produced from each of the 30 fields was recorded. The researcher wanted to determine the optimal level of nitrogen to apply to *any* wheat field, but, of course, she was limited to running experiments on a limited number of fields. After determining the amount of nitrogen that yielded the largest production of wheat in the study fields, the researcher then concluded that similar results would hold for wheat fields possessing characters somewhat the same as the study fields. Is the experimenter justified in reaching this conclusion?

4. **Problem: Determining public opinion concerning a question, issue, product, or political candidate.**

   Similar applications of statistics are brought to mind by the frequent use of the *New York Times/CBS News, Washington Post/ABC News, CNN, Harris, and Gallup* polls. How can these pollsters determine the opinions of the more than 200 million Americans of voting age? They certainly do not contact every potential voter in the United States. Rather, they sample the opinions of a small number of potential voters, often as few as 1,500, to estimate the reaction of every person of voting age in the country. The amazing result of this process is that if the selection of the voters is done in an unbiased manner and voters are asked unambiguous, nonleading questions, the fraction of those persons contacted who hold a particular opinion will closely match the fraction in the total population holding that opinion at a particular time. Convincing supportive evidence of this assertion will be presented in a later handout.

These problems illustrate the process of Learning from Data. First, there was a problem of question to be addressed. Next, for each problem a study or experiment was proposed to collect meaningful data to answer the problem. The quality assurance department had to decide both how many bulbs needed to be tested and how to select the sample of 1,000 bulbs from the total production of bulbs to obtain valid results. The polling organizations must decide how many voters to sample and how to select these individuals in order to obtain information that is representative of the population of all voters. Similarly, it was necessary to carefully plan how many participants in the weight-gain study were needed and how they were to be selected from the list of all participants. Furthermore, what variables should the researchers have measured on each participant? Was is necessary to know each participants age, sex, physical fitness, and other health-related variables, or was weight the only important characteristic? The results of the study may not be relevant to the general population if a large number of the participants in the study had a particular health condition or were of a particular ethnic group. In the wheat experiment, it was important to measure both the soil characteristics of the 30 fields and the environmental conditions, such as temperature and rainfall, to obtain results that could be generalized to fields not included in the study. The design of a study or experiment is crucial to obtaining results that can be generalized beyond the participants in the study.

Finally, having collected, summarized, and analyzed the data, it is important to report the results in unambiguous terms to interested people. For the lightbulb example, management and technical staff would need to know the quality of their production batches. Based on this information, they could determine whether adjustments in the process are necessary. Therefore, the results of the statistical analyses cannot be presented in ambiguous terms; decisions must be made from a well-defined knowledge base. The results of the weight-gain study would be of vital interest to physicians who have patients participating in the smoke-cessation program. If a significant increase in weight was recorded for those individuals who had quit smoking, physicians may have to recommend diets so that the former smokers would not from one health problem (smoking) to another (elevated blood pressure due to being overweight). It is crucial that a careful description of the participants-that is, age, sex, and other health-related information-be included in the report. In the wheat study, the experiment would provide wheat growers with information that would allow them to economically select the optimum amount of nitrogen required for their wheat fields to achieve maximum production. Therefore, the study report must contain information concerning the amount of moisture and types of soils present in the study fields. Otherwise, the conclusions about optimal wheat production may not pertain to farmers growing wheat under considerably different conditions.

To infer validly that the results of a study are applicable to a larger group than just the participants in the study, we must carefully define the **population** to which inferences are sought and design a study in which the **sample** has been appropriately selected from the designated population.

DEFINITION: A *population* is the set of all measurements of interest to the researcher collecting data.

DEFINITION: A *sample* is any subset of the measurements selected from the population.

Why Study Statistics?

One of the many reasons to study statistics is that you to know how to evaluate published numerical facts. Every person is exposed to manufacturer's claims for products; to the results of sociological, consumer, and political polls; and to the published results of scientific research. Many of these results are inferences based on sampling. Some inferences are valid; others are invalid. Some of these results are inferences based on samples of adequate size; others are not. Yet all these published results bear the ring of truth. Some people (particular statisticians) say that statistics can be made to support almost anything. Others say it is easy to lie with statistics. Both statements are true. It is easy, purposely or unwittingly, to distort the truth using statistics when presenting the results of sampling to the uninformed. It is thus crucial that you become an informed and critical reader of data-based reports and articles.

A second reason for studying statistics is that your profession or employment may require you to interpret the results of sampling (surveys or experimentation) or to employ statistical methods of analysis to make inferences in your work. For example, practicing physicians receive large amounts of advertising describing the benefits of new drugs. These advertisements frequently display the numerical results of experiments that compare a new drug with an older drug. Do such data really imply that the new drug is more effective, or is the observed differences in results due simply to random variation in the experimental measurements.

Recent trends in the conduct of court trials indicate an increasing use of probability and statistical inference in evaluating the quality of evidence. The use of statistics in the social, biological, and physical sciences is essential because all these sciences make use of observations of natural phenomena, through sample surveys or experimentation, to develop and new theories. Statistical methods are employed in business when sample data are used to forecast sales and profit. In addition, they are used in engineering and manufacturing to monitor product quality. The sampling of accounts is a useful tool to assist accountants in conducting audits. Thus, statistics plays an important role in almost all areas of science, business, and industry; persons employed in these areas need to know the basic concepts, strengths, and limitations of statistics.

The information about careers in statistics can be found at the website amstat.org/careers:

Statisticians are in high demand in a wide variety of fields. As the largest professional association for statisticians in the world, the ASA serves as the main clearinghouse for information about jobs, careers, and employment for the statistical profession. There is an abundance of information at the website, including the following topics:

1. What is statistics?
2. What do Statisticians do?
3. Which Industries Employ Statisticians?
4. How do I Become a Statistician?

For only $15 per year, you as a student can be a member of ASA. Information on joining is given at the website: **www.amstat.org**

Common Misconceptions and Confusions Found in Research Articles

The article, "What Educated Citizens Should Know About Statistics and Probability", by Jessica Utts, in *The American Statistician*, May 2003, contains a number of statistical ideas that need to be understood by users of statistical methodology in order to avoid confusion in the use of their research findings. Misunderstandings of statistical results can lead to major errors by government policymakers, medical workers, and consumers of this information. The article selected a number of topics for discussion. We will summarize some of the findings in the article. A complete discussion of all these topics will be given throughout this course.

1. One of the most frequent misinterpretations of statistical findings is when a statistically significant relationship is established between two variables and it is then concluded that a change in the explanatory variable *causes* a change in the response variable. As will be discussed throughout this course, this conclusion can be reached only under very restrictive constraints on the experimental setting. Utts examined a recent *Newsweek* article discussing the relationship between the strength of religious beliefs and physical healing. Utts' article discussed the problems in reaching the conclusion that the stronger a patient's religious beliefs, the more likely patients would be cured of their ailment. Utts shows that there are numerous other factors involved in a patient's health, and the conclusions that religious beliefs **cause** a cure can not be validly reached.

2. A common confusion in many studies is the difference between *(statistically) significant* findings in a study and *(practically) significant* findings. This problem often occurs when large data sets are involved in a study or experiment. This type of problem will be discussed in detail throughout this course. We will use a number of examples that will illustrate how this type of confusion can be avoided by the researcher when reporting the findings of their experimental results. Utts' article illustrated this problem with a discussion of a study that found a statistically significant difference in the average heights of military recruits born in the spring and in the fall. There were 507,125 recruits in the study and difference in average height was about $\frac{1}{4}$ inch. So, even though there may be a difference in the actual height of recruits in the spring and the fall, the difference is so small ($\frac{1}{4}$ inch) that it is of no practical importance.

3. The size of the sample also may be a determining factor in studies in which statistical significance is *not* found. A study may not have selected a sample size large enough to discover a difference between the several populations under study. In many government-sponsored studies, the researchers do not receive funding unless they are able to demonstrate that the sample sizes selected for their study are of an appropriate size to detect specified differences in populations if in fact these differences exist. Methods to determine appropriate sample sizes will be provided in the handouts on hypotheses testing.

4. Surveys are ubiquitous, especially during the years in which presidential elections are held. In fact, market surveys are nearly as widespread as political polls. There are many sources of bias that can creep into the most reliable of surveys. The manner in which people are selected for inclusion in the survey, the way in which questions are phrased, and even the manner in which questions are posed to the subject may affect the conclusions obtained from the survey. We will discuss these issues in Handout 2.

5. Many students find the topic of probability to be very confusing. One of these confusions conditional probability where the probability of an event occurring is computed under the condition that a second event has occurred with certainty. For example, a new diagnostic test for the pathogen, *E. coli* in meat is proposed to the U.S. Department of Agriculture (USDA). The USDA evaluates the test and determines that the test has both a low *false positive* and a low *false negative* rate. That is, it is very unlikely that the test will declare the meat contains *E. coli* when in fact it does not contain *E. coli*. Also, it is very unlikely that the test will declare the meat does not contain *E. coli* when in fact it does contain *E. coli*. Although the diagnostic test has a very low false positive rate and a very low false negative test, the probability that *E.coli* is in fact present in the meat when the test produces a positive test result is *very* low for those situations in which a particular strain of *E. coli* occurs very infrequently. In Handout 13, we will demonstrate how conditional probability concepts can be applied to this type of situation to produce a true assessment of the performance of a diagnostic test.

6. Another concept that is often misunderstood is the role of the degree of variability in interpreting what is a "normal" occurrence of some naturally occurring event. Utts' article provided the following example. A company was having an odor problem with its waste water treatment plant. They attributed the problem to "abnormal" rainfall during the period in which the odor problem was occurring. A company official stated the facility experienced 170% to 180% of its "normal" rainfall during this period, which resulted in the water in the holding ponds taking longer to exit for irrigation. Thus, there was more time for the pond to develop an odor. The company official did not point out that yearly rainfall in this region is extremely variable. In fact, the historical range for rainfall is between 6.1 and 37.4 inches with a median rainfall of 16.7 inches. The rainfall for the year of the odor problem was 29.7 inches, which was well within the "normal" range for rainfall in this area. There is a confusion between the terms "average" and "normal" rainfall. The concept of natural variability is crucial to correct interpretation of statistical results. In this example, the company official should have evaluated the percentile for an annual rainfall of 29.7 inches in order to demonstrate the abnormalities of such a rainfall. We will discuss the ideas of data summary and percentiles in Handout 6.

The types of problems expressed above and in Utts' article represent common and important misunderstandings that can occur when researchers use statistics in interpreting the results from their studies. We will attempt throughout this course to discuss possible misinterpretations of statistical results and how to avoid them in your data analyses. Furthermore, it is hoped that after completing this course, you will be a discriminating reader of statistical findings, the results of surveys, and project reports.

Some Current Applications of Statistics

Reducing the Threat of Acid Rain

The accepted causes of acid rain are sulfuric and nitric acids; the sources of these acidic components of rain are hydrocarbon fuels, which spew sulfur and nitric oxide into the atmosphere when burned. Here are some of the many effects of acid rain:

- Acid rain, when present in spring snow melts, invades breeding areas for many fish, which prevents successful reproduction. forms of life that depend on ponds and lakes contaminated by acid rain begin to disappear.
- In forests, acid rain is blamed for weakening some varieties of trees, making them more susceptible to insect damage and disease.
- In areas surrounded by affected bodies of water, vital nutrients are leached from the soil.
- Man-made structures are also affected by acid rain. Experts from the U.S. estimate the acid rain has caused nearly $15 billion of damages to buildings and other structures by the beginning of this century. The problem continues.

Solutions to the problems associated with acid rain will not be easy. The National Science Foundation (NSF) has recommended that the U.S. strive for a 50% reduction in sulfur-oxide emissions. Perhaps that is easier said than done. High sulfur coal is a major source of these emissions, but in states dependent on coal fired power plants, a shift to lower sulfur coal is not always possible. Instead, devices must be developed to remove these contaminating oxides from the burning process before they are released into the atmosphere. Fuels for internal combustion engines are also major sources of the nitric and sulfur oxides of acid rain. Clearly, better emission control is needed for motor vehicles.

Reducing the oxide emissions from coal burning power plants and motor vehicles will require greater use of existing emission control devices and the development of new technology. Of course, the major goal should be the development of cleaner energy sources which eliminate the use of carbon based fuels. Statisticians will play a key role in monitoring atmospheric conditions, testing the effectiveness of proposed emission control devices, and developing alternative energy sources.

Determining the Effectiveness of a New Drug Product

The development and testing of the Salk vaccine for protection against poliomyelitis (polio) provide an excellent example of how statistics can be used in solving practical problems. Most parents and children growing up before 1954 can recall the panic brought on by the outbreak of polio cases during the summer months. Although relatively few children fell victim to the disease each year, the pattern of outbreak of polio was unpredictable and caused great concern because of the possibility of paralysis or death. The fact that very few of today's youth have even heard of polio demonstrates the great success of the vaccine and the testing program that preceded its release on the market.

It is standard practice in establishing the effectiveness of a particular drug product to conduct an experiment, clinical trial, with human participants. For some clinical trials, assignments of participants are made at random, with half of the subjects receiving the new drug product and the half receiving a solution or tablet that does not contain the medication (called a *placebo*). One statistical problem concerns determining the number of participants to be included in the clinical trial. This problem was particularly important in the testing of the Salk vaccine because data from previous years suggested that the incidence rate for polio might be less than 50 cases for every 100,000 children. Hence, a large number of participants had to be included in the clinical trial in order to detect a difference in the incidence rates for those treated with the vaccine and those receiving the placebo.

With the assistance of statisticians, it was decided that a total of 400,000 children should be included in the Salk clinical trial begun in 1954. No other clinical trial had ever been attempted on such a large group of participants. Through a public school inoculation program, the 400,000 participants were treated and then observed over the summer to determine the number of children contracting polio. Although fewer than 200 cases of polio were reported for the 400,000 children in the clinical trial, more than three times as many cases appeared in the group receiving the placebo. These results, together with some statistical calculations, were sufficient to indicate the effectiveness of the Salk polio vaccine. However, these conclusions would not have been possible if the statisticians and scientists had not planned for and conducted such a large clinical trial.

The development of the Salk vaccine is not an isolated example of the use of statistics in the testing and developing of drug products and medical devices. In recent years, the Food and Drug Administration (FDA) has placed stringent requirements on pharmaceutical firms to establish the effectiveness of proposed new drug products and medical devices. Thus, statistics has played an important role in the development of birth control devices, rubella vaccines, chemotherapeutic agents in the treatment of cancer, and the investigation of gene based treatments of various devices.

Use and Interpretation of Scientific Data in the Courts

Libel suits related to consumer products have touched each one of us; you may have been involved as a plaintiff or defendant in a suit or you may know of someone who was involved in such litigation. We all help to fund the costs of this litigation indirectly through insurance premiums and increase costs of products. The testimony in civil suits concerning salary discrimination, drug product, medical suit, and so on, frequently leans heavily on the interpretation of data from one or more scientific studies. This is how and why statistics and statisticians have become involved in the courtroom.

For example, epidemiologists have used statistical concepts applied to data to determine whether there is a statistical "association" between a specific characteristic, such as the leakage i silicone breast implants, and a disease condition, such as an autoimmune disease. An epidemiologist who finds an association should try to determine to determine whether the observed statistical association from the study is due to random variation or whether it reflects an actual association between the characteristics and the disease. Courtroom arguments about the interpretations of these types of associations involve data analyses using statistical methodologies as well as a clinical interpretation of the data. Many examples exist in which models are used in court cases. In salary discrimination suits, a lawsuit is filed claiming that an employer underpays employees based on the employees' age, ethnicity, or sex. Statistical models are developed to explain salary differences based on many factors, such as work experience, years of education, and work performance. The adjusted salaries are then compared across age groups or ethnic groups to determine whether significant salary differences remain after adjusting for the relevant work performance factors.

Estimating Bowhead Whale Population Size

Raftery and Zeh (1998) discuss the estimation of the population size and rate of increase in bowhead whales, *Balaena mysticetus*. The importance of such a study derives from the fact that bowheads were the first species of great whale for which commercial whaling was stopped; thus, their status may indicate the recovery prospects of other great whales. Also, the International Whaling Commission uses these estimates to determine the aboriginal subsistence whaling quota for Alaskan Native Americans. To obtain the necessary data, researchers conducted a visual and acoustic census of Point Barrow, Alaska. The researchers then applied statistical models and estimation techniques to the data obtained in the census to determine whether the bowhead population had increased or decreased since commercial whaling was stopped. The statistical estimates demonstrated that the bowhead population was increasing at a healthy rate, indicating that stocks of great whales that have been decimated by commercial hunting can recover after hunting is discontinued.

Ozone Exposure and Population Density

Ambient ozone pollution in urban areas is one of the nation's most pervasive environmental problems. Whereas the decreasing stratospheric ozone layer may lead to increasing increases of skin cancer, high ambient ozone intensity has been shown to cause damage to the human respiratory system as well as to agriculture crops and trees. The Houston, Texas area has ozone concentrations rated second only to Los Angeles that exceed the National Ambient Air Quality Standard. Carroll et al. (1997) describe how to analyze the hourly ozone measurements collected in Houston from 1980 to 1993 by 9 to 12 monitoring stations. Beside the ozone level, each station also recorded three meteorological variables: temperature, wind speed, and wind direction.

The statistical aspects of the project had three major goals:

1. Provide information (and/or tools to obtain such information) about the amount and pattern of missing data, as well as about the quality of the ozone and the meteorological measurements.

2. Build a model of ozone intensity to predict the ozone concentration at any given location within Houston at any given time between the years 1980 and 1993.

3. Apply this model to estimate exposure indices that account for either a long-term exposure or a short-term high-concentration exposure; also, relate census information to different exposure indices to achieve population exposure indices.

The spatial-temporal model the researchers built provided estimates demonstrating that the highest ozone levels occurred at locations with relatively small populations of young children. Also, the model estimated that the exposure of young children to ozone decreased by approximately 20% from 1980 to 1993. An examination of the distribution of population exposure had several policy implications. In particular, it was concluded that the current placement of monitors is not ideal if one is concerned with assessing population exposure to ozone. This project involved all four components of Learning from Data: planning where the monitoring stations should be placed within the city, how often data should be collected, and what variables should be recorded; conducting spatial-temporal graphing of the data;

creating spatial-temporal models of the ozone data, meteorological data, and demographic data; and finally, writing a report that could assist local and federal officials in formulating policy with respect to decreasing ozone levels.

Assessing Public Opinion

Public opinion, consumer preference, and election polls are commonly used to assess the opinions and preferences of a segment of the public for issues, products, or candidates of interest. The American public are exposed to the results of these polls daily in newspapers, in magazines, on the radio,and on television. For example, the results of polls related to the following subjects were printed in local newspapers over a 2-day period:

- Consumer confidence related to future expectations about the economy
- Preferences for candidates in upcoming elections and caucuses
- Attitudes toward cheating on federal income tax returns
- Preference polls related to specific products (for example, foreign vs American cars, Coke vs Pepsi, McDonald's vs Wendy's)
- Opinions of voters toward proposed changes in social security and medicare
- Reactions of Texas residents toward same sex unions

A number of questions can be raised about polls. Suppose we consider a poll on the public's opinion toward a proposed reduction of funding for public education in Michigan. *What was the population of interest the pollster?* Was the pollster interested in all residents in Michigan or just those citizens who have children of school age? *Was the sample in fact selected from this population?* If the population of interest was all persons who have children of school age, did the pollster make that all the individuals sampled had children of school age? *What questions were asked and how were the questions phrased?* Was each person asked the same question? Were the questions phrased in such a manner as to bias the responses? Can we believe the results of these polls? Do the results "represent" how the general public *currently* feels about the issues raised in the polls?

Opinion and preference polls are an important, visible application of statistics for the consumer. We will discuss this topic in more detail in Handout 2.

Preparing Data for Statistical Analysis

In practice, processing data from data may consume 75% of the total effort from the receipt of the raw data to the presentation of results from the analysis. What are the steps in processing the data, why are they so important, and why are they so time-consuming. To answer these questions, let us begin by listing the major data-processing steps in the cycle, which begin with receipt of the data and end when the statistical analysis begins.

Steps in Preparing Data for Analysis

1. Receiving the raw data source
2. Creating the database from the raw data source
3. Editing the database
4. Correcting and clarifying the raw data source
5. Finalizing the data base
6. Creating data files from the data base

We will discuss each of the six steps.

1. **Receiving the raw data source.** For each study to be summarized and analyzed, the data arrive in some form, which will be referred to as the **raw data source**. For a clinical trial, the raw data source is usually case report forms that have been used to record study data for each patient entered into the study. For other types of studies, the raw data source may be sheets of paper from a laboratory notebook, a thumb drive, hand tabulations, and some other form of electronic memory. It is important to retain the raw data source because it is the beginning of the **data trial**, which leads for the raw data to the conclusions drawn from a study. Many consulting operations involved with the analysis and summarization of many different studies keep a log that contains vital information related to the study and raw data source. General information contained in a study log is provided in the following list.

   Log of Study Data

   1. Data received and from whom
   2. Study investigator
   3. Statistician and others assigned to team
   4. Brief description of study
   5. Treatments (compounds, preparations, etc.) studied
   6. Raw data source
   7. Responses measured
   8. Reference number for study
   9. Estimated completion data
   10. Other pertinent information

Later, when the study has been analyzed and results have been communicated, additional information can be added to the log on how the study results were communicated, where these results are recorded, what data files have been saved, and where these files are stored.

2. **Creating the database from the raw data source.** For most studies that are scheduled for a statistical analysis, a machine-readable database is created. The steps taken to create the database and the eventual form of the database vary from one organization to another, and depending on the software systems to be used in the statistical analysis. When the data are to be *key-entered* from a paper record, the raw data first checked for legibility. Any illegible numbers or letters or other problems should be brought to the attention of the study coordinator. Then a coding guide that assigns column numbers and variable names to the data is filled out. Certain codes for missing values (for example, those data not available) are also defined at this point. Also, it is helpful to gave a brief description of each variable. The data file keyed in at the terminal is referred to as the **machine-readable database.** A listing of the contents of the database should be obtained and checked carefully against the raw data source. Any errors should be corrected at the terminal and verified against an updated data listing.

   Often data are received in a machine-readable form. In these situations, the data file is considered to be the database. You must, however, have a coding guide to "read" the database. Using the coding guide, obtain a listing of the contents of the database and check it *carefully* to see that all numbers and characters look reasonable and that proper formats were used to create the file. Any problems that arise must be resolved before proceeding further.

3. **Editing the database.** The types of edits done and the completeness of the editing process really depend on the type of study and how concerned you are about the accuracy and completeness of the data prior to analysis. For example, it is wise to examine the minimum, maximum, and frequency distribution for variable to make certain that none of the values look unreasonable.

   Certain other checks should be made. Plot the data (scatter plots or box plots) and look for problems. Also, certain **logic checks** should be done, depending on the structure of the data. For example, if data are recorded for patients during several different visits, then the data for Visit 2 cannot be earlier than the data for Visit 1; similarly, if a patient is lost to follow-up after Visit 2, we cannot have any data for that patient at later visits.

4. **Correcting and clarifying the raw data source.** Questions frequently arise concerning the legibility or accuracy of the raw data during any one of the steps from the receipt of the raw data to the communication of the results from the statistical analysis. It is helpful to keep a list of these problems or discrepancies in order to define the data trail for a study. If a correction (or clarification) is required to the raw data source, this should be indicated on the form and the appropriate change made to the raw data source. If no corrections are required, this should be indicated on the form as well. Keep in mind that the machine-readable database should be changed to reflect any changes made to the raw data source.

5. **Finalizing the database.** You may have been led to believe that all data for a study arrive at one time. This, of course, is not always the case. For example, with a marketing survey, different geographic locations may be surveyed at different times, and hence those responsible for data processing do not receive all the data at once. All these subsets of data, however, must

be processed through the cycles required to create, edit, and correct the database. At this time, the database should be reviewed again and final corrections made before beginning the analysis. This is because, large data sets, the analysis and summarization chores take considerable staff and computer time. It is better to agree on a final database analysis than to have to repeat all analyses on a changed database at a later date.

6. **Creating data files from the database.** Generally, one or two sets of data files are created from the machine-readable database. The first set, referred to as **original files**, reflects the basic structure of the database. A listing of the files is checked against the database listing to verify that the variables have been read with correct formats and missing value codes have been retained. For some studies, the original files are actually used for editing the database.

   A second set of data files, called **work files**, may be created from the original files. Work files are designed to facilitate the analysis. They may require restructuring of the original files, a selection of important variables, or the creation or addition of new variables by insertion, computation, or transformation. A listing of the work files is checked against that of the original files to ensure variables are checked by hand calculation to verify program code.

   If original and work files are SAS data sets, you should utilize the documentation features provided by SAS. At the time an SAS data set is created, a descriptive label for the data set of up to 40 characters should be assigned. The label can be stored with the data set, imprinted wherever the contents procedure is used to print the data set's contents. All variables can be given descriptive names, up to 8 characters in length, which are meaningful to those involved in the project. In addition, variable labels up to 40 characters in length can be used to provide additional information. Title statements can be included in SAS code to identify the project and describe each job. For each file, a listing (proc print) and dictionary (proc contents) can be retained.

   Even if appropriate statistical methods are applied to data, the conclusions drawn from the study are only as good as the data on which they are based. So you be the judge. The amount of time spent on these data-processing chores before analysis really depends on the nature of the study, the quality of the raw data source, and how confident you want to be about the completeness and accuracy of the data.

Guidelines for a Statistical Analysis and Report

In this section, we briefly discuss a few guidelines for performing a statistical analysis and list some important elements of a statistical report used to communicate results. The statistical analysis of a large study can usually be broken down into three types of analyses: (1) preliminary analyses; (2) primary analyses; and (3) backup analyses.

The **preliminary analyses**, which are often descriptive or graphic, familiarize the statistician with the data and provide a foundation for all subsequent analyses. These analyses may include frequency distributions, histograms, box plots, descriptive statistics, an examination of comparability of the treatment groups, correlations, or univariate and bivariate plots.

Primary analyses address the objectives of the study and the analyses on which conclusions are drawn. **Backup analyses** include alternative methods of examining the data that confirm the results of the primary analyses; they may also include new statistical methods that are not as readily accepted as the more standard methods. Several guidelines for analyses follow.

Preliminary, Primary, and Backup Analyses

1. Perform the analyses with software that has been extensively tested.
2. Label the computer output to reflect which study is analyzed, what subjects (animals, patients, etc.) are used in the analysis, and a brief description of the analysis preferred.
3. Use variable labels and value labels (for example, = = none, 1 = mild, 2 = severe) on the output.
4. Provide a list of the data used in each analysis.
5. Check the output *carefully* for all analyses. Did the program submission run successfully (check SAS log for errors)? Are the sample sizes, means, and degrees of freedom correct? Other checks may be necessary as well.
6. Save all preliminary, primary, and backup analyses that provide the informational base from which study conclusions are drawn.

After the statistical analysis is completed, conclusions must be drawn and the results communicated to the intended audience. Sometimes it is necessary to communicate these results as a formal written statistical report. A general outline for a statistical report follows.

General Outline for a Statistical Report

1. Summary
2. Introduction
3. Experimental design and study procedures
4. Descriptive statistics
5. Statistical methodology used in analysis
6. Results and conclusions
7. Discussion
8. Data listings (usually contained in an Appendix to report)

Documentation and Storage of Results

The final part of this cycle of data processing, analysis, and summarization concerns the documentation and storage of results. For formal statistical analyses that are subject to careful scrutiny by others, it is important to provide detailed documentation for all data processing and the statistical analyses so the data trail is clear and the database or work files are readily accessible. Then the reviewer can follow what has been done, redo it, or extend the analyses. The elements of a documentation and storage file depend on the particular setting in which you work. The contents for a general documentation storage file are as follows.

Study Documentation and Storage File

1. Statistical report
2. Study description
3. Random code (used to assign subjects to treatment groups)
4. Important correspondence
5. File creation information
6. Preliminary, primary, and backup analyses
7. Raw data source
8. A data management sheet, which includes the log, as well as information on the storage of the data files

The major thrust behind the documentation and storage file is that we want to provide a clear data and analysis "trail" for our own use or for someone else's use, should there be a need to revisit the data. For any given situation, ask yourself whether such documentation is necessary and, if so, how detailed it must be. A good test of the completeness and understandability of your documentation is to as a colleague who is unfamiliar with your project but knowledgeable in your field to try to reconstruct and even redo the primary analyses you did. If he or she can navigate through your documentation trial, you have done the job.

Example of Designed Experiment

The following example is from R.D. Snee (1983), "Graphical Analysis of Process Variation Studies," *Journal of Quality Technology*, **15**, 76-88. In most industrial processes there are numerous sources of variation in the physical characteristics of the product being produced. Frequently studies are conducted to investigate what aspects of the production process are the major causes of the variation. For example, a chemical analysis is performed on the raw materials prior to their injection into the process. The amount of DMZ in the raw materials is to be determined. This analysis involves different specimens of the raw materials and is performed by numerous operators using a combustion furnace.

In order to investigate sources of variability for a chemical analysis, an experiment was designed and analyzed to ensure that relevant sources of variation could be identified and measured. The process engineer and operators discussed the situation and agreed that the four possible major sources of variation are:

1. Operator - O: Variation due to operators systematically differing in the their adherence to the analytic procedures

2. Specimen - S(O): Variation in specimens of raw materials analyzed by the same operator

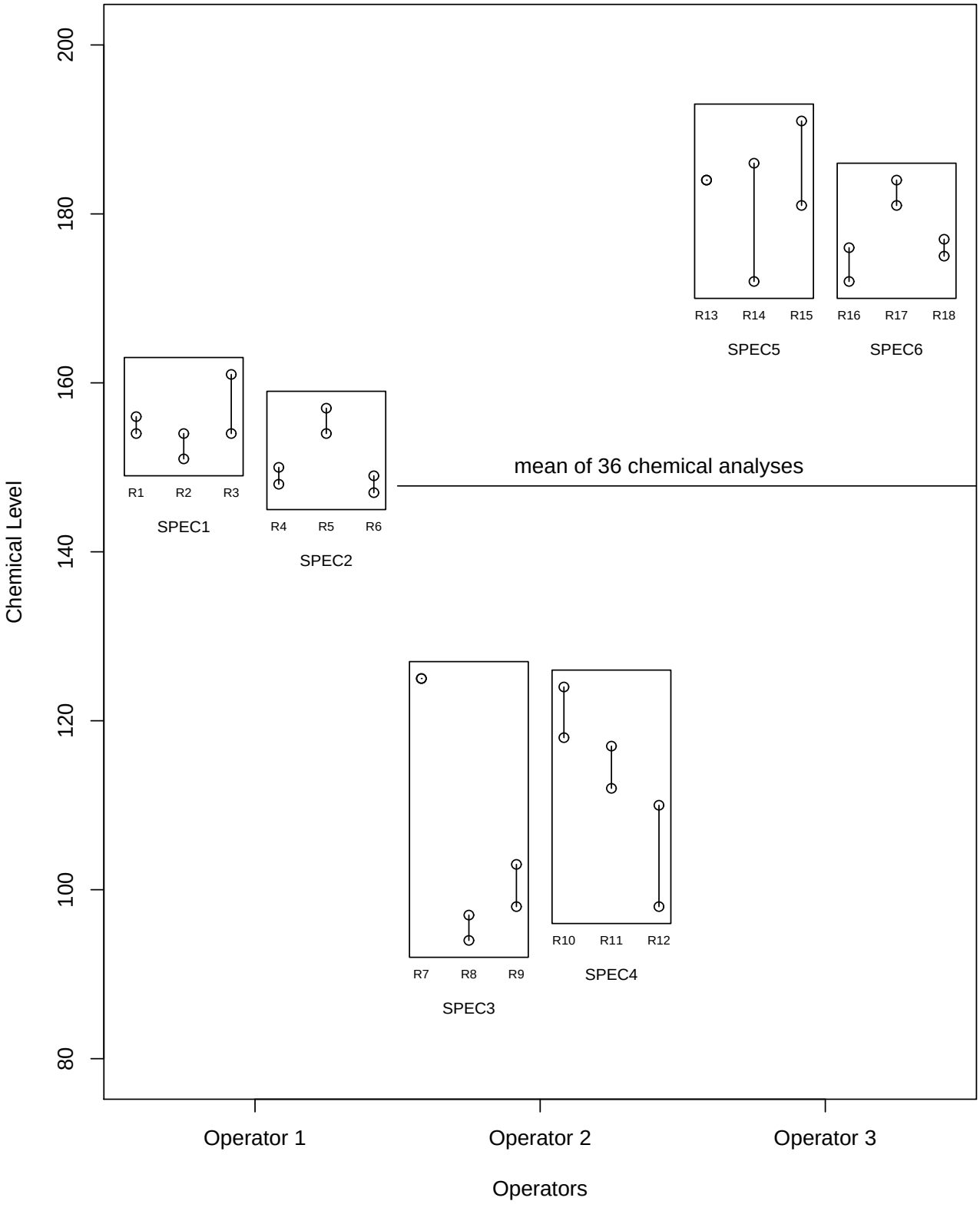
3. Combustion Run - R(S,O): Variation in measurements from run to run in the furnace using the same specimen and operator due to variations in the conditions within the combustion furnace on any given running

4. Chemical Analysis - A(R,S,O): Variation in the measurements of the chemical analysis performed on the material from a fixed combustion run using same specimen and operator due to equipment or procedural variation

The experiment was designed to measure the relative sizes of each of the four potential sources of variation in the amount of DMZ in the material. Three operators were randomly selected to perform the analysis. Each operator analyzed two specimens. Each of the six specimens were split into 3 units. The individual units were then placed in a combustion furnace (Run). After removing the specimens from the combustion furnace, the units were titrated in duplicate (Chemical Analysis). The resulting determinations of DMZ from the experiments are displayed in the following table and figure.

			Chemical Analysis				
Operator	Specimen	Run	1	2	Run Mean	Specimen Mean	Operator Mean
1	1	1	156	154	155	155	152.917
		2	151	154	152.5		
		3	154	161	157.5		
	2	1	148	150	149	150.833	
		2	154	157	155.5		
		3	147	149	148		
2	3	1	125	125	125	107	110.083
		2	94	97	95.9		
		3	98	103	100.5		
	4	1	118	124	121	113.167	
		2	112	117	114.5		
		3	98	110	104		
3	5	1	184	184	184	183	180.25
		2	172	186	179		
		3	181	191	186		
	6	1	172	176	174	177.5	
		2	181	184	182.5		
		3	175	177	176		

Figure 3: Chemical Variation Plot



The Figure 3 highlights a major problem with the chemical analysis procedure. There are definite differences in the analytic results of the three operators.

- Operator 1 exhibits very consistent results for each of the two specimens and each of the three combustion runs.
- Operator 2 produces analytic results which are lower on the average than those of the other two operators.
- Operator 3 shows good consistency between the two specimens, but the repeat analysis of two of the combustion runs on specimen 5 appear to have substantially larger variation than for most of the other repeat analysis in the data set.
- Operator 2 likewise shows good average consistency for the two specimens, but large variations both for the triplicate combustion runs for each specimen and for at least one of the repeat analysis for the fourth specimen.

A statistical analysis (See STAT 642) of the four variance components reveals the following per cent allocation of the total variation in the measurements:

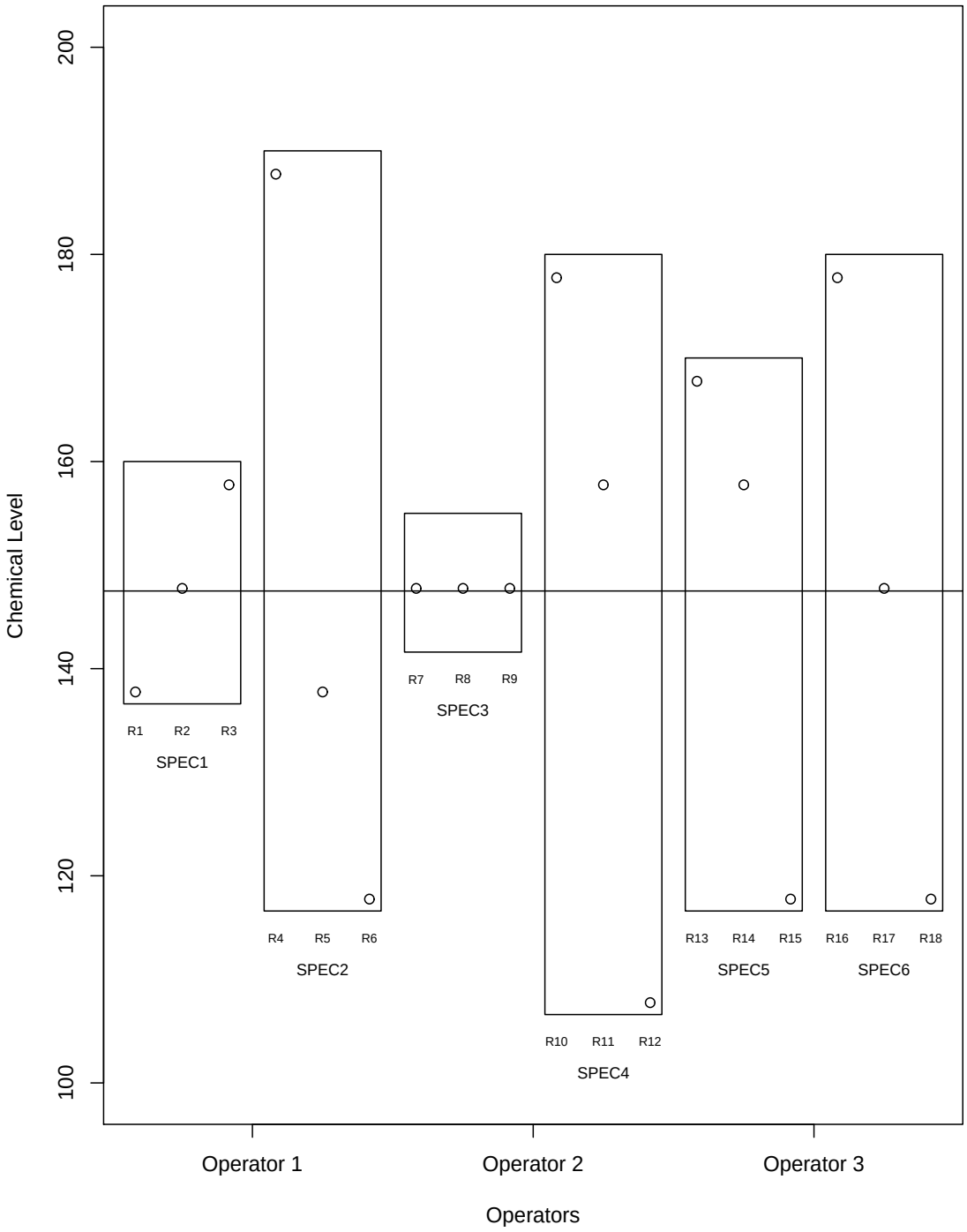
1. 94.80% Operator
2. 0.00% Specimen
3. 3.83% Combustion Run
4. 1.29% Chemical Analysis
5. 0.08% Other Sources

Hypothetical Experiment: A situation in which all all the variation is due to R(S,O), Runs within Operator and Specimen.

1. There is no variation due to Operator: The three Operator Means are equal
2. There is no variation due to Specimen within Operator: The two Specimen means within each Operator are identical
3. There is no variation due to Chemical Analysis within Operator, Specimen, Runs: The two values for the chemical analysis are identical for each selection of a Run, Operator, and Specimen

			Chemical Analysis				
Operator	Specimen	Run	1	2	Run Mean	Specimen Mean	Operator Mean
1	1	1	137.75	137.75	137.75	147.75	147.75
		2	147.75	147.75	147.75		
		3	157.75	157.75	157.75		
	2	4	187.75	187.75	187.75	147.75	
		5	137.75	137.75	137.75		
		6	117.75	117.75	117.75		
2	3	7	147.75	147.75	147.75	147.75	147.75
		8	147.75	147.75	147.75		
		9	147.75	147.75	147.75		
	4	10	177.75	177.75	177.75	147.75	
		11	157.75	157.75	157.75		
		12	107.75	107.75	107.75		
3	5	13	167.75	167.75	167.75	147.75	147.75
		14	157.75	157.75	157.75		
		15	117.75	117.75	117.75		
	6	16	177.75	177.75	177.75	147.75	
		17	147.75	147.75	147.75		
		18	117.75	117.75	117.75		

Figure 4: Variation Due Only to R(S,O)



The following is the R code used to produce the Variation Plot:

```
run = seq(1,18)
Res = c(137.75,147.75,157.75,187.75,137.75,117.75,
147.75,147.75,147.75,177.75,157.75,107.75,
167.75,157.75,117.75,177.75,147.75,117.75)
spec = seq(1,6)
postscript("u:/meth1/lectures/chemplot.ps",horizontal=F)
plot(run,Res,type="p",xlab="Operators",ylab="Chemical Level",
        main="Figure 4: Variation Due Only to R(S,O)",cex=.99,
        ylim=c(100,200),xaxt="n")
rect(0.75,136.6,3.25,160)
text(1,134,"R1",cex=.55)
text(2,134,"R2",cex=.55)
text(3,134,"R3",cex=.55)
text(2,131,"SPEC1",cex=.75)

rect(3.75,116.6,6.25,190)
text(4,114,"R4",cex=.55)
text(5,114,"R5",cex=.55)
text(6,114,"R6",cex=.55)
text(5,111,"SPEC2",cex=.75)

rect(6.75,141.6,9.25,155)
text(7,139,"R7",cex=.55)
text(8,139,"R8",cex=.55)
text(9,139,"R9",cex=.55)
text(8,136,"SPEC3",cex=.75)

rect(9.75,106.6,12.25,180)
text(10,104,"R10",cex=.55)
text(11,104,"R11",cex=.55)
text(12,104,"R12",cex=.55)
text(11,101,"SPEC4",cex=.75)

rect(12.75,116.6,15.25,170)
text(13,114,"R13",cex=.55)
text(14,114,"R14",cex=.55)
text(15,114,"R15",cex=.55)
text(14,111,"SPEC5",cex=.75)


rect(15.75,116.6,18.25,180)
text(16,114,"R16",cex=.55)
text(17,114,"R17",cex=.55)
text(18,114,"R18",cex=.55)
text(17,111,"SPEC6",cex=.75)

axis(side=1,at=c(3.5,9.5,15.5),
labels = c("Operator 1","Operator 2","Operator 3"))
segments(0,147.5,19,147.5)
graphics.off()
```

Studying the relationship Between Variable: The Challenger Disaster

The following example is from Hogg-Ledoleter (1992), *Applied Statistics for Engineers and Physical Scientist*. On January 28, 1986 the *Challenger* space shuttle was launched from Cape Kennedy in Florida on a January morning. Meteorologists on the previous day had predicted temperatures at launch to be around $30°F$. The night before launch there was much debate among engineers and NASA officials whether such a low-temperature launch was safe. Several engineers advised against a launch because they thought that O-ring failures were related to temperature. Data on O-ring failures experienced in previous launches were available and were studied the night before the launch. There were seven previous launches in which O-ring failures occurred. A plot of the number of O-ring failures versus temperature is given below:

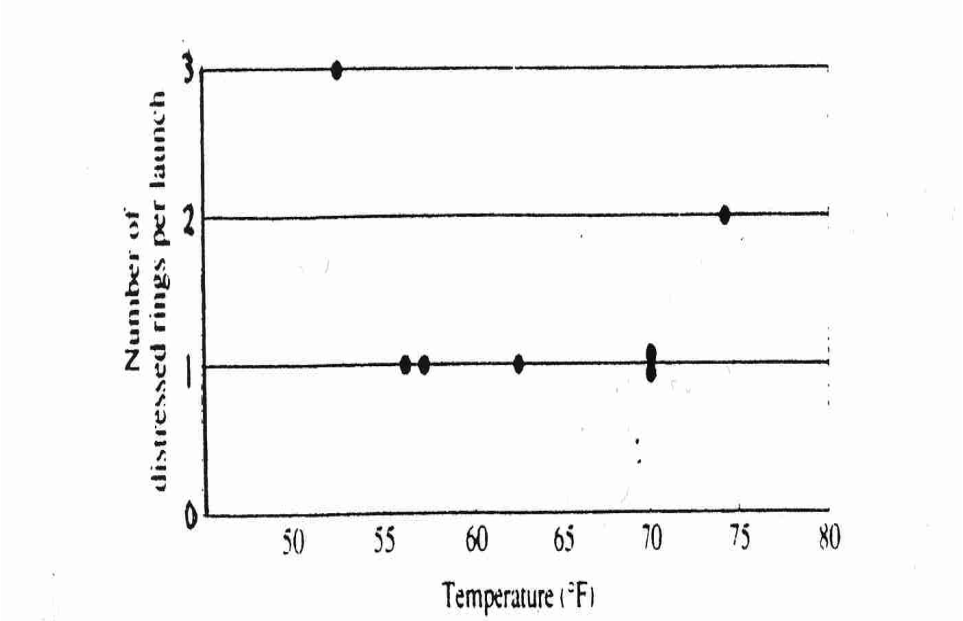


FIGURE 1.5-1 Scatter plot of number of distressed rings per launch against temperature.

From this plot alone there does not seem to be a strong relationship between the number of O-ring failures and temperature. Based on this information, it was decided to launch. The launch resulted in disaster: the loss of seven astronauts, billions of dollars, and a serious setback in the space program. The major problem with the above plot is that the engineers did not display all the data that were relevant to the question of whether O-ring failure is related to temperature. They only looked at the launches where there were failures; they ignored the launches where there were no failures. A scatter plot of the number of O-ring failures per launch against temperature using data from all previous shuttle launches is displayed here:

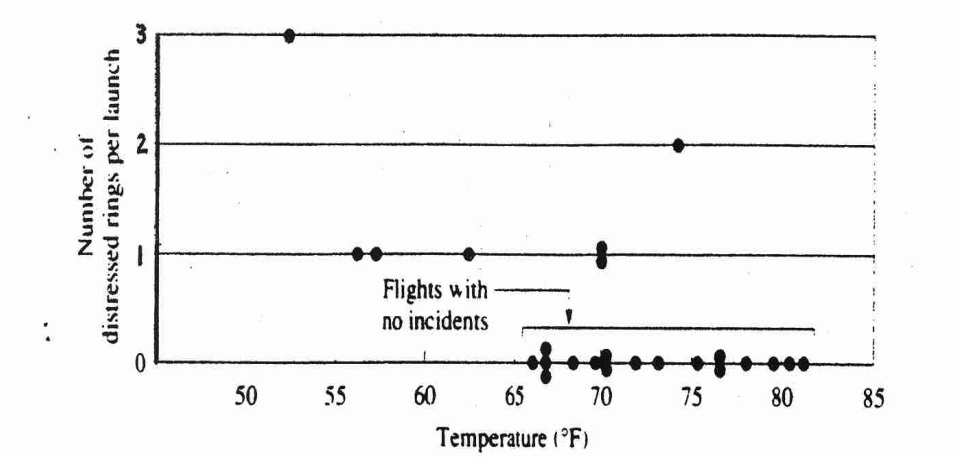


FIGURE 1.5-2 Scatter plot of number of distressed rings per launch against temperature (all data).

This plot reveals a relationship between failures and temperature.

No. Distressed O-rings per Launch	No. of Launches	Temp. at Launch
0	17	T$\geq$ 66
1	5	$55 \leq T \leq 70$
2	1	T = 75
3	1	T = 52
Total	24	

Thus, if $T > 60°F$, then only $\frac{4}{21} = 19\%$ of the launches had O-ring distress, whereas,
if $T < 60°F$, then $\frac{3}{3} = 100\%$ of the launches had O-ring distress.

Furthermore, an extrapolation is required, that is, a prediction of the probability of an O-ring failure at temperatures which are outside the range of previous launch temperatures. The temperature at *Challenger's* launch was only $31°F$, while the lowest temperature recorded at a previous launch was $51°F$. It is always very dangerous to extrapolate inferences to a region of values for which there is not data. If NASA officials had looked at this plot, certainly the launch would have been delayed.

This example illustrates why it is so important to have statistically minded engineers involved in important decisions. Ron Snee, a noted applied statistician has stated many times: **"In God We Trust; Others Must Have Data."**.

This example raises two important points. The importance of scatter plots where we plot one variable against another variable. Secondly, is the importance of plotting *relevant data*. In the *Challenger* study a scatter plot was used in reaching the decision to launch; however, not all the relevant data were utilized. It takes knowledge of statistics to make good decisions, as well as knowledge of the relevant subjects, common sense, and an ability to question the relevance of information.

How risk models failed Wall Street, Washington

Houston Chronicle, October 2, 2008

■ 'Value at risk' involves a huge conceptual error

By JAMES G. RICKARDS

CROOKED mortgage brokers, greedy investment bankers, oblivious rating agencies and gullible investors have all been faulted in the financial crisis, and there is bipartisan agreement that regulators were asleep at the switch. It's all well and good to call for substantial new oversight. But if regulators were oblivious to the danger, the question is why.

In the case of Fannie Mae and Freddie Mac, the answer seems easy: Their massive lobbying machines thwarted every legislative attempt at reform. But what about the Fed, the Treasury and the Securities and Exchange Commission, agencies that are not above politics but are known for their professionalism and expertise? Surely they had the capability and motivation to avoid a calamity of the type that is occurring. Why did they fail?

The problem is that Wall Street and regulators relied on complex mathematical models that told financial institutions how much risk they were taking at any given time. Since the 1990s, risk management on Wall Street has been dominated by a model called "value at risk" (VaR). VaR attributes risk factors to every security and aggregates these factors across an entire portfolio, identifying those risks that cancel out. What's left is "net" risk that is then considered in light of historical patterns. The model predicts with 99 percent probability that institutions cannot lose more than a certain amount of money. Institutions compare this "worst case" with their actual capital and, if the amount of capital is greater, sleep soundly at night. Regulators, knowing that the institutions used these models, also slept soundly. As long as capital was greater than the value at risk, institutions were considered sound — and there was no need for hands-on regulation.

Lurking behind the models, however, was a colossal conceptual error: the belief that risk is randomly distributed and that each event has no bearing on the next event in a sequence. This is typically explained with a coin-toss analogy. If you flip a coin and get "heads" and then do it again, the first heads has no bearing on whether the second toss will be heads or tails. It's a common fallacy that if you get three heads in a row, there's a better-than-even chance that the next toss will be tails. That's simply not true. Each toss has a 50-50 chance of being heads or tails. Such systems are represented in the bell curve, which makes clear that events of the type we have witnessed lately are so statistically improbable as to be practically impossible. This is why markets are taken by surprise when they occur.

But what if markets are not like coin tosses? What if risk is not shaped like a bell curve? What if new events are profoundly affected by what went before?

Both natural and man-made systems are full of the kind of complexity in which minute changes at the start result in divergent and unpredictable outcomes. These systems are sometimes referred to as "chaotic," but that's a misnomer; chaos theory permits an understanding of dynamic processes. Chaotic systems can be steered toward more regular behavior by affecting a small number of variables. But beyond chaos lies complexity that truly is unpredictable and cannot be modeled with even the most powerful computers. Capital markets are an example of such complex dynamic systems.

Think of a mountainside full of snow. A snowflake falls, an avalanche begins and a village is buried. What caused the catastrophe? The value-at-risk crowd focuses on each snowflake and resulting cause and effect. The complexity theorist studies the mountain. The arrangement of snow is a good example of a highly complex set of interdependent relationships; so complex it is impossible to model. If one snowflake did not set off the avalanche, the next one could, or the one after that. But it's not about the snowflakes; it's about the instability of the system. This is why ski patrols throw dynamite down the slopes each day before skiers arrive. They are "regulating" the system so that it does not become unstable.

The more enlightened among the value-at-risk practitioners understand that extreme events occur more frequently than their models predict. So they embellish their models with "fat tails" (upward bends on the wings of the bell curve) and model these tails on historical extremes such as the post-Sept. 11 market reaction. But complex systems are not confined to historical experience. Events of any size are possible, and limited only by the scale of the system itself. Since we have scaled the system to unprecedented size, we should expect catastrophes of unprecedented size as well. We're in the middle of one such catastrophe, and complexity theory says it will get much worse.

Financial systems overall have emergent properties that are not conspicuous in their individual components and that traditional risk management does not account for. When it comes to the markets, the aggregate risk is far greater than the sum of the individual risks; this is something that Long-Term Capital Management did not understand in the 1990s and that Wall Street seems not to comprehend now. As long as Wall Street and regulators keep using the wrong paradigm, there's no hope they will appreciate just how bad things can become. And the new paradigm of risk must be understood if we are to avoid lurching from one bank failure to the next.

Richards was general counsel of Long-Term Capital Management from 1994 to 1999. He works for Omnis Inc., a McLean, Va., as a consultant on national security and capital markets. This article originally appeared in The Washington Post.

Computer model helps predict violence

By JON BARDIN
Los Angeles Times

LOS ANGELES — In August 2010, shortly after WikiLeaks released tens of thousands of classified documents that cataloged the harsh realities of the war in Afghanistan, a group of friends — all computer experts — gathered at the New York City headquarters of the Internet company Bitly Inc. to try and make sense of the data.

The programmers used simple code to extract dates and locations from about 77,000 incident reports that detailed everything from simple stop-and-search operations to full-fledged battles. The resulting map revealed the outlines of the country's ongoing violence: hot spots near the Pakistani border but not near the Iranian border, and extensive bloodshed along the country's main highway. They did it all in just one night.

Now one member of that group has teamed up with mathematicians and computer scientists and taken the project one major step further: They have used the WikiLeaks data to predict the future.

Based solely on written reports of violence from 2004 to 2009, the researchers built a model that was able to foresee which provinces would experience more violence in 2010 and which would have less. They could also anticipate how much the level of violence went up or down.

The project, whose results were published online recently by the Proceedings of the National Academy of Sciences, is part of a growing movement to understand and predict episodes of political and military conflict using automated computational techniques.

The availability of huge amounts of data combined with steady increases in computing power has prompted experts to bring the rigor of objective quantitative analysis to realms that were once considered fundamentally subjective, including literature and the study of social groups.

"For the first time, we have large data sets from places like Facebook and Twitter that we can analyze with high-powered computers and get meaningful results," said Paulo Shakarian, a computer scientist at the U.S. Military Academy at West Point, who is working on an algorithm to predict the location of insurgent weapons caches. "Iraq and Afghanistan are the very first conflicts where we have been collecting as much data as we possibly can."

In the case of the WikiLeaks data, the researchers sought to find a general pattern to the violence in Afghanistan and use it to predict how it would change in each province in 2010 — the year President Barack Obama increased the number of U.S. troops in the country.

"The model we employed is both complex and simple," said Guido Sanguinetti, an expert in computational sciences at the University of Edinburgh in Scotland and the study's senior author. "It doesn't take in any knowledge of military operations or political events, and it treats all types of violence exactly the same, whether it's a stop-and-search or a big battle."

Even with these ostensibly key details missing, the researchers found that they could predict 2010's events with striking accuracy.

And the model wasn't tripped up by Obama's decision to send 30,000 additional troops, which introduced a new dimension to the Afghanistan conflict.

"Our findings seem to prove that the insurgency is self-sustaining," Sanguinetti said. "You may throw a large military offensive, but this doesn't seem to disturb the system."

The study authors said they were most surprised that the model could predict activity even in Afghanistan's relatively quiet northern provinces, where there were fewer data points available to analyze.

"This shows that the escalation we see isn't just attributed to the noise in the data," said study leader Andrew Zammit Mangion, a computational sciences researcher at the University of Edinburgh. Instead, he said, patterns existed nearly everywhere.

Michael Ward, a political scientist at Duke University who has shown that location data can improve predictions of conflicts, said the study pointed the way to future research.

"Suppose you could say, 'This is the effect on violence if you build different types of infrastructure,'" he said. "They don't do that, but they've set up the framework to do it."

The study also shows why it's important to make as much data public as possible, Ward said. Without WikiLeaks, he said, a study like this would have been far more difficult to carry out.

Clionadh Raleigh of Trinity College Dublin, who uses data to predict violence in Africa based on factors such as the outcomes of local elections, said the Afghanistan model could be made even better by including variables such as the political party in power.

"Violence, in general, is a really good predictor of future violence," she said. But even better would be "to figure out what stops the cycle of conflict."

Quantitative rigor is making its way into some surprising fields of study. In 2010, a few months after WikiLeaks' data dump, Google released a database of every single word contained in thousands of books published between 1800 and 2000 — about 4 percent of all books ever printed. That has enabled some intrepid researchers to close in on the final frontier: Studying literature with advanced math.